

Analysis of the Relation Between Mobility Signatures and Product Usage of Bank Customers on Product Recommendation Systems

by

Çağın Ürküp

A Dissertation Submitted to the
Graduate School of Sciences and Engineering
in Partial Fulfillment of the Requirements for
the Degree of
Master of Science

in

Industrial Engineering



September 7, 2017

**Analysis of the Relation Between
Mobility Signatures and Product
Usage of Bank Customers on
Product Recommendation Systems**

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a master's thesis by

Çağın Ürküp

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Committee Members:

Assoc. Prof. Sibel Salman

Prof. Burcin Bozkaya

Assoc. Prof. Özden Gür Ali

Date: _____

ABSTRACT

Conventional financial decision support systems are based on distinct individual attributes like gender, marital status, age and occupation, while being unaware of the spending habits or spatio-temporal mobility of the individuals. The rapid growth of mobile payment and geo-aware systems, and the emergence of big data present the opportunity to investigate human consumption patterns across space and time. We analyze a one-year transaction dataset of a leading bank in Turkey to understand the relation between mobility behaviors and product usage patterns for Individual Consumer Loan and Retirement Savings products. After data preprocessing, we filter the customers who provide more information in terms of mobility. In order to evaluate the product recommendation model in different circumstances, we divide the data into different scenarios with respect to the last 3 months' product usage pattern, time period, demographic, financial and mobility features. We conduct computational experiments to determine the best feature selection methods and the resulting features, and the best classification algorithms for each dataset independently. Finally, we analyze the classification results and report on most interesting product usage patterns and customer characteristics.

ÖZETÇE

Geleneksel finansal karar destek sistemleri, bireylerin zaman-mekansal hareketlilik ve harcama davranışlarından habersiz bir şekilde, yaş, medeni durum, iş durumu ve cinsiyet gibi belirgin bireysel özelliklere dayanır. Mobil ödeme ve jeografik farkındalık sistemlerinin hızlı gelişimi, ve büyük data olgusunun oluşması, insanların harcama paternlerinin uzay-zaman açısından araştırılması ve ürün tavsiye sistemlerinin kullanılması fırsatlarını ortaya çıkarmıştır. Bu tez çalışmasında Türkiye'nin önde gelen bankalarından birinin bir yıllık banka işlem datası analiz edilerek, bireysel ihtiyaç kredisi ve bireysel emeklilik ürünlerinin kullanılma patternleri ile bireylerin hareketsetel davranışları arasındaki ilişkiyi araştırılmıştır. İlk olarak çeşitli data ön işlemlerinden sonra müşterilerin hareketlilik açısından daha çok data sahibi olanları filtrelenmiştir. Sonrasında veri seti için ayrı ayrı sayısal deneyler yapılarak özellik seçim algoritmaları ve ürün önerisi sistemleri arasından en iyi performans gösterenler belirlenmiştir. Son olarak ürün önerisi sonuçlarını analiz edilmiştir ve en ilgi çekici ürün kullanım düzenleri ve müşteri karakteristikleri raporlanmıştır.

ACKNOWLEDGMENTS

First, I would like to express my deep gratitude to my advisor, F. Sibel Salman for her incredible and precious guidance during my undergraduate and graduate studies. She generously shared her knowledge, passion, time and experience with me with great kindness. I also appreciate her endless and invaluable support during my transition between academia and business. I also want to thank Prof. Burcin Bozkaya for suggesting me my thesis topic and providing the opportunity for working on the data in Sabancı University. I am also grateful about his guidance provided to me with his experience through my studies. I thank my committee member Assoc. Prof. Özden Gür Ali for kindly accepting to be in my thesis committee and for her valuable feedback.

During my undergraduate and graduate studies, I had the chance to meet awesome people whose friendships are priceless. Especially, I would like to express my gratitude to Kaan, who is one of the main reasons of my success. I could never complete my thesis without his support and friendship. I want to thank Ezgi for her precious support, understanding and patience during my undergraduate study.

Many special thanks go to my roommate Kemoş, who made me feel like I am always at home and he was always there for me through the good and bad times. I would like to thank my office buddy from ENG 218, Gökçen Gökmen. Even though we barely saw each other, we have built a valuable friendship.

I extremely feel lucky to have the chance of taking lectures from Prof. Emre Alper Yıldırım, who is one of the best instructors I have ever seen. I would like to thank Prof. Metin Türkay for sharing his valuable experiences with us. I am also grateful to Prof. Ceyda Oğuz for her valuable support and trust in me and my friends.

Finally, I am grateful to my family for their unhesitant support throughout my studies and life in general. I dedicate this thesis to my family, primarily my parents, Nurten Ürküp and Hasan Ürküp, who are just perfect.



TABLE OF CONTENTS

List of Tables	ix
List of Figures	xi
Chapter 1: Introduction	1
Chapter 2: Literature Review	4
2.1 Data Preprocessing and Feature Selection	4
2.2 Product Recommendation	9
2.3 Mobility Behavior in Data Mining	15
Chapter 3: Data and Preprocessing	23
3.1 Data	23
3.2 Features	25
3.2.1 Descriptive Statistics	34
3.3 Data Preprocessing	50
3.4 Dataset Separation	51
3.4.1 Time Interval and Feature Type	51
3.4.2 Separation of Customers	53
Chapter 4: FSE Methods and Classifiers	55
4.1 Performance Evaluation	55
4.2 Feature Selection and Extraction Methods	58
4.3 Classification Algorithms	66

Chapter 5:	Computational Results	69
5.1	Individual Consumer Loan	69
5.1.1	Computational Results of ICL	69
5.1.2	Inferences about ICL	80
5.2	Retirement Savings	85
5.2.1	Computational Results of RS	85
5.2.2	Inferences about RS	97
Chapter 6:	Conclusion	99

LIST OF TABLES

2.1	Studies on Product Recommendation	14
2.2	Studies on Mobility Behavior	22
3.1	Regularity Measure Comparison Periods	27
3.2	Feature Explanations	32
3.3	Descriptive Statistics of All Features	49
3.4	Dataset Types Explanations	52
4.1	ICL Class-Zero Optimum Assignments	60
4.2	ICL Class-Ones Optimum Assignments	61
4.3	ICL Class-All Optimum Assignments	62
4.4	RS Class-Zero Optimum Assignments	63
4.5	RS Class-Ones Optimum Assignments	64
4.6	RS Class-All Optimum Assignments	65
5.1	ICL Class-Zero Results	70
5.2	ICL Class-Ones Results	72
5.3	ICL Class-All Results	73
5.4	ICL Class-Zero Results on New Customers	74
5.5	ICL Class-Ones Results on New Customers	75
5.6	ICL Class-All Results on New Customers	76
5.7	ICL Feature Importances	79
5.8	RS Class-Zero Results	87
5.9	RS Class-Ones Results	88
5.10	RS Class-All Results	89

5.11 RS Class-Zero Results on New Customers 90

5.12 RS Class-Ones Results on New Customers 91

5.13 RS Class-All Results on New Customers 92

5.14 RS Feature Importances 95



LIST OF FIGURES

4.1	Confusion Matrix	56
5.1	ICL Class-Zero 6 Month	77
5.2	ICL Class-Ones 9 Month	77
5.3	ICL Class-All 3 Month	77
5.4	ICL Class-All 6 Month	77
5.5	ICL Class-All 9 Month	78
5.6	ICL Class-All 11 Month	78
5.7	ICL Feature Effects Graph	80
5.8	Cluster Centers Throughout Turkey	82
5.9	ICL Feature Effects Pie Chart	84
5.10	ICL cluster+d-l-r Example	84
5.11	ICL radial+d-l-r Example	84
5.12	RS Class-Zero 6 Month	94
5.13	RS Class-Ones 3 Month	94
5.14	RS Class-All 3 Month	94
5.15	RS Class-All 6 Month	94
5.16	RS Class-All 9 Month	95
5.17	RS Class-All 11 Month	95
5.18	RS Feature Effects Graph	96
5.19	RS Feature Effects Pie Chart	96

Chapter 1

INTRODUCTION

Digital technologies cause the rapid growth and increasing availability of big data. Day by day, millions of data is being gathered with full of information inside. Most of the businesses have the same struggle: how to use all this information. Businesses have enormous data sources and they are investing in systems that can analyze their data and find the hidden patterns through various technical approaches. Better understanding of their customers affect their business decisions, which leads to better comparative advantage and increase in profits. However, the huge size of transactional databases reduces the efficiency and quality of recommendations. On the other side, the customers are also struggling in terms of an increasing variety of products and information never seen before. Rodrigues and Ferreira (2016) state that this confusion causes an increase in diversity on customers' demand, which directly challenges the businesses to provide the right products to right customers. This is where the recommendation systems come up and cope with this challenge. Recommendation systems are tools to provide suggestions to the customers according to customers' preferences, that are obtained through data analysis (Riyaz and Varghese (2016)). The recommendation system reveals the hidden patterns in big data and simplify the recommendation process for businesses. On the other side, customers are not exposed to products that they will never buy, instead products that they are likely to purchase are suggested to them, which also simplify the process from the customers' perspective. To sum up, recommendation systems not only help businesses, but also assist customers.

Various big data types have been investigated in the literature on behalf of creation of recommendation systems. For example, Sanchez-Moreno et al. (2016) use previously listened songs data to obtain user's preferences and recommend unheard songs to users related to the preferences. Another example is by Liu and Shih (2005). They use e-commerce data to recommend related products to customers for their future purchases. This type of product recommendation studies mostly gather financial and historical data of the customers in order to recommend the best products. Moreover, there exists some studies on location predictability of people, that use mobility behaviors of the customers via digital traces. Traditional financial decision systems rely on explicit individual traits like gender, job type, marital status and age, while being oblivious to the habits of the individual or spatio-temporal mobility involved (Singh et al. (2015)). Emerging trends in mobile payment and geo-aware systems and the resulting big data present an opportunity to study human consumption patterns across space and time. While some studies use phone call, wi-fi location, check-in data, etc. to study mobility behaviors, some others use expenditure data with geographical information, such as credit card expenses. In our time, any data including location information leads to patterns on the mobility behavior of users. For example, Noulas et al. (2012) use Foursquare check-in data to predict the next locations of the users. As another example, Isaacman et al. (2010) use mobile phone data to demonstrate that there are different mobility patterns in different in two cities. Therefore, mobility traces can also be used to gather information and generate recommendations about users.

Digitalization of expenses and mobile banking developments turn banks into largest data sources. Banks have enormous data sources with various data types including financial, demographic and mobility information about customers, but without any analysis of the data, the data does not mean anything alone (Kung et al. (2014)). The mobility data, can be expressed as digital traces of human behavior, is gathered from mainly credit card expenses and ATM (Automatic Teller Machine) transactions, which allow banks to trace their users through space and time. The discovery

of these data types occurring in data sources creates tremendous new opportunities in the fields of data analysis, which are experiencing developments due to the availability of large-scale datasets reflecting human interactions, communications, and mobility (Sobolevsky et al. (2014)). For instance, Singh et al. (2015) extract both demographic, financial and mobility information from a bank transaction data and seek the relation between financial well-being and these features. Another study that uses the opportunity in this area is Sobolevsky et al. (2014), who use a bank transaction data to extract the relation between mobility patterns and nationality.

The main contribution of this study is seeking the relation between mobility signatures and product usage patterns using bank transaction data. We analyze one year bank transaction data and we select two products to investigate this relation. Nowadays, banks have millions of customers, hundreds of products and enormous data sets that are expending everyday. Although, in general the data is in very good quality, banks cannot give the meaning to it. Banks want to recommend related products to customers so that they will increase their revenue: nevertheless, customers' responses to recommended products by the banks are fractional. In order to increase this important ratio, the data should be analyzed properly, which will lead to the extraction of the customer product usage patterns, and then related recommendations should be performed.

In our study, we implement various data mining approaches, feature selection/extraction methods and classification algorithms on our dataset. Chapter 2 provide the necessary background and literature on data preprocessing, feature selection/extraction, product recommendation and mobility behavior and data mining techniques. The performance evaluation criteria, the feature selection/extraction methods and classification algorithms are presented in Chapter 4. The computational experiments and deductions are examined in Chapter 5. Chapter 6 concludes the thesis, summarizes the main contributors and inferences we obtain during our study.

Chapter 2

LITERATURE REVIEW

In this section, a review of the literature on data preprocessing, feature selection, product recommendation and mobility behavior is provided. Although each of these problems is extensively studied in the literature, to the best of our knowledge there does not exist any research on product recommendation by analyzing the relation between an individual's mobility signatures and product usage.

2.1 Data Preprocessing and Feature Selection

We present an overview of studies related to data preprocessing, feature extraction and feature selection methods. Missing value replacement, bootstrapping, data transformation, normalization and resampling are common methods used in data preprocessing in data mining (Wang and Huang (2009), Cohen et al. (2009), Sevilla and Sola (1997), Singh et al. (2015)), the advantages of these methods and mostly used techniques are discussed further in this section. For feature selection/extraction, filter methods, wrapper methods and embedded methods are the most widely used techniques in the literature (Chandrashekar and Sahin (2014), Sun et al. (2005), Kohavi and John (1997), Ding and Peng (2005)). These techniques are also discussed in this section.

Data preprocessing is the very first step in data mining, and perhaps the most crucial one. Without getting to know the data carefully in advance, the classification task could be misleading (Wang and Huang (2009)). Since the preprocessing step is performed before feature selection and model implementation, we first examine the data preprocessing methods. Wang and Huang (2009) use missing value replacement as a preprocess by changing the missing values with the averages or the medians of the

values. This replacement helps the prediction method to handle the missing values easily. This constitutes data manipulation but as long as the proportion of missing data is small, it does not have significant effects on the prediction. Moreover, Wang and Huang (2009) use the max-min normalization method on the selected features, which basically scales each feature value according to the maximum and minimum value of the corresponding feature set.

The class imbalance problem is an interesting topic in machine learning and data mining (Weiss and Provost (2001)). If the classes are imbalanced in data, then the classifier can be negatively affected and would not give good results. The classifier can ignore the minority class and predict all the sample as a majority class (Drummond and Holte (2003)). The percentage of successful predictions might seem high in this case, but the classifier actually does nothing for the minority class. The reason that the percentage is so high is that the classes are imbalanced. Repeatedly taking samples of a data set in a controlled fashion is the basic idea in re-sampling. Then, desired summary statistics are computed for each sample, and samples are combined to estimate a property of the entire data set more robustly (Cohen et al. (2009)).

In the literature, there are basically two standard ways of re-sampling. The bootstrapping method is the straightforward approach. To explain bootstrapping further, let us say, from a population of size N , we pick k members from the population and compute the desired statistic O_0 . Then, the subsample is replaced and another random k members are picked. The new statistic O_1 will change. If this process is repeated tens of thousands of times, the distribution of the resulting O_i 's is called the sampling distribution. Then the sampling distribution will be normally distributed because of the Central Limit Theorem. Thus the mean of a large sampling distribution generates an accurate measure O^* (Cohen et al. (2009)). The second standard re-sampling method is the jackknife method. In this method, a summary statistic O_i is repeatedly recomputed by leaving out some data points from the full data set to measure the influence of certain subpopulations. To generate a good estimator for the underlying desired statistics, the resulting set of observations are used as a

sampling distribution like in the bootstrapping method (Cohen et al. (2009)). Resampling is commonly used in the preprocessing step for a better data mining structure. Wang and Huang (2009) and Singh et al. (2015) both use resampling to improve their classifiers and data mining techniques.

Another crucial preprocessing method is normalization, which is basically the process of organizing the attributes of a relational database to reduce data redundancy and improve data integrity. Sevilla and Sola (1997) conclude that normalization of input data prior to network training has two clear advantages. The first one is that the estimation errors can be reduced. The second one is that the calculation time needed in the training process to obtain such results is reduced. Most of the studies in the literature use normalization on their data sets as a preprocessing method e.g., Wang and Huang (2009), Singh et al. (2015).

Data transformation is another data preprocessing method. Some classifiers cannot handle both numeric and nominal features simultaneously. When a data set includes both nominal and numeric attributes, generally the nominal features are converted to numerical features. For example, if a nominal feature has three outcomes, low, medium and high, three columns can be generated with 0/1 numeric values. The medium column will have 1 values where the nominal feature has value medium and 0 for others. After this process is performed for all the data points, one nominal column will be replaced by three numeric columns and now it can be embedded to the models that only use numeric features. The opposite also can be done if the classifier only requires nominal attributes (Cohen et al. (2009)).

Chandrashekar and Sahin (2014) provide a survey about feature selection methods. They claim that feature selection methods improve prediction quality, reduce computational times and patterns can be recognized easily. The feature selection concept is basically selecting a subset of the whole variables of the dataset. The dependent variables should not be selected simultaneously in feature selection, because when one of them is selected, the other one cannot provide additional information to the system. Thus, among all the dependent variables, choosing only one is the

optimal selection; choosing many will slow the system and could decrease the prediction accuracy. Therefore, eliminating dependent variables is crucial for data mining quality. Besides, variable selection is another essential process for data mining. In variable selection, input variables are examined and new features are created with the combination of existing features to better estimate the output variable (class label). For example, combining the data of the number of withdraw amounts and the amount of credit card payments, cash/credit card expenditure ratio can be obtained for a person and can be used as a new feature on the dataset (Chandrashekar and Sahin (2014)).

There are two main variable elimination methods in the literature. These methods are filter and wrapper (Chandrashekar and Sahin (2014)). Variable ranking techniques are mainly used in filter methods, which is basically choosing the variables by ordering them. Ranking is done according to the simplicity and success, and then the ones under a threshold value are eliminated (Chandrashekar and Sahin (2014)). Moreover, relevancy is another important concept for choosing the variables. A variable can be independent from the input data but it cannot be independent from the class label, the target variable. If it is independent from the class label, then it should be eliminated because it has no effect on the variable to be predicted. Pearson correlation coefficient and mutual information (entropy) are the most commonly used filtering methods to find relevance and eliminate attributes (Law et al. (2004)).

Wrapper methods try to find the best combination of feature subsets that maximizes the given objective function. Since there are 2^N combinations and this is a combinatorial problem, heuristics are proposed to find the best feature subsets. For example, Kohavi and John (1997) use a tree structured branch and bound algorithm to evaluate the subsets, but the search time grows exponentially as the number of features increases above 40. It is challenging for exhaustive search algorithms to deal with these problems as their size increases. Therefore, more simplified algorithms like sequential search or heuristic search algorithms are recommended by Chandrashekar and Sahin (2014).

Sequential selection algorithms work iteratively and evaluate the objective function value at each iteration. Sequential Feature Selection (SFS) starts with an empty set and adds one feature to the current subset if it increases the objective value and stops when the required number of features are collected. Sequential Backward Selection (SBS) is the opposite of SFS and it starts with the whole feature set and removes the subsets that increase the objective function value until the stopping criteria. Sequential Floating Forward Selection (SFFS) is the combination of SFS and SBS. It starts with an empty set and adds the features according to the objective function, but it also removes the features that decreases the objective function (Pudil et al. (1994), Reunanen (2003)). The drawback of the SFFS method is that, it does not consider the dependent features, so it can add highly correlated features to the subset if they keep increasing the objective function. To avoid this drawback, Sun et al. (2005) use Adaptive Sequential Forward Floating Selection (ASFFS) method, where two parameters are used to specify the number of features to be added and removed during the add and remove phases.

An alternative wrapper method uses a Genetic Algorithm (GA) as a heuristic search algorithm. In general, a GA works with a parent, offspring, crossing over and mutation operators where parents are exposed to crossing over so that diversified and qualified offspring are generated to explore better solutions. Cordon et al. (2006) use CHC-GA algorithm trying to maximize the given objective function by selecting the best subset of given features. As Cordon et al. (2006) state, the CHC-GA algorithm is fast and performs an efficient search on the dataset while maintaining diversity.

Ding and Peng (2005) propose max-relevancy, min-redundancy (mRMR) as an embedded feature selection method that consists of two stages. In the first stage, the optimum number of features giving the lowest cross-validation classification error is determined. In the second stage, wrapper methods are used to evaluate different subsets with size found in the previous stage with smallest classification error. The method is basically trying to maximize relevancy between the class label and minimize the redundancy between each selected feature simultaneously. The mutual

information (MI) formula is used to compute the relation between two variables. As MI value increases, features become more relevant and as it decreases, they become more independent. The main idea is to maximize the MI between the feature and the class label, while minimizing the MI between selected features. Ding and Peng (2005) basically extracted the redundancy from relevancy in the objective function, which gives them equal importance. This makes sense because as relevancy increases, the objective function value increases and as redundancy increases, the objective function value decreases.

2.2 Product Recommendation

In this section, studies on product recommendation are investigated in detail. Some studies use customer lifetime values with recency frequency and monetary features to cluster the customers and then apply product recommendation (Liu and Shih (2005) and Rodrigues and Ferreira (2016)). Some other studies use the collaborative filtering method and train the algorithm with similarity measures of products or customers (Riyaz and Varghese (2016) and Wei et al. (2016)). We also find studies that apply content-based filtering approaches for product recommendation e.g., Yewatkar et al. (2016). Some studies work on problems where there is only a small-sized available data and gray sheep customers are involved e.g., Sanchez-Moreno et al. (2016). Moreover, several studies use hybrid recommendation systems combining collaborative filtering and content-based filtering methods (Rodrigues and Ferreira (2016)). To conclude, this section combines studies that use different kinds of product recommendation approaches and various algorithms. The most commonly used approaches are collaborative filtering, content-based filtering and hybrid recommender algorithms.

While multiple studies have connected an individual's product usage to her demographic data and product similarity measures (Choi et al. (2006), Park and Chang (2009)), no publicly reported studies have looked at fine-grained data of the individual's current mobility behavior (mobility signatures) to predict product usage. Choi et al. (2006) create a system where customers input their preference through an

interface and customers' similarity measures are calculated with these inputs. The similarity measure of the products are also calculated and the system recommends a list of k products as an output to the customer. They use multi-attribute decision making (MADM) to find the utility values of the products. This system is different from ours since the system takes inputs directly from the customers, analyzes the information and gives an output of recommended products. In comparison, we do not take direct preferences from customers, we use bank data, and instead of a list, we predict the usage pattern of the products that we analyze. Park and Chang (2009) propose a new customer profile model, using product features, and individual and group interests with weights. Then, the product groups are generated according to their corresponding similarity measures. The similarity between a customer and each product group is calculated and the recommendation is done according to the similarity measures with the most highly ranked k products.

Riyaz and Varghese (2016) use collaborative filtering to recommend products on Amazon's data. They start with extracting user preferences from the data. Then, they find the similar items based on the user preferences and as the final step they recommend products to customers. They basically create a list with customerID, productID and a rating for each product and customer couple. Afterwards, they use the Pearson correlation coefficient measure to calculate the similarity between the products. Next, item-to-item collaborative filtering is performed. When generating the recommendations, they use the similarities between items. They basically assume that if two items have the same rating from some users, then they are similar items and the user may like the other item that he has never bought before. For example, if items A, B and C are rated similarly by customer X, then they are similar items and if user Y liked item B, then the system recommends A and C to user B (Riyaz and Varghese (2016)).

Yewatkar et al. (2016) propose to use a smart system for shopping in supermarkets to reduce the long queues, to provide more information for the products and to help the customers choose the right product. The system basically uses RFID tags and

readers. Each product has an RFID tag and the customers can easily scan and learn what the product really is and they can make their payments with the carts given to them. The proposed product recommendation algorithm is generated by a Bayesian network and an iterative dichotomiser 3 algorithm, which is a decision tree classifier. The product recommendation is based on the history of the user preferences and the similarity between the products (Yewatkar et al. (2016)).

Liu and Shih (2005) use Customer Lifetime Values (CLV), an approximation of how much profit the customer will bring to the bank, to recommend products to their customers. They evaluate the CLV values according to a weighted Recency, Frequency and Monetary (RFM) combination. Analytical hierarchy process (AHP) is used to obtain the weights of each element. Then, customers are clustered according to the RFM values using the k-means clustering method. Afterwards, an associate rule mining approach is performed on the data to extract recommendation rules and frequent purchase patterns from each group of customers. The extracted frequent purchase patterns represent the common purchasing behavior of customers with similar product purchases. Therefore, the approach recommends products to customers according to the frequency of purchase patterns of customers with similar product purchases (Liu and Shih (2005)).

Rodrigues and Ferreira (2016) propose a hybrid product recommendation algorithm which combines content-based, collaborative filtering and data mining techniques, based on shared CLV. First they start with Extraction Transformation Loading (ETL) to collect and clean transactional data. This step involves extraction of the data, treatment of inconsistencies, handling of missing values and errors, integrity faults, etc. The final step of ETL is loading the data into the staging area, which is the database. They also apply sigmodal normalization to the data not to dismiss large outlier values. They cluster the customers using k-means with similar CLV. They determine the number of clusters by the elbow shape method, change the number k and observe the graph when an elbow shape occurs. For better understanding of the clusters, they divide each cluster into groups according to five different RFM

values (very low, low, etc.). Afterwards, they use the Apriori associate rule mining algorithm to identify the relationships among the set of items in the data. By using this algorithm, they obtain all candidate products in a sorted form and the N highest ranked products are used for the recommendation. N is determined between 2 to maximum number of items in the baskets. They also state that both frequency and accuracy of the predictions are important. Thus, too frequent recommendations and unwanted recommended products can lead to unhappiness of the customers. This is why they keep the period of recommended products low and they focus on mostly false positive errors, the recommendation of irrelevant products. They outperform the base recommendation system with the hybrid recommendation system in all clusters (Rodrigues and Ferreira (2016)).

Wei et al. (2016) study collaborative filtering and deep learning based recommendation for complete cold start (CCS) and incomplete cold start (ICS) items, which have no rating record available and small record available, respectively. After data preprocessing with normalization and filtering, they extract content features from content descriptions by deep learning neural network algorithms. They apply collaborative filtering timeSVD++ (a single value decomposition algorithm) and an advanced neural network model, SADE, on their datasets. Their four main steps are configuring system parameters, preparing train and testing datasets, training the model by getting the data in train the set and finally performing the recommendations. They conclude that recommendations of the CCS model outperform the ICS case (Wei et al. (2016)).

Sanchez-Moreno et al. (2016) study the collaborative filtering method for music recommendation using playing coefficients that they generate for both artists and users. They mainly focus on gray sheep customers, who cannot directly be assigned to a customer cluster because they show different preferences. Their proposed algorithm is also designed of the case where there is only limited information provided, because they claim that user item and user preference information are not generally available. Their strategy is to determine a coefficient for every user representing the degree to

which they are gray sheep users. They extract the number of times a song played by each user and use this information as a user preference. They derive the ratings for each user and for each artist and a normalized listening coefficient for each user and each artist again. The second coefficient that they derive is the user playing coefficient (UPC) which refers to the number of times that a user plays an artist. They embed these coefficients and ratings into the k-nearest neighborhood algorithm and obtain the recommendations. They use the mean absolute error, normalized mean absolute error and root mean square error techniques to assess the performance of the recommendations. They compare the results of their algorithm without UPC features and they observe that the model with UPC features outperforms the other algorithms (Sanchez-Moreno et al. (2016)).

Throughout the literature review on product recommendation methods, we observe that most of the studies use either collaborative filtering (CF), content-based filtering (CBF) or a combination of these two methods, called hybrid recommendation system. The main difference between CF and CBF is as follows. CF generates customer profiles by examining customer purchases or transactions and find similarity between users. On the other hand, CBF generates product profiles and the recommendation is performed according to the similarity of the products (Park and Chang (2009)). The main idea behind CF is that if person A has the same interests with person B on an issue, then A is more likely to have the same opinion with B on another issue, compared to a random person. The main idea of the CBF is; if person A bought item X in the past, and if item X and Y are similar, then person X is more likely to buy also item Y. There are both weaknesses and strengths of these methods. It is not possible to recommend unique products in CBF because if there is no relation with other products, then the item will not be recommended. On the other hand, the nature of the product cannot be analyzed in CF because the method uses customer profiles. The hybrid method is used in some studies to overcome these weaknesses during the recommendation process (Park and Chang (2009), Rodrigues and Ferreira (2016)). We provide an overview of the studies on product recommendation in Table

Authors	Data Type	Algorithm	Evaluation Criteria	Success Rate
Choi et al. (2006)	Purchasing Data	MADM	Directly Asked Users	5.75 / 7
Park and Chang (2009)	Purchasing Data	Content Based Filtering	Precision, Recall, F1, MAE	0.64 P, 0.33 R, 0.41 F1
Riyaz and Varghese (2016)	Amazon Product Data	Collaborative Filtering	Not Specified	Not Specified
Yewatkar et al. (2016)	Supermarket Data	Naive Bayesian & Decision Tree	Not Specified	Not Specified
Liu and Shih (2005)	E-commerce Data	Collaborative Filtering	Precision, Recall, F1	0.34 P, 0.76 R, 0.47 F1
Rodrigues and Ferreira (2016)	Purchasing Data	Association Rule Mining	Precision, Recall, F1	F1 0.22
Wei et al. (2016)	Rating Data	Collaborative Filtering & Neural Network	RMSE	Not Specified
Sanchez-Moreno et al. (2016)	Music Data	Collaborative Filtering	MAE, NMAE, RMSE, AUC	AUC 0.76

Table 2.1: Studies on Product Recommendation

2.1, where the cited articles' characteristics are compared.

2.3 Mobility Behavior in Data Mining

In this section, studies that combine mobility behavior and data mining methods are investigated in detail. Some studies try to extract mobility behaviors by using various data mining methods and defining novel features related to mobility (e.g., Gonzalez et al. (2008), Isaacman et al. (2010)). Singh et al. (2015) try to embed different mobility behaviors in algorithms and observe the effects of each behavior to financial well-being. On the other hand, Kung et al. (2014) study the work/home commuting distances and obtain promising outcomes. In summary, this section combines studies that use or extract mobility behaviors with data mining techniques.

Song et al. (2004) aim to predict the location of individuals by using Wi-Fi data. They use Order- k Markov predictors and LZ-based predictors and compared the performances with different metrics. Order- k Markov predictor basically predicts the next position of an individual by considering the most recent k positions. LZ-based predictors are based on a popular incremental parsing algorithm. It is similar to order- k Markov predictors but k can go to infinity in LZ-based predictors. They concluded that order- k Markov predictors and LZ-based predictors perform equally well.

In Singh et al. (2015), spatio-temporal behaviors are extracted from bank transaction data and are used to find the relation with financial well-being. In order to achieve this, they process credit card transaction data to generate features representing spatio-temporal aspects of the customer shopping activity. The main mobility features used are; diversity, loyalty and regularity. *Diversity* is the extend to which the customers' shopping activity is diverse in both time and space dimensions. *Loyalty* is the percentage of transactions happening in the individuals' top three most frequently visited stores. *Regularity* is the similarity of individuals' activity over shorter and longer time periods. The authors measure the financial well-being of a customer according to three main features: overspending, trouble and late payments. *Overspending* is the

ratio of expenditures to the income of individuals. The higher this ratio is, the more customers overspend. *Trouble* refers to whether users have defaults. *Late payments* are the late payment of credit card balances. They perform bootstrap aggregation with SMOTE approach as the classifier (Singh et al. (2015)).

Noulas et al. (2012) build a recommendation system for previously unvisited venues from behavioral, social and spatial data. They use check-in data of Foursquare. Then, they rank and classify the venues for each user according to category. They use rank distances, place popularities, geographic distances, activity transitions and place transitions as mobility features. They utilize these features to obtain global mobility of the users. A random walk approach is implemented with steady state conditions and the results are compared with various different algorithms.

Gonzalez et al. (2008) study the trajectory of 100,000 anonymous mobile phone users whose positions are tracked for a six-month period and they observe that, in contrast with the random trajectories predicted by the prevailing Le'vy flight and random walk models, human trajectories show a high degree of temporal and spatial regularity; each individual being characterized by a time independent characteristic travel distance and a significant probability to return to a few highly frequented locations. They use a distribution of displacement function to predict the probability of occurrence of that particular displacement. Moreover, they also calculate the *radius of gyration* for each user, which is interpreted as the characteristic distance traveled by a user when observed up to time t . They group the customers according to the radius of gyration and observe their characteristics. Gonzalez et al. (2008) conclude that the trajectories become more anisotropic as the radius of gyration increases.

Song et al. (2010) investigate what the role of randomness in human behavior is and to what degree individual human actions are predictable. They try to solve the relation between the regular and the random human mobility to characterize the predictability of human dynamics by analyzing mobile phone data. To achieve this, they define an indicator called *entropy*, which is a fundamental quantity to capture the degree of predictability characterizing a time series. The entropy measure is classified

into three main types: random entropy, temporal-uncorrelated entropy and actual entropy. *Random entropy* directly depends on the number of distinct locations visited by the users. *Temporal – uncorrelated entropy* refers to both distinct locations and the frequencies of these locations. *Actual entropy* depends on distinctness, frequencies and the order. At the end, they conclude that despite humans’ deep-rooted desire for change and spontaneity, humans’ daily mobility is characterized by a deep-rooted regularity (Song et al. (2010)).

Kung et al. (2014) describe a methodology for inferring the home/work locations and aggregate commute patterns from mobile phone calls in different countries/cities, in comparison with the car-only GPS traces. They first apply spatial-temporal filtering and eliminate some of the outliers from data. Afterwards they construct individual’s travel portfolios, which are referred to a list of cell towers that the individual frequented, ranked by the time the individual spent in each place. Cell towers are the devices which phone users connect to make a call. Phones connect to the cell tower with the highest signal, meaning that lowest distance in general. Then, the authors identify the home and work locations of individuals by identifying the daytime and nighttime locations in which the user spends the maximum time. Such locations account for more than 50% of the total observed day time and night time times in the user’s travel portfolio. Thus, they simply filter these consequences by day and night, which refer to work and home, respectively. Once the home and work locations are determined for each user, they compute the commute distances by using the great circle distance approach between home and work locations. They also compare their commute distances from another article to prove that the distances are the same. Finally, they determine the morning and evening commute times separately for each user. For the home-work case, they measure the distance between the last call users did the night before and the first call users did in the morning from the work location. Likewise for the work-home distance they follow the same logic. They discover that individuals display limited travel range during the night and different locations show distinct commuting distance profiles. The fact that they are able to observe

common commute features despite the highly diverse nature of their datasets, which consists of countries and cities, offer a compelling demonstration that there are some aspects of human commuting that are universal. Eventually Kung et al. (2014) find a satisfactory correlation between the commute distance and the radius of gyration, which makes sense because radius of gyration includes also the commute distances in itself, and even for some people who generally just change location between work and home, the two measures become almost the same.

Singh et al. (2013) examine Friends and Family dataset, a socio-mobile dataset, to find a relation between socio-mobility and spending behavior. The data is collected for each couple, which eliminates the challenging task of disentangling the spending of the different family members. Thus, this study is more focused on couple level variables with social and spending features. The spending behavior data is collected directly from the participants. Big Five personality questionnaire is used on participants to extract their personalities and these personalities are stated as: extroversion, agreeableness, conscientiousness, neuroticism, and openness. The spending behavior of couples are examined through their diversity, loyalty and overspending characteristics. Afterwards, the Pearson correlation is used to extract the relation between spending behavior and socio-mobile features. The observations of the study are: spending diversity increases as users visit various places, as bluetooth contacts decreases, the loyalty of the users increases, and finally, users with high social activities tend to overspend. They also compare their model with a baseline model and a model without social feature dataset and they outperform other two models with the one with social features.

Isaacman et al. (2010) use anonymous and aggregate statistics of approximate locations of cell phones used in New York and Los Angeles to demonstrate different mobility patterns in two cities. A new daily range concept, the maximum distance that a person travels in one day, is introduced. They calculate the daily ranges for each day of the data and they extract median daily ranges and maximum daily ranges for each customer. The median daily range is an approximation of the daily commute

distance for the users. Moreover, they group the data according to the billing zip codes of the users. For example, in the New York city the users are divided into Brooklyn, Queens, New Jersey areas, etc. They also categorize the ranges according to the weekdays and weekends. They acknowledge that the users spending less time in a location might not use their phone, which means that the data may not represent actual commuting patterns. Isaacman et al. (2010) observe that the median daily ranges of people living in Los Angeles area is two times farther than people living in New York. Nevertheless, the maximum daily range of New Yorkers are 2 to 6 times greater than people living in Los Angeles.

Candia et al. (2008) study the mean collective behavior at large scales and occurrence of anomalous events is their main focus. They use standard percolation theory tools to determine the spatio-temporal anomalies. They define an approximation formula to represent the expected number of calls recorded at any location between some time intervals. They also use another formula to represent the standard deviation of the number of calls. Then, they adopt a simple logic and take the difference between the actual number of calls and the mean number of calls and check whether the difference is greater than the standard deviation multiplied by some constant sets the fluctuation level. When an actual number of calls exceeds this limit, they infer that the call is an anomalous call (Candia et al. (2008)).

Sohn et al. (2010) use mobile phone data to predict the step count and the type of mobility of the users. They basically calculate Euclidean distances for each measurement and determine the similarity of the measures according to these distances. They apply 5-fold cross validation to the whole data and use the average value of the results. They compare the boosted logistic regression with Naive Bayes, Support Vector Machine, AdaBoost (Freund and Schapire (1996)), MultiBoost (Webb (2000)). Sohn et al. (2010) observe that the boosted logistic classifiers perform better than others. They also present a confusion matrix for the mobility type behavior. It is clearly seen that overall accuracy is pretty good with 85% but the model struggles to predict the walking mobility of the users. They also compare the two-stage classi-

fication approach with a single multi-class approach and observe that the two-stage classifiers perform better in accuracy.

Schneider et al. (2013) study daily mobility patterns by analyzing the temporal and spatial trajectories as individual networks. In each dataset, ubiquitous daily mobility patterns, which are statistically reproduced by an analytical model, are observed. They argue that daily trips of humans follow a universal law. From the data, first they take out any additional information about the purpose of the activity, the travel time and the activity duration as well as the distances between the visited locations; thus, only the trip direction is examined. Afterwards, the distribution of the number of different visited locations and the number of different daily networks are investigated. Then, they basically examine the number of feasible daily networks that satisfies the need for sleep and the consistency of trips constraints. The first constraint implies that the trips start and end in the same location, which is called home. The consistency implies that each of the locations is visited at least once. These small feasible networks are called subnetworks. The subnetworks that occur more often in randomized versions of the entire network are called motifs. *Motifs* are basically the most common patterns that individuals follow in the data. For example, home-bakeryX - work - restaurantY - home is a motif. Even though data that is used in this study consists of different cities from different countries, the frequencies of observing a specific motif are similar (Schneider et al. (2013)).

Sobolevsky et al. (2014) study bank card transaction data of BBVA, the second largest Spanish bank, to examine how the mobility patterns vary depending on nationality. They first divide the customers into two groups, which are Spaniard and non-Spaniard. Then, they use a particular optimization algorithm for partitioning, based on an iterative combination of splits, merges and multi-node shifts from one community to another. Afterwards, they create the network of the customers by creating links between every pair of customers. For every transaction, an edge is created in between the current location of the customer and the executed location of the transaction. Also, the weight of the transaction depends on the amount of

the money transacted. They build a country-wide network of money flow by using this approach. They perform this network creation approach for both Spaniard and non-Spaniard customer groups and analyze both networks and observe that there are some differences between the two groups. Next, they model the mobility networks of the two groups by using a new link that they create by gravity law approximation. The *gravity law approximation* is a measure combining the weights and the distance between two locations and results in an outcome number. They conclude that Spaniards are more local, but foreign visitors travel longer distances within the country. Spaniards spend their time mostly in their primary local areas. On the other hand, non-Spaniards are attracted by the major destinations in the country and show more tendency to travel and explore. The final conclusion that they give is that people from nearby countries show more local mobility patterns, but people from distant countries are the opposite.

Krumme et al. (2012) study the predictability of customers' visitation patterns given their economic transaction data. In doing so, they calculate the *temporally – uncorrelated entropy*, which is the entropy of visitation of customers to the stores neglecting specific ordering of visits, and the *sequence – dependent entropy*, which is the compressibility of the sequence of stores. They observe that both entropy distributions are close to each other. They believe that this result is affected by the temporal resolution they select in the calculation of sequence-dependent entropy. From this perspective, comparing the individuals are related to the selection of temporal resolution. As a result, they conclude that the sequence-dependent entropy decreases with low temporal resolution (Krumme et al. (2012)). All the studies covered in this section are briefly illustrated in Table 2.2, where the characteristics of articles are summarized.

Authors	Data Type	Mobility Feature	Purpose of the Study
Noulas et al. (2012)	Check-in Data	Rank Distance, Place Popularity, Geographic Distance	Predicting Next Location
Song et al. (2004)	Wi-Fi Data	Entropy	Predicting Next Location
Singh et al. (2013)	Socio-Mobile Data	Diversity, Loyalty	Classifying Spending Behavior
Gonzalez et al. (2008)	Mobile Phone Data	Radius of Gyration	Extracting Human Trajectory Pattern
Song et al. (2010)	Mobile Phone Data	Random Entropy, Temporal-Correlated Entropy, Actual Entropy	Extracting Regular and Random Human Mobility Relation
Kung et al. (2014)	Mobile Phone Data	Commute Distances	Extracting Commute Patterns
Singh et al. (2015)	Bank Transaction Data	Diversity, Loyalty, Regularity	Extracting Relation with Financial Well-being
Isaacman et al. (2010)	Mobile Phone Data	Daily Range, Median Daily Range	Demonstrating Different Mobility Patterns
Sohn et al. (2010)	Mobile Phone Data	Similarity Measure	Predicting Type of Mobilities
Schneider et al. (2013)	Mobile Phone Data	Correlation Between Motifs	Extracting Daily Mobility Patterns
Sobolevsky et al. (2014)	Bank Transaction Data	Link Between Two Locations	Extracting Mobility Patterns and Nationality Relation
Krumme et al. (2012)	Mobile Phone Data	Entropy	Predicting Customer Visitation Patterns

Table 2.2: Studies on Mobility Behavior

Chapter 3

DATA AND PREPROCESSING

In this section, we present the details of our dataset and explain the features we extract from the dataset. Subsequently, we address the data preprocessing methods we perform on the dataset for a better classifier performance. We end this section by discussing the data separation process we implement to observe the differences between different training/testing datasets.

3.1 Data

In this study, we analyze the one-year customer transaction data of one of the largest banks in Turkey. Customer name and unique identifier (such as the social security number, national ID number) fields are removed from the data before the analysis. Thus, we perform the analysis with an anonymized and de-identified dataset in this study. The dataset includes demographic, financial and behavioral information about the customers. The data covers all the bank transactions belonging to 100,000 customers that occurred in one year period, including 15 tables, 217 columns and 52,267,342 rows in total. The unique ID of the customers and transaction dates are implicitly included for every data point in the dataset. These 15 tables are as follows:

- ATM (Automatic Teller Machine) transaction table includes transactions such as withdrawal, deposit and money transfers with location data.
- Crosswise sales table represents the activity and ownership of bank products used by customers.

- Call center table includes all call center call information about the customers, including frequency and the number of the operations performed during the calls.
- Electronic funds transfer table includes all the transfers going out of the customer's account including transfer amounts.
- Remittance table includes the remittance transactions including amount and currency information.
- There are two campaign tables that include all the campaigns that the bank has offered to the customer, the offering channel and the customer responses.
- Credit card receipt table consists of credit card balance and last payment date.
- Credit card receipt payment table includes total credit card debt and the corresponding currencies.
- Credit card expenditure table contains all the credit card expenditures with corresponding amounts, locations, currencies and spending category.
- Credit card information table is comprised of the number of credit cards that the customer owns including other banks and also credit card limits of the customer, again including both our bank and other banks.
- Web banking table includes all transactions that the customer performs via the web platform, including the type of the transaction, currency, channel, amount and total number of transactions.
- Customer information table includes some demographic and financial information about the customers, such as age, gender, occupation, marital status, income, education, bank age and risk scores.

- Auto payment table includes the amount and the total number of auto payments that the customer uses.
- Branch table consists of the number of the visits that the customer has done to the branches, including location data and the total number of transactions.

3.2 Features

In order to analyze the relationship between product usage and customer's behaviors, we use several different types of measures. These measures are divided into three main categories: demographic, financial and mobility. The spatio-temporal mobility behavior of each customer is measured based on three features, which are generated with four variations based on different definition of bins.

Diversity is a measure of customer's shopping behavior that may vary over time and space, where we generate the bins as segments over space and time. In other words, spreading the expenses equitable to different bins means that the customer is diverse. Given the bin definition, the fraction of transactions that fall within bin j for user i is denoted as p_{ij} . The parameter p_{ij} is calculated for each user independently. Then, we calculate the spatial diversity of user i as the normalized entropy of all transactions counted in all N bins. The diversity formula is given below, where M denotes the number of non-empty bins:

$$D_i = \frac{-\sum_{j=1}^N p_{ij} \log p_{ij}}{\log M} \quad (3.1)$$

The values of D_i are between 0 and 1. Higher spatial diversity corresponds to numbers closer to 1. A customer with high spatial diversity performs her transactions almost equitably across different locations. The reason of normalizing the measurement by $\log M$ is to keep the focus on quantifying the relative spread across bins and to give equal chance to customers with different number of transactions to score high in the diversity measure.

Loyalty refers to the percentage of transactions that occur over a customer's k most frequent bins. With the same definition of bins, we use k as 5 in this study for each user. Let f_i be the combined fraction of all transactions of user i that occur in top 5 most frequent bins of user i . The loyalty for each customer is calculated as follows:

$$L_i = \frac{f_i}{\sum_{j=1}^N p_{ij}} \quad (3.2)$$

The values of the loyalty measure change between 0 and 1. A large loyalty value indicates a high spatial loyal behavior. For example, a user with large spatial loyalty spends most of her transactions within top three of her visited locations.

Regularity measures the similarity level in customer's behavior over term periods (monthly, weekly, etc.). Since we have to calculate regularity measures for different monthly intervals described in Chapter 3.3, the compared regularity periods change for each interval for each dataset. We calculate 11 month (months 1-11) and 9 month (months 3-11) data tables' regularity measures choosing 3 month (months 9-11) transactions as the compared time period. The 6 month data table's comparison period is taken as 2 months (months 10-11) and finally the 3 month (months 9-11) data table's comparison time period is taken as 1 month (11th month). For instance, 9 month regularity measure is calculated by comparing the similarity of the same purchase patterns in 3 month transaction data. These assignments are illustrated in Table 3.1. We simply use SD_i and SL_i vectors and calculate the normalized Euclidean distances between them. As a consequence, the calculation of temporal regularity becomes as follows:

$$R_i = 1 - \frac{\sqrt{(D_i^1 - D_i^T)^2 + (L_i^1 - L_i^T)^2}}{\sqrt{2}} \quad (3.3)$$

Data Type	Comparison Periods
11 Months	3 Months
9 Months	3 Months
6 Months	2 Months
3 Months	1 Month

Table 3.1: Regularity Measure Comparison Periods

In Eq.3.3, D_i^1 and D_i^T represent the diversity of the user in the first month and in the entire period, respectively. Likewise, L_i^1 and L_i^T refer to the spatial loyalty of the user in the same time periods. The regularity measure, R_i , results between 0 and 1, where 1 corresponds to perfect regularity and 0 to imperfect regularity. For example, if the two vectors D_i and L_i are identical, then the first month's measures and overall measures will result in the same way, which will lead to a perfect regularity with 1. In other words, if a customer spends her money equitably in her top 3 most frequent places both in the reference month and in the month to be calculated month, then the regularity measure will be close to 1. On the other hand, if she is both loyal and not diverse in reference month but not loyal and diverse in the month to be calculated month, then the regularity measure will be close to zero because there is no regularity relation between those two periods in terms of diversity and loyalty.

The bins are created according to two interpretations for space and two for time:

Temporal-Hourly: The time period is taken as 8 hourly. For example, 12am-8am is a bin and 8am-16pm is another bin. Features X40, X41 and X42 are calculated by using temporal 8 hourly time period bins. For instance the hourly_diversity takes all the 8 hourly bins (j) for each user (i) that have transactions in it, and summed them according to the Eq. 3.1. So as the number of transactions are spread equitably among all the bins, then this value ends up close to 1, on the other hand, if the transactions are gathered on fewer bins, then the value is closer to 0. As a result, as the user have his transactions more balanced in terms of these 8 hourly bins, then her hourly_diversity value would be higher.

Temporal-Weekly: The time period is taken as weekly. For example, 33th week of the year is a bin and 34th week of the year is another bin. Features X43, X44 and X45 are calculated with temporal weekly time period bins. For instance, weekly_loyalty measure is calculated according to Eq. 3.2 for each user (i) and for top 5 most frequent weekly bins. Thus, if the user has most of the transactions in her top 5 most frequent bins among her all transactions, then the customer is accepted as loyal in terms of weekly bins.

Spatial-Radial: A radial structure is implemented where bins are concentric circles that have 1, 3, 5, 15, 50, 150, and 500 kilometer radii. The center of the radial structure is taken as the average of the transaction coordinates for each customer. Using the average of the transaction coordinates provides strong information about the customer's most frequent location. Therefore, spatial radial measure specifies how far the customer travels from her most frequent position. X46, X47 and X48 are calculated with spatial-radial period bins. For instance, for spatial_radial_diversity the number of transactions are spread equitably among all the spatial-radial bins, then the spatial_radial_diversity value ends up close to 1, on the other hand, if the transactions are gathered on fewer bins, then the value is closer to 0. As a result, as the user have her transactions more balanced in terms of these spatial-radial bins, then her spatial_radial_diversity value would be higher.

Spatial-Cluster: A clustering structure is used and transactions are split to 50 clusters according to the transaction coordinates across Turkey. For example, the center of Istanbul and the center of Ankara are cluster seeds because of the excessive number of transactions on those coordinates. X49, X50 and X51 are calculated with spatial-cluster period bins. For instance, spatial_cluster_loyalty is calculated according to Eq. 3.2 for each user (i) and for top 5 most frequent spatial-cluster bins. Thus, if the user has most of the transactions in her top 5 most frequent bins among her all transactions, then the customer is accepted as loyal in terms of spatial-cluster bins.

Feature Number	Feature Name	Data Type	Feature Type
X1	Customer ID	Integer	Demographic
X2	Age	Double Precision	Demographic
X3	Education Status	Text	Demographic
X4	Gender	Text	Demographic
X5	Marital Status	Text	Demographic
X6	Job	Text	Demographic
X7	Income	Double Precision	Financial
X8	Bank Age	Double Precision	Financial
X9	Average of Risk Scores	Integer	Financial
X10	Variance In Risk Scores	Double Precision	Financial
X11	Last Risk Score	Integer	Financial
X12	Number of Active Products	Integer	Financial
X13	Total Number of Call Center Calls	Integer	Financial
X14	Number of Call Center Calls Last Month	Integer	Financial
X15	Number of Total Credit Cards	Integer	Financial
X16	Number of Other Bank's Products	Integer	Financial
X17	Bank Credit Card Limit	Integer	Financial
X18	Total Credit Card Limit	Integer	Financial
X19	Average Interval Between Transactions	Double Precision	Financial
X20	Variance of Intervals Between Transactions	Double Precision	Financial
X21	Average Balance of the Account	Double Precision	Financial

Feature Number	Feature Name	Data Type	Feature Type
X22	Last Known Balance of the Account	Double Precision	Financial
X23	Average Amount of Auto Payments	Double Precision	Financial
X24	Number of Auto Payments	Integer	Financial
X25	Days Remaining to Last Payment Date	Integer	Financial
X26-X34	Average Expenses	Double Precision	Financial
X35	Variance of Expenses	Double Precision	Financial
X36	Monthly Total Expenses	Double Precision	Financial
X37	Mobile Banking Usage	Integer	Financial
X38	Total Number of Credit Card Transactions	Integer	Financial
X39	Total Number of ATM transactions	Integer	Financial
X40	hourly_diversity	Double Precision	Mobility
X41	hourly_loyalty	Double Precision	Mobility
X42	hourly_regularity	Double Precision	Mobility
X43	weekly_diversity	Double Precision	Mobility
X44	weekly_loyalty	Double Precision	Mobility
X45	weekly_regularity	Double Precision	Mobility
X46	spatial_radial_diversity	Double Precision	Mobility
X47	spatial_radial_loyalty	Double Precision	Mobility
X48	spatial_radial_regularity	Double Precision	Mobility
X49	spatial_cluster_diversity	Double Precision	Mobility
X50	spatial_cluster_loyalty	Double Precision	Mobility

Feature Number	Feature Name	Data Type	Feature Type
X51	spatial_cluster_regularity	Double Precision	Mobility
X52	Home-Work Distance	Double Precision	Mobility
X53	Max Average Monthly Distance	Double Precision	Mobility
X54	Max Variance Monthly Distance	Double Precision	Mobility
X55	hour+d+l+r	Double Precision	Mobility
X56	hour+d+l-r	Double Precision	Mobility
X57	hour+d-l-r	Double Precision	Mobility
X58	hour+d-l+r	Double Precision	Mobility
X59	week+d+l+r	Double Precision	Mobility
X60	week+d+l-r	Double Precision	Mobility
X61	week+d-l-r	Double Precision	Mobility
X62	week+d-l+r	Double Precision	Mobility
X63	radial+d+l+r	Double Precision	Mobility
X64	radial+d+l-r	Double Precision	Mobility
X65	radial+d-l-r	Double Precision	Mobility
X66	radial+d-l+r	Double Precision	Mobility
X67	cluster+d+l+r	Double Precision	Mobility
X68	cluster+d+l-r	Double Precision	Mobility
X69	cluster+d-l-r	Double Precision	Mobility
X70	cluster+d-l+r	Double Precision	Mobility

Feature Number	Feature Name	Data Type	Feature Type
Y	Specific Product Activity	Integer	Financial

Table 3.2: Feature Explanations

Table 3.2, which is constructed for the 9 month data period (9DFM) including months 3 to 11, contains all the features we extract from the bank dataset. The first feature, X1, is the ID of the customer, which is not used as a feature during the prediction process. X2 is the age of the customer. X3 is the education status of the customer (e.g., high school, university). X4 is the gender of the customer. X5 is the martial status of the customer (e.g., single, married). X6 is related to the job of the customer (e.g., private, public). All the features so far contain demographic information about the customers and from now on till X39, the features contain financial information. X7 is the monthly income of the customer. X8 is the duration that the customer uses the bank in terms of years. X9 is the variance of the risk scores of the customer through the data period. X10 is the mean of the risk scores of the customer through the data period. X11 is the last risk score of the customer in the data period, which corresponds to the 11th month in our data. X12 is the number of products that the user actively uses. In the dataset, there are two different concepts about product usage; product activity and product ownership. The activity definition varies according to the product. For example, for retirement savings product, the activity status is directly equal to the ownership status. This is because when the customer stops using this product, automatically she also loses the ownership status. On the other hand, for some products like time deposits, an inactivity of consecutive 3 months turns active status to inactive. In our dataset, when we examine the activity and ownership columns in detail, we observe that, they mostly represent the same values with a minor variation (in 98% of the cases the activity and ownership values are the same, for the rest, the activity values is 1 or 2 less than the ownership values). Thus, we eliminate the ownership column and continue with directly activity values.

X13 is the total number of call center calls the customer made. X14 is the total number of call center calls the customer made in the last month of the period, which corresponds to the 11th month. X15 is the customer's total number of credit cards including other banks. X16 is the number of products that the customer uses from other banks. X17 is the credit card limit of the customer in our bank. X18 is the total credit card limit of the customer including other banks' credit card limits. X19 is the average time between two consecutive transactions of the customer. X20 is the variance of these consecutive transaction intervals. X21 is the average cash in the customer's account during the time period. X22 is the amount of cash in the 11th month. X23 is the average amount that the customer pays via auto payments. X24 is the total number of auto payments that the customer uses. X25 is the due date of the last payment of credit card date. X26-X34 are the average monthly expenses during the periods. There are 9 monthly expenses in this data, because the table is constructed for a 9 month dataset. For instance, X26 and X31 indicate the average monthly expenses in 1st and 6th months, respectively. X35 is the variance of these monthly expenses. X36 is the sum of these monthly expenses. X37 indicates the total number of mobile banking usage of the customer. X38 is the total number of credit card transactions the customer performed during the period. X39 is the total number of ATM transactions the customer performed during the period.

We have mentioned all demographic and financial features so far and the remaining features are mobility features except the class label. X40, X41 and X42 represent customer's diversity, loyalty and regularity measures calculated according to the hourly bins of transactions. X43, X44 and X45 represent customer's diversity, loyalty and regularity measures calculated according to the weekly bins of transactions. X46, X47 and X48 represent customer's diversity, loyalty and regularity measures calculated according to the spatial-radial bins of transactions. X49, X50 and X51 represent customer's diversity, loyalty and regularity measures calculated according to the spatial-cluster bins of transitions. X52 is the Euclidean distance between the customer's home and work locations. X53 is the average of monthly maximum distances

that the user has taken during each month. X54 is the variance of these monthly maximum distances. X55-X70 values are the addition-subtraction combinations of all the diversity, loyalty and regularity values (X40-X51). For instance, hourly+d+l-r represents the summation of positive hourly diversity, positive hourly loyalty and negative hourly regularity. The reason that we have created the summation combinations of these variables is; we believe that each combination represents a specific type of mobility life style. For instance, as the week+d+l+r variable is greater, this means that the customer is spending most of her transactions equitably among her transaction locations, she is also loyal to her most frequent transaction places and she follows this weekly pattern over and over again. Thus, this is a sign of a standard life. Another example is low cluster+d-l-r. This means that the customer spends his transactions unequally among transaction points but she is loyal to her most frequent places and she follows this kind of pattern regularly. This type is also another sign of a customer with a standard life cycle. The last example can be low radial+d-l-r. This variable indicates that the user is again having her transactions unequal to the radial bins but she is loyal to the most frequent bins and follows the same regular pattern during his life cycle. Therefore, this is another sign of a standard life cycle. On the other hand, all the opposite values for these variables, like higher radial+d-l-r, etc., describes a non-standard and a more random life cycles. To sum up, each combination refers to a specific type of a life cycle. Finally, Y is the output variable (class label) of our study. Y indicates the active usage of a specific product (such as individual consumer loan) in the 12th month for each customer. All the elements in the feature set are used to predict the class label except for the ID column.

3.2.1 Descriptive Statistics

In this section, we summarize all the 90 features in Table. 3.3, with corresponding mean, standard deviation (Std Dev), minimum (Min), maximum (Max) and % percentiles.

Measure	Value	Measure	Value	Measure	Value
Age		Total Monthly Expense		radial+d-l+r	
Count	10,000	Count	10,000	Count	10,000
Mean	37.44	Mean	1,085.49	Mean	0.48
Std Dev	9.33	Std Dev	2,568.85	Std Dev	0.23
Min	19	Min	0.01	Min	0
25%	31	25%	180	25%	0.36
50%	36	50%	431.31	50%	0.52
75%	43	75%	1,018.80	75%	0.63
Max	83	Max	85,931.37	Max	1.32
Income		Mobile Bank Usage		cluster+d+l+r	
Count	10,000	Count	10,000	Count	10,000
Mean	3,847.53	Mean	145.20	Mean	2.05
Std Dev	7,246.77	Std Dev	696.01	Std Dev	0.74
Min	502	Min	0	Min	0
25%	1,500	25%	11	25%	2.20
50%	2,385	50%	74	50%	2.31
75%	4,000	75%	190	75%	2.38
Max	250,000	Max	65,319	Max	2.78

Measure	Value	Measure	Value	Measure	Value
Bank Age		Total Number of Credit Card Transactions		cluster+d+l-r	
Count	10,000	Count	10,000	Count	10,000
Mean	8.24	Mean	58.37	Mean	0.57
Std Dev	4.69	Std Dev	69.88	Std Dev	0.33
Min	0	Min	1	Min	-0.28
25%	4	25%	21	25%	0.36
50%	8	50%	41	50%	0.60
75%	11	75%	71	75%	0.80
Max	33	Max	2,469	Max	1.44
Average Risk Scores		Total ATM Transactions		cluster+d-l-r	
Count	10,000	Count	10,000	Count	10,000
Mean	3.01	Mean	48.70	Mean	-1.04
Std Dev	0.66	Std Dev	39.68	Std Dev	0.46
Min	0	Min	0	Min	-2
25%	2.91	25%	23	25%	-1.34
50%	3	50%	41	50%	-1.13
75%	3.36	75%	64	75%	-0.88
Max	5	Max	615	Max	0.02

Measure	Value	Measure	Value	Measure	Value
Variance Scores	Risk	hourly_diversity		cluster+d-l+r	
Count	10,000	Count	10,000	Count	10,000
Mean	0.20	Mean	0.79	Mean	0.44
Std Dev	0.27	Std Dev	0.27	Std Dev	0.20
Min	0	Min	0	Min	0
25%	0	25%	0.84	25%	0.40
50%	0.09	50%	0.88	50%	0.49
75%	0.27	75%	0.91	75%	0.56
Max	4.36	Max	0.96	Max	1.26
Last Risk Score		hourly_loyalty		Education Unknown	
Count	10,000	Count	10,000	Count	10,000
Mean	2.99	Mean	0.38	Mean	0.09
Std Dev	0.83	Std Dev	0.16	Std Dev	0.01
Min	0	Min	0	Min	0
25%	3	25%	0.34	25%	0
50%	3	50%	0.40	50%	0
75%	3	75%	0.47	75%	0
Max	5	Max	0.89	Max	1

Measure	Value	Measure	Value	Measure	Value
Number of Active Product		hourly_regularity		PhD	
Count	10,000	Count	10,000	Count	10,000
Mean	6.30	Mean	0.80	Mean	0
Std Dev	2.26	Std Dev	0.30	Std Dev	0.05
Min	0	Min	0	Min	0
25%	5	25%	0.84	25%	0
50%	6	50%	0.92	50%	0
75%	8	75%	0.95	75%	0
Max	18	Max	1	Max	1
Total Number of Call Center Calls		weekly_diversity		Uneducated	
Count	10,000	Count	10,000	Count	10,000
Mean	83.54	Mean	0.81	Mean	0.01
Std Dev	158.75	Std Dev	0.27	Std Dev	0.08
Min	0	Min	0	Min	0
25%	19	25%	0.88	25%	0
50%	50	50%	0.91	50%	0
75%	104	75%	0.92	75%	0
Max	591	Max	0.94	Max	1

Measure	Value	Measure	Value	Measure	Value
Last Month Call Center Calls		weekly_loyalty		Primary School	
Count	10,000	Count	10,000	Count	10,000
Mean	7.72	Mean	0.51	Mean	0.05
Std Dev	24.19	Std Dev	0.18	Std Dev	0.22
Min	0	Min	0	Min	0
25%	0	25%	0.51	25%	0
50%	0	50%	0.55	50%	0
75%	5	75%	0.59	75%	0
Max	98	Max	0.90	Max	1
Number of Total Credit Cards		weekly_regularity		Master's Degree	
Count	10,000	Count	10,000	Count	10,000
Mean	2.40	Mean	0.81	Mean	0.04
Std Dev	1.54	Std Dev	0.30	Std Dev	0.19
Min	0	Min	0	Min	0
25%	1	25%	0.85	25%	0
50%	2	50%	0.92	50%	0
75%	3	75%	0.96	75%	0
Max	15	Max	1	Max	1

Measure	Value	Measure	Value	Measure	Value
Number of Other Bank's Products		spatial_radial_diversity		High School	
Count	10,000	Count	10,000	Count	10,000
Mean	1.44	Mean	0.39	Mean	0.46
Std Dev	1.53	Std Dev	0.21	Std Dev	0.50
Min	0	Min	0	Min	0
25%	0	25%	0.24	25%	0
50%	1	50%	0.43	50%	0
75%	2	75%	0.56	75%	1
Max	14	Max	0.88	Max	1
Bank Credit Card Limit		spatial_radial_loyalty		Secondary School	
Count	10,000	Count	10,000	Count	10,000
Mean	8,209.71	Mean	0.87	Mean	0.06
Std Dev	9,616.58	Std Dev	0.29	Std Dev	0.24
Min	0	Min	0	Min	0
25%	2,000	25%	0.94	25%	0
50%	5,000	50%	0.98	50%	0
75%	10,600	75%	1	75%	0
Max	150,000	Max	1	Max	1

Measure	Value	Measure	Value	Measure	Value
Total Credit Card Limit		spatial_radial_regularity		University Degree	
Count	10,000	Count	10,000	Count	10,000
Mean	19,670.27	Mean	0.77	Mean	0.29
Std Dev	30,651.14	Std Dev	0.29	Std Dev	0.45
Min	0	Min	0	Min	0
25%	4,150	25%	0.77	25%	0
50%	9,450	50%	0.89	50%	0
75%	22,800	75%	0.95	75%	1
Max	958,750	Max	1	Max	1
Average Interval Between Transactions		spatial_cluster_diversity		Male	
Count	10,000	Count	10,000	Count	10,000
Mean	8.89	Mean	0.51	Mean	0.71
Std Dev	8.65	Std Dev	0.23	Std Dev	0.46
Min	0	Min	0	Min	0
25%	3.73	25%	0.42	25%	0
50%	6.35	50%	0.57	50%	1
75%	10.86	75%	0.68	75%	1
Max	141.50	Max	0.93	Max	1

Measure	Value	Measure	Value	Measure	Value
Interval Between Transactions	Variance Transac-tions	spatial_cluster_loyalty		Female	
Count	10,000	Count	10,000	Count	10,000
Mean	1,665,680	Mean	0.75	Mean	0.29
Std Dev	1,955,880	Std Dev	0.27	Std Dev	0.46
Min	0	Min	0	Min	0
25%	593,915.40	25%	0.72	25%	0
50%	1,111,539	50%	0.84	50%	0
75%	2,019,977	75%	0.92	75%	1
Max	31,944,020	Max	1	Max	1
Average Balance of Account		spatial_cluster_regularity		Single	
Count	10,000	Count	10,000	Count	10,000
Mean	1,941.01	Mean	0.80	Mean	0.23
Std Dev	2,995.28	Std Dev	0.30	Std Dev	0.42
Min	0	Min	0	Min	0
25%	483.20	25%	0.86	25%	0
50%	1,000.29	50%	0.91	50%	0
75%	2,160.66	75%	0.95	75%	0
Max	57,409.89	Max	1	Max	1

Measure	Value	Measure	Value	Measure	Value
Last Balance of Account		Home-Work Distance		Marital Status Unknown	
Count	10,000	Count	10,000	Count	10,000
Mean	1,498.69	Mean	27.53	Mean	0.03
Std Dev	2,623.34	Std Dev	112.26	Std Dev	0.17
Min	0	Min	0	Min	0
25%	256.88	25%	2.08	25%	0
50%	726.91	50%	6.02	50%	0
75%	1,673.60	75%	12.95	75%	0
Max	54,876.37	Max	1,377.81	Max	1
Average Amount of Auto Payments		Max Average Monthly Distance		Divorced	
Count	10,000	Count	10,000	Count	10,000
Mean	9.93	Mean	30.46	Mean	0.04
Std Dev	35.35	Std Dev	84.54	Std Dev	0.20
Min	0	Min	0	Min	0
25%	0	25%	0	25%	0
50%	0	50%	0	50%	0
75%	0	75%	18.22	75%	0
Max	969.92	Max	1,273.69	Max	1

Measure	Value	Measure	Value	Measure	Value
Number of Auto Payments		Max Variance Monthly Distance		Widow	
Count	10,000	Count	10,000	Count	10,000
Mean	3.02	Mean	1,341.48	Mean	0
Std Dev	11.29	Std Dev	7,312.47	Std Dev	0.07
Min	0	Min	0	Min	0
25%	0	25%	0	25%	0
50%	0	50%	0	50%	0
75%	0	75%	1.79	75%	0
Max	175	Max	220,884.01	Max	1
Days Remaining to Last Payment Date		hour+d+l+r		Married	
Count	10,000	Count	10,000	Count	10,000
Mean	13.74	Mean	1.95	Mean	0.69
Std Dev	7.26	Std Dev	0.71	Std Dev	0.46
Min	0	Min	0	Min	0
25%	7.67	25%	2.12	25%	0
50%	12.33	50%	2.20	50%	1
75%	18.83	75%	2.26	75%	1
Max	29	Max	2.56	Max	1

Measure	Value	Measure	Value	Measure	Value
Month 1 Average Expense		hour+d+l-r		Marriage Cancellation	
Count	10,000	Count	10,000	Count	10,000
Mean	159.05	Mean	1.20	Mean	1
Std Dev	533.57	Std Dev	0.46	Std Dev	0.01
Min	0	Min	0	Min	0
25%	0	25%	1.20	25%	0
50%	59.80	50%	1.36	50%	0
75%	135.03	75%	1.46	75%	0
Max	20,500	Max	1.70	Max	1
Month 2 Average Expense		hour+d-l-r		Job Unkown	
Count	10,000	Count	10,000	Count	10,000
Mean	174.27	Mean	-0.40	Mean	0
Std Dev	616.71	Std Dev	0.22	Std Dev	0.05
Min	0	Min	-1.39	Min	0
25%	0	25%	-0.53	25%	0
50%	57.45	50%	-0.42	50%	0
75%	135.91	75%	-0.31	75%	0
Max	20,000	Max	0.48	Max	1

Measure	Value	Measure	Value	Measure	Value
Month 3 Average Expense		hour+d-l+r		Retired	
Count	10,000	Count	10,000	Count	10,000
Mean	191.73	Mean	0.35	Mean	0.04
Std Dev	675	Std Dev	0.16	Std Dev	0.21
Min	0	Min	0	Min	0
25%	7.89	25%	0.31	25%	0
50%	63.45	50%	0.37	50%	0
75%	147.18	75%	0.43	75%	0
Max	20,800	Max	1.14	Max	1
Month 4 Average Expense		week+d+l+r		Housewife	
Count	10,000	Count	10,000	Count	10,000
Mean	171.94	Mean	2.10	Mean	0.01
Std Dev	593.05	Std Dev	0.76	Std Dev	0.07
Min	0	Min	0	Min	0
25%	15	25%	2.31	25%	0
50%	67.38	50%	2.38	50%	0
75%	146.26	75%	2.43	75%	0
Max	20,000	Max	2.58	Max	1

Measure	Value	Measure	Value	Measure	Value
Month 5 Average Expense		week+d+l-r		Self Employment	
Count	10,000	Count	10,000	Count	10,000
Mean	172.71	Mean	1.11	Mean	0.10
Std Dev	571.13	Std Dev	0.41	Std Dev	0.30
Min	0	Min	0	Min	0
25%	16.58	25%	1.15	25%	0
50%	65.58	50%	1.26	50%	0
75%	144.56	75%	1.33	75%	0
Max	20,700	Max	1.46	Max	1
Month 6 Average Expense		week+d-l-r		Working Retired	
Count	10,000	Count	10,000	Count	10,000
Mean	181.92	Mean	-0.50	Mean	0.02
Std Dev	574.39	Std Dev	0.21	Std Dev	0.14
Min	0	Min	-1.19	Min	0
25%	21.83	25%	-0.62	25%	0
50%	68.34	50%	-0.55	50%	0
75%	151.86	75%	-0.47	75%	0
Max	22,112	Max	0.20	Max	1

Measure	Value	Measure	Value	Measure	Value
Month 7 Average Expense		week+d-l+r		Not Working	
Count	10,000	Count	10,000	Count	10,000
Mean	170.87	Mean	0.49	Mean	0.01
Std Dev	561.41	Std Dev	0.20	Std Dev	0.07
Min	0	Min	0	Min	0
25%	21.52	25%	0.48	25%	0
50%	67.09	50%	0.53	50%	0
75%	150	75%	0.58	75%	0
Max	20,000	Max	1.20	Max	1
Month 8 Average Expense		radial+d+l+r		Student	
Count	10,000	Count	10,000	Count	10,000
Mean	166.50	Mean	2.02	Mean	0
Std Dev	541.96	Std Dev	0.75	Std Dev	0.02
Min	0	Min	0	Min	0
25%	18.60	25%	2.07	25%	0
50%	64	50%	2.27	50%	0
75%	139	75%	2.42	75%	0
Max	20,000	Max	2.77	Max	1

Measure	Value	Measure	Value	Measure	Value
Month 9 Average Expense		radial+d+l-r		Government Worker	
Count	10,000	Count	10,000	Count	10,000
Mean	186.15	Mean	0.30	Mean	0.06
Std Dev	639.68	Std Dev	0.25	Std Dev	0.24
Min	0	Min	-0.51	Min	0
25%	22.42	25%	0.09	25%	0
50%	67.26	50%	0.31	50%	0
75%	147.74	75%	0.49	75%	0
Max	18,500	Max	1.03	Max	1
Variance of Monthly Expenses		radial+d-l-r		Private Sector Worker	
Count	10,000	Count	10,000	Count	10,000
Mean	260,573.90	Mean	-1.25	Mean	0.75
Std Dev	2,100,472	Std Dev	0.49	Std Dev	0.43
Min	0	Min	-2	Min	0
25%	1,403.06	25%	-1.52	25%	0
50%	7,022.86	50%	-1.38	50%	1
75%	42,718.25	75%	-1.20	75%	1
Max	79,656,840	Max	0	Max	1

Table 3.3: Descriptive Statistics of All Features

3.3 Data Preprocessing

The very first data preprocessing step applied on the dataset is handling the null data points. Null data distorts both feature selection and prediction algorithms (Wang and Huang (2009)). There are various algorithms in the literature that replace the missing values (Martin-Fernandez et al. (2003)). Nevertheless, missing value replacement somehow perturbs the data. Missing value replacement approaches try to make strong estimations for the missing values, but this process should be performed accurately, because changing the data points directly affects the prediction results (Kim et al. (2005)). By relying upon the size of our data and the sparsity of null data points, we removed the records which had missing data elements, which corresponds to less than 1% of the data. After overcoming the missing value problem, in order to use the data as an input to the classification algorithms, we transform all the categorical features into numeric features. We implement this process in Python and transform all the categorical variables into multiple representative columns with binary elements. For instance, the gender column is transferred into male and female columns, where male customers have 1 in the male column and 0 in the female column. This process provides classification algorithms to figure out the data in numeric format and increases the number of the features up to 90.

As a next step, a detailed filtering approach is performed on customers to obtain more desirable customers. The data includes all transactions of 100,000 customers, with various ages, jobs, incomes, etc. As a matter of fact, some customers are very active with frequent transactions, and some are not. Therefore, we filter those 100,000 customers into a smaller subset, who might be more useful for the objectives of this study. We begin filtering the customers whose number of total credit card transactions is greater than 10 and bank POS (Point of Sales) rate is greater than 60%, meaning that 60% of the transactions belongs to our bank's POS machines. This query eliminates 40,000 out of 100,000 customers. On top of that we filter those with monthly income greater than 500 *TL* (Turkish Lira) and less than 300,000 *TL*. After eliminating the income outliers and inactive customers, we sum the number of credit

card transactions and number of ATM transactions including location data information for all customers and sort customers in descending order of this summation. Afterwards, we select the top 10,000 customers from this list as our desired set of customers using PostgreSQL. We perform this last filtering method because we want to select the most active customers due to their high amount of transactions, they will provide more information in terms of mobility and financial behavior. Since the objective of the study is to show how customer's mobility behavior affects product recommendation decisions, filtering the customers in this manner will aid the study in line with our objective.

3.4 Dataset Separation

3.4.1 Time Interval and Feature Type

At this point, null data points are gone and passive customers are removed. From now on, the aim is to gather all required features in one table with desired customers and time filters for each table. In other words, 15 different data tables should be merged in a way that satisfies all desired characteristics and supports the prediction algorithm through its features. To better understand the relation of mobility behavior and product usage, we basically separate the datasets into 4 different datasets by choosing monthly intervals between months 9-11 (3 month data), months 6-11 (6 month data), months 3-11 (9 month data) and months 1-11 (11 month data). We cannot take the month interval as 12 months, because we use the 12th month data as our output variable. We also have to divide the data into four sets: data with only demographic features, demographic-mobility features, demographic-financial features and demographic-mobility-financial features. Taking into account that we also have 4 monthly intervals except demographic dataset, the study requires 13 different datasets illustrated in the Table 3.4.

Dataset	Transactions Between Months	Features
D	Does not Depend on Month	Demographic
11DM	1-11	Demographic & Mobility
11DF	1-11	Demographic & Financial
11DFM	1-11	Demographic & Mobility & Financial
9DM	3-11	Demographic & Mobility
9DF	3-11	Demographic & Financial
9DFM	3-11	Demographic & Mobility & Financial
6DM	6-11	Demographic & Mobility
6DF	6-11	Demographic & Financial
6DFM	6-11	Demographic & Mobility & Financial
3DM	9-11	Demographic & Mobility
3DF	9-11	Demographic & Financial
3DFM	9-11	Demographic & Mobility & Financial

Table 3.4: Dataset Types Explanations

All the features described in Sec.3.2 are calculated independently from each other because of the monthly intervals. For example, the number of active products column for 3DF table considers activity in months 9, 10 and 11; the number of active products column for 6DFM considers activity in months 6, 7, 8, 9, 10 and 11. After completing 13 different tables with the corresponding features, we normalize all the features in their own to reduce data redundancy and improve data integrity (Sevilla and Sola (1997)). We implement max-min normalization (mentioned in Wang and Huang (2009)) using Python as shown in Eq.3.4. Here, $\text{Min}(X)$ and $\text{Max}(X)$ refer to the minimum and maximum values of the corresponding feature set (Column X). Here, x is the value in the feature set to be normalized by the minimum and maximum values and N is the normalized version of x . Max-min normalization scales all the variables between 0 and 1, where 0 (1) corresponds to the minimum (maximum) element in the feature set. Eventually, the data is ready for Feature Selection/Extraction (FSE)

and classification methods.

$$N = \frac{x - \text{Min}(X)}{\text{Max}(X) - \text{Min}(X)} \quad (3.4)$$

3.4.2 Separation of Customers

We believe that predicting the activeness status change of the products for a customer, i.e. changing from active to passive or passive to active, is challenging in comparison to prediction of active to active or passive to passive customers. Therefore, we generate three different training/testing scenarios for FSE and prediction processes. The first scenario contains customers who did not use the product actively for the last 3 months. We use only these customers in both FSE, training/testing processes and call this scenario **Class-Zero**. The second scenario contains two groups, a group of Class-Zero customers where the last 3 months' activity is 0, and a group of mixed customers with the last 3 months' activity being either 0, 1, 2 or 3. For this scenario, FSE and prediction model training are performed with mixed customers, but the model is tested with only Zero customers. We call this scenario **Class-Ones**. The last scenario is comprised of mixed customers and all methods are performed on mixed customer datasets, called **Class-All**. To sum up, we have 4 types of monthly intervals, 3 types of mobility behaviors except single demographic, and 3 types of scenarios. This combination results in 36 different datasets, and then we have to add the single demographic datasets (D) to each scenario which are independent of monthly intervals. As a result, we have a total of 39 different scenarios to test for each of the two product types.

We select the two products; namely, individual consumer loan (ICL) and retirement savings (RS) products. ICL refers to the product such that the customer borrows a sum of money from the bank and then, pays it back in installments, including an interest rate during a time period. For example, a customer could borrow 10,000 today and pay 1,000 during the next 11 months to the bank. The amounts are calculated according to the interest rates during the periods. RS product involves the payment

of a particular amount of money to the bank monthly during working years until the retirement date. After the retirement, the bank will pay a particular money to the customer either as a one lump sum amount or in monthly installments. This product aims for the customer to save some money for retirement years.



Chapter 4

FSE METHODS AND CLASSIFIERS

In this section, we first clarify how we evaluate the performances of the FSE methods, how the number of features and classification algorithms are determined on each of our dataset independently. Afterwards, we briefly explain all of the FSE methods, the number of features and the classifiers we implement. The most basic definition of classification is pattern recognition. Classification algorithms basically use supervised learning techniques to learn the patterns in the data via training sets, where the output variable is known, and then use the extracted pattern on the testing datasets to predict its output variable. In our study, the classification algorithms classify the customers into two depending on whether a customer will use the product in the 12th month, or not.

4.1 Performance Evaluation

We follow a simple logic in evaluating the performances of the FSE methods: the better the classifier performs, the better the FSE is (Chandrashekar and Sahin (2014)). This argument is reasonable since the information transferred to the classification algorithm is determined by the FSE method. If the FSE method chooses the best features to explain the class label, then the classifier can perform better using this information. Therefore, we use the performance of the classification algorithms to interpret the performance of different FSE methods. There are five common performance evaluation metrics in big data analysis: accuracy, receiver operating characteristic area under the curve (AUC), precision, recall and F1.

Before explaining the details of the evaluation metrics, first some common terms should be defined. A 2x2 matrix is used as a measurement in data analysis, called the

confusion matrix as shown in Fig.4.1. The first and the second rows of the confusion matrix indicate the actual values of the output feature in the testing dataset. The first row means that the customer does not use the product, the second row means that the customer uses the product in the 12th month. The first and the second columns indicate the predicted values of the output feature in the testing dataset, predicted as not using and using, respectively. Four main terms are used to interpret the confusion matrix: true negative (TN), false positive (FP), false negative (FN) and true positive (TP). **TN** is the number of predictions of the customers who do not use the product correctly. **FP** is the number of the customers who do not use the product as if they use the product. **FN** is the number of the customers that use the product as if they do not use the product. **TP** is the number of the customers that use the product correctly. All the terms are shown in Fig.4.1.

		Predicted	
		0	1
Actual	0	TN	FP
	1	FN	TP

Figure 4.1: Confusion Matrix

Accuracy is the ratio of correctly predicted results to the sample size. The formula of accuracy is given in Eq.4.1.

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \quad (4.1)$$

Precision refers to correctness ratio of positive predictions. The formula of precision is given in Eq.4.2.

$$Precision = \frac{TP}{TP + FP} \quad (4.2)$$

Recall refers to how well the predictor performs on the customers who actually use the product. The formula of recall is given in Eq.4.3.

$$Recall = \frac{TP}{TP + FN} \quad (4.3)$$

F1 is derived from both precision and recall. The formula of F1 is given in Eq.4.4.

$$F1 = 2 \frac{Precision Recall}{Precision + Recall} \quad (4.4)$$

AUC is a simple two dimensional graph, where the X axis is False Positive Rate (FPR) and Y axis is True Positive Rate (TPR). **TPR** is calculated identical to recall measure, the ratio of TP over all predicted customers as one (TP+FP). **FPR** calculation is illustrated in Eq. 4.5 (Zweig and Campbell (1993)). Joining these two points in two dimensional space generates the AUC curve.

$$FPR = \frac{FP}{FP + TN} \quad (4.5)$$

Accuracy, precision, recall, F1 and AUC are the evaluation metrics that we use in all performance evaluation parts for various FSE methods, changes in the number of features of FSE methods, different classifiers and the classifier parameter selection. Since there are 5 metrics, we basically take the average of these 5 scores and determine the best performing methods according to this average score.

4.2 Feature Selection and Extraction Methods

The impact and importance of FSE are mentioned in the literature, Chandrashekar and Sahin (2014) state that choosing the best features will boost the classifier's performance. Therefore, achieving success in the prediction process requires choosing the best prediction algorithms using best FSE methods and considering the optimal number of features for each scenario independently. The reason of independence is that all datasets have different features with different values so the processes should be performed independently. For instance, applying FSE on the dataset with the 3 month interval including mobility features for Class-All may result differently from applying the same algorithm on the dataset with the 9 month interval without mobility features for Class-Zero. Hereby, we determine the success criteria as the performance of different prediction algorithms with selected features. We use 10 prediction algorithms, given in Sec.4.3 and compare their average performances for each FSE method. To sum up, our hypothesis is that, the better the classifiers perform, the better the FSE method is. The performance metrics of the prediction algorithms will be described in Sec.4.3 in detail.

We implement four FSE algorithms and evaluate the average performances of 10 prediction algorithms after using these FSE methods for each of the 24 cases independently. The FSE methods we implement are: `f_regression`, Recursive Feature Elimination (RFE) and principal component analysis (PCA). F-regression is a univariate linear regression test. The cross correlation between each feature and the class label is computed. Afterwards the score is converted to a score called F-Score and according to the input value of the number of features, top ranked features are selected. Another FSE method is RFE, which first selects all features in the dataset and recursively eliminates them into smaller and smaller sets of features according to the weights that the algorithm assigns based on relevances to the class label. Smaller weights are eliminated recursively until the desired number of features is obtained. The last FSE method we implement is PCA, which is a common technique to find patterns in high dimensional data with small effort. PCA helps reducing the number

of dimensions in data while keeping the information as much as it can. In this way it can find hidden and simplified structures. It uses eigenvalue decomposition and reduces data dimensions to find the best features to explain as much of the variability in the data as possible (Shlens (2014)).

In order to determine the best FSE method for each scenario (78 in our study) shown in Tables 4.1 to 4.6, we take 5 runs (with small changes on prediction methods' parameters) using the 10-fold cross validation technique mentioned in Kohavi (1995) for each scenario. K-fold cross validation reduces data dependency of classifiers in data mining by creating "k" different training and testing subsets with the given dataset and taking the average of all the prediction results. We use "k" as 10, because k does not increase running times excessively and 10 random training/testing samples are sufficient to eliminate data dependency. The split ratio we use in the study for training and testing samples are 7000 and 3000 customers, respectively.

Afterwards, we analyze the performances of each FSE method on each scenario and we compare these values to obtain the best scenario-method assignments, shown in the first two columns of Tables 4.1 to 4.6. The tables are the results for the ICL product. For instance, for 9DFM dataset of Class-Ones, we compare the average classifier performances of all FSE methods and select the best FSE method for this particular scenario as RFE. Afterwards, we extract which number of features are selected, corresponding FSE methods that have the best performance, represented in the third columns of Tables 4.1 to 4.6. For instance, after obtaining that RFE performs best for 9DFM dataset of Class-Ones for ICL product, by using the same evaluation metrics described in Sec.4.3, we determine that the best performing number of features for this scenario is 20. At this point, we have the best performing FSE methods and the number of features for every scenario. Thus, the next step is to determine the best prediction algorithms using the results of the FSE process.

Individual Consumer Loan Product				
Class-Zero				
Dataset	Best Method	Best # of Features	Best Classifier	Best Parameter
D	PCA	13	Log	8
11DM	RFE	45	Log	1
11DF	RFE	30	Out	3
11DFM	f_regression	35	Out	4
9DM	RFE	30	Out	1.2
9DF	RFE	45	Log	1.2
9DFM	f_regression	7	Out	2.5
6DM	RFE	8	Log	1
6DF	PCA	13	Log	8
6DFM	RFE	45	Log	1.5
3DM	RFE	30	Out	1.2
3DF	RFE	10	Log	1
3DFM	RFE	25	Out	3

Table 4.1: ICL Class-Zero Optimum Assignments

Individual Consumer Loan Product				
Class-Ones				
Dataset	Best Method	Best # of Features	Best Classifier	Best Parameter
D	RFE	45	Rand	12
11DM	PCA	20	Rand	12
11DF	f_regression	18	Rand	20
11DFM	f_regression	30	ET	12
9DM	f_regression	40	ET	14
9DF	RFE	20	ET	20
9DFM	PCA	40	DT	9
6DM	RFE	25	Rand	12
6DF	RFE	20	ET	20
6DFM	PCA	35	DT	9
3DM	f_regression	30	ET	12
3DF	RFE	25	ET	20
3DFM	RFE	35	Rand	1.2

Table 4.2: ICL Class-Ones Optimum Assignments

Individual Consumer Loan Product				
Class-All				
Dataset	Best Method	Best # of Feature	Best Classifier	Best Parameter
D	f_regression	13	ET	18
11DM	f_regression	30	ET	9
11DF	f_regression	18	ET	20
11DFM	RFE	18	ET	18
9DM	f_regression	35	Rand	20
9DF	RFE	45	ET	20
9DFM	RFE	40	Rand	20
6DM	f_regression	30	Rand	25
6DF	f_regression	40	Rand	20
6DFM	RFE	18	ET	18
3DM	RFE	35	Rand	25
3DF	f_regression	40	Rand	25
3DFM	RFE	20	ET	18

Table 4.3: ICL Class-All Optimum Assignments

Retirement Savings Product				
Class-Zero				
Dataset	Best Method	Best # of Features	Best Classifier	Best Parameter
D	PCA	25	Log	2
11DM	f_regression	5	Out	1.5
11DF	f_regression	20	Log	4
11DFM	RFE	4	Log	2
9DM	RFE	7	OVO	3.5
9DF	RFE	35	Log	1
9DFM	RFE	13	OVO	2
6DM	f_regression	30	Out	4
6DF	f_regression	25	OVO	3
6DFM	RFE	20	Out	3
3DM	PCA	5	Log	1
3DF	f_regression	7	Log	1.2
3DFM	PCA	5	Out	4

Table 4.4: RS Class-Zero Optimum Assignments

Retirement Savings Product				
Class-Ones				
Dataset	Best Method	Best # of Features	Best Classifier	Best Parameter
D	PCA	30	Rand	16
11DM	PCA	40	ET	20
11DF	RFE	30	ET	14
11DFM	RFE	25	ET	2
9DM	RFE	25	Rand	14
9DF	RFE	40	ET	4
9DFM	PCA	45	Rand	20
6DM	f_regression	20	ET	20
6DF	PCA	25	Rand	14
6DFM	f_regression	30	ET	2
3DM	f_regression	35	ET	14
3DF	RFE	40	ET	4
3DFM	f_regression	40	Rand	20

Table 4.5: RS Class-Ones Optimum Assignments

Retirement Savings Product				
Class-All				
Dataset	Best Method	Best # of Features	Best Classifier	Best Parameter
D	RFE	9	Rand	20
11DM	RFE	25	ET	12
11DF	RFE	39	Rand	20
11DFM	f_regression	45	Rand	20
9DM	RFE	10	DT	8
9DF	RFE	45	ET	14
9DFM	f_regression	20	Rand	20
6DM	f_regression	30	ET	16
6DF	f_regression	20	Rand	20
6DFM	RFE	25	DT	8
3DM	RFE	35	ET	20
3DF	RFE	45	Rand	8
3DFM	f_regression	45	ET	14

Table 4.6: RS Class-All Optimum Assignments

4.3 Classification Algorithms

We implement 10 prediction algorithms with various characteristics and parameters to find the most appropriate ones for our study for each scenario. All the classifiers are implemented using libraries in [http : //scikit – learn.org](http://scikit-learn.org). The classifiers are listed below:

- Logistic regression classifier (Log)
- Decision tree classifier (DT)
- Random forest classifier (Rand)
- Extra trees classifier (ET)
- OneVsOne multi-class classifier (OVO)
- OneVsRest multi-class and multi-label classifier (OVR)
- Error correcting output code classifier (ECOC)
- Stochastic gradient descent classifier (SGD)
- Gaussian Naive Bayes classifier (GNB)
- Multi-layer perceptron neural network classifier (NN)

Logistic Regression Classifier is a statistical regression method to analyze datasets in which there are multiple independent variables related to an outcome (Ng and Jordan (2002)).

Decision Tree Classifier creates a model to predict the value of the output variable by learning simple decision rules inferred from the data features (Safavian and Landgrebe (1991)).

Random Forest Classifier fits a number of decision tree classifiers on various sub-samples of the dataset and takes the average of these tree results to improve prediction accuracy and prevent over-fitting (Pal (2005)).

Extra Trees Classifier follows the same procedure as random forest classifier but the number of decision trees are randomly chosen in the extra trees classifier (Desir et al. (2012)).

OneVsOne Classifier constructs one classifier for each pair of classes. In other words, it divides an m class problem into $m(m-1)/2$ binary problems. Each problem is modeled by a binary classifier, and using these classifiers the prediction is performed (Galar et al. (2011)).

OneVsRest Classifier divides an m class problem into m binary problems, instead of $m(m-1)/2$. Each problem is again modeled by a binary classifier and using these classifiers the prediction is performed (Galar et al. (2011)).

ECOC Classifier is a meta method that combines many binary classifiers to perform the prediction (Pujol et al. (2006)).

SGD Classifier is a gradient descent optimization method for minimizing an objective function which is comprised of the sum of differentiable functions (Bottou (2010)).

GNB Classifier is a supervised learning algorithm based on applying the Bayes' theorem between every pair of features (George and Langley (1995)).

MLP Neural Network Classifier is a feed forward neural network algorithm that maps sets of input data onto a set of appropriate outputs. MLP consists of multiple layers of nodes and each layer is connected to the next layer (Gardner and Dorling (1998)).

We follow the same approach as we do in Sec.4.2 to determine the best classifiers for each scenario and use the evaluation metrics detailed in Sec.4.1. We first fix the dataset-FSE method-number of features assignments that we obtain, and we implement all the classifiers listed above to each scenario. Afterwards, we analyze the results and determine the dataset-classifier assignments. Then, we basically fix all

the information we have and change the main parameter of the classifiers in Python using the iteration number as 10 for a loop. Eventually, we obtain the best performing dataset-FSE method-top 3 number of features-classifier-classifier parameter assignments.



Chapter 5

COMPUTATIONAL RESULTS

5.1 Individual Consumer Loan

In this section, we will examine the prediction results of ICL in detail and we will interpret the results in terms of demographic data, financial data and mobility behaviors.

5.1.1 Computational Results of ICL

After all the preprocessing and analysis we perform for ICL as described in Sec.4, eventually we obtain the results for all scenarios with corresponding optimum assignments. The average AUC, accuracy, precision, recall, F1 scores and a confusion matrix sample is shown in Tables 5.1, 5.2 and 5.3 for each scenario. The sample AUC graphs of ICL are shown in Figures 5.1 to 5.6. Afterwards, to evaluate the performance of the model on a completely different customer set and to prove that our model is not over fitting, we test our model on the next 3,000 customers which we have never trained our model with. Those results for ICL product are shown in Tables 5.4, 5.5 and 5.6. After comparing these results with the ones that we have before, we observe that our model is not over fitting and there is not a significant difference between the performances.

For Class-Zero, we train and test the model with customers who have not used ICL for the last 3 months. As it can be predicted, the percentage from these customers, who start using the product in 12th month is extremely low, only 2%. Thus, for this scenario, we are trying to accurately predict 2% of the customers. Achieving a high performance on this scenario is difficult due to having a heavily imbalanced dataset. Moreover, for Class-Zero model, since we train the model without the customers who

actually use the product during the time period of the dataset, we believe that the model could not identify the pattern of the customers using the product. This is why the model has difficulties to predict the customers who will use ICL in the 12th month. When we examine the confusion matrices in detail in Table 5.1, we can see that the model predicts almost 55% of customers who will not use ICL, and 56% of the customers (almost 0.8% of total customers) who will use the product. Consequently, the model still predicts more than half of each type of customers but unfortunately, it is not enough to claim it performs well. As a result, our model could not perform well on this scenario regardless of the dataset type. Besides, the AUC graph of Class-Zero with 6 months time period is shown in Fig. 5.1. It cannot be seen easily from that graph, but the performance ranking is $DFM \geq DF \geq DM \geq D$.

Individual Consumer Loan							
Class-Zero							
Dataset	AUC	Accuracy	Precision	Recall	F1	Confusion Sample	Matrix
D	0.568	51.400	0.016	0.571	0.032	[[1489 1452]	[26 33]]
11DM	0.571	54.684	0.025	0.559	0.048	[[1716 1225]	[26 33]]
11DF	0.574	61.656	0.024	0.530	0.046	[[1854 1087]	[27 32]]
11DFM	0.582	59.911	0.025	0.560	0.048	[[1771 1170]	[25 34]]
9DM	0.578	57.500	0.016	0.500	0.032	[[1696 1245]	[29 30]]
9DF	0.578	59.867	0.024	0.502	0.047	[[1769 1172]	[30 29]]
9DFM	0.588	59.233	0.023	0.575	0.044	[[1789 1152]	[24 35]]
6DM	0.585	50.533	0.018	0.643	0.035	[[1466 1475]	[21 38]]
6DF	0.585	55.733	0.024	0.571	0.046	[[1720 1221]	[26 33]]
6DFM	0.607	57.767	0.027	0.583	0.051	[[1721 1220]	[24 35]]
3DM	0.588	50.933	0.021	0.646	0.040	[[1476 1465]	[20 39]]
3DF	0.595	59.778	0.023	0.503	0.043	[[1777 1164]	[29 30]]
3DFM	0.629	61.233	0.028	0.572	0.053	[[1769 1172]	[25 34]]

Table 5.1: ICL Class-Zero Results

For Class-Ones, we train the model with mixed customers and test the model with customers who has not used ICL for the last 3 months. In this scenario, AUC, accuracy and recall is more important to evaluate, 0.87%, 84% and 0.64%, respectively. The reason is; this is an imbalanced dataset such that only 3% of the customers start using the product in the 12th month, so it is nearly impossible to achieve a high precision measure. Suppose that the model predicts 90% of both customers who will use/not use the product accurately. This will result in a confusion matrix with approximately these numbers; $\begin{bmatrix} 1385 & 160 \\ 43 & 6 \end{bmatrix}$. As we examine this confusion matrix, we see that it has impressive results in terms of accuracy and recall. On the other hand, the precision measure of this confusion matrix is only 21%. Thus, considering the precision measure as a performance indicator for this type of dataset is not appropriate.

Furthermore, we believe that using mixed customers in the training dataset provides useful information about product usage patterns. As we examine the confusion matrices in detail in Table 5.2, we can observe that the model predicts more than 85% of customers who will not use ICL, and 65% of the customers (almost 1.6% of total customers) who will use the product. Considering that the dataset is imbalanced, these numbers indicate that the model performs well in this scenario for each dataset type. Moreover, the AUC graph of Class-Ones with 9 months time period data is shown in Fig. 5.2. It can be seen that the performance ranking is $DFM \geq DF \geq DM \geq D$.

For Class-All, we train and test the model with mixed customers. These results are shown in Table 5.3. In this scenario, the model performs best in terms of AUC, precision, recall and F1 measures, with values 0.89%, 0.80%, 0.76% and 0.78%, respectively. The accuracy performance is also impressive with 82% , but Class-Ones has a better accuracy. All the measures are more important to evaluate the performance of Class-All scenario, because these datasets are not imbalanced. We see that 44% of the customers use the product in the 12th month, and 56% do not. Therefore, all the measures play an important role. Furthermore, the AUC graph of Class-All for each monthly data type is shown in Figures 5.3 to 5.6. It can be seen that the performance ranking is $DFM \geq DF \geq DM \geq D$.

Individual Consumer Loan							
Class-Ones							
Dataset	AUC	Accuracy	Precision	Recall	F1	Confusion Matrix	Sample
D	0.629	59.482	0.044	0.591	0.082	[[1000 541]	[21 28]]
11DM	0.732	86.052	0.121	0.512	0.196	[[1339 202]	[24 25]]
11DF	0.841	89.806	0.159	0.600	0.251	[[1384 157]	[18 31]]
11DFM	0.867	89.554	0.164	0.635	0.260	[[1380 161]	[17 32]]
9DM	0.759	80.585	0.110	0.579	0.185	[[1255 286]	[19 30]]
9DF	0.844	86.717	0.135	0.663	0.224	[[1339 202]	[17 32]]
9DFM	0.870	86.926	0.131	0.680	0.220	[[1322 219]	[16 33]]
6DM	0.781	83.943	0.102	0.645	0.176	[[1231 310]	[16 33]]
6DF	0.853	88.841	0.134	0.658	0.223	[[1343 198]	[17 32]]
6DFM	0.884	88.570	0.132	0.683	0.221	[[1327 214]	[15 34]]
3DM	0.789	86.133	0.138	0.622	0.226	[[1339 202]	[17 32]]
3DF	0.859	88.241	0.134	0.680	0.224	[[1321 220]	[16 33]]
3DFM	0.894	89.126	0.189	0.784	0.305	[[1380 161]	[11 38]]

Table 5.2: ICL Class-Ones Results

Individual Consumer Loan							
Class-All							
Dataset	AUC	Accuracy	Precision	Recall	F1	Confusion Sample	Matrix
D	0.669	62.431	0.537	0.646	0.587	[[886 787]	[477 850]]
11DM	0.835	80.233	0.799	0.636	0.696	[[1505 168]	[523 804]]
11DF	0.918	82.856	0.874	0.767	0.817	[[1399 274]	[304 1023]]
11DFM	0.924	86.922	0.858	0.828	0.843	[[1470 203]	[227 1100]]
9DM	0.847	79.067	0.739	0.713	0.715	[[1385 288]	[413 914]]
9DF	0.924	86.056	0.850	0.787	0.817	[[1471 202]	[280 1047]]
9DFM	0.931	87.211	0.846	0.802	0.823	[[1480 193]	[236 1091]]
6DM	0.869	79.733	0.761	0.683	0.709	[[1404 269]	[434 893]]
6DF	0.931	85.456	0.843	0.821	0.832	[[1469 204]	[228 1099]]
6DFM	0.940	85.651	0.838	0.854	0.846	[[1470 203]	[210 1117]]
3DM	0.878	79.500	0.767	0.715	0.730	[[1426 247]	[421 906]]
3DF	0.932	86.056	0.855	0.812	0.833	[[1500 173]	[276 1051]]
3DFM	0.942	85.556	0.868	0.830	0.849	[[1468 205]	[232 1095]]

Table 5.3: ICL Class-All Results

Individual Consumer Loan							
Class Zero							
Dataset	AUC	Accuracy	Precision	Recall	F1	Confusion Sample	Matrix
D	0.584	54.695	0.023	0.541	0.045	[[1719 1225]	[25 31]]
11DM	0.568	53.510	0.019	0.488	0.037	[[1699 1245]	[28 28]]
11DF	0.578	59.217	0.025	0.557	0.047	[[1789 1155]	[24 32]]
11DFM	0.586	59.857	0.023	0.484	0.043	[[1769 1175]	[29 27]]
9DM	0.558	51.415	0.019	0.581	0.036	[[1492 1452]	[23 33]]
9DF	0.564	59.925	0.017	0.574	0.033	[[1773 1171]	[23 33]]
9DFM	0.584	61.670	0.025	0.540	0.048	[[1856 1088]	[24 32]]
6DM	0.567	52.548	0.019	0.625	0.037	[[1469 1475]	[21 35]]
6DF	0.598	55.717	0.034	0.561	0.064	[[1722 1222]	[26 30]]
6DFM	0.619	57.755	0.030	0.593	0.057	[[1722 1222]	[21 35]]
3DM	0.569	50.943	0.025	0.658	0.048	[[1479 1465]	[20 36]]
3DF	0.581	59.789	0.028	0.516	0.054	[[1780 1164]	[27 29]]
3DFM	0.613	61.221	0.021	0.584	0.040	[[1769 1175]	[24 32]]

Table 5.4: ICL Class-Zero Results on New Customers

Individual Consumer Loan							
Class Ones							
Dataset	AUC	Accuracy	Precision	Recall	F1	Confusion Matrix	Sample
D	0.619	59.495	0.056	0.575	0.102	[[1005 542]	[18 25]]
11DM	0.744	86.042	0.107	0.526	0.178	[[1342 205]	[21 22]]
11DF	0.823	89.820	0.170	0.611	0.266	[[1386 161]	[15 28]]
11DFM	0.851	89.570	0.174	0.645	0.274	[[1382 165]	[13 30]]
9DM	0.772	80.575	0.096	0.569	0.164	[[1258 289]	[18 25]]
9DF	0.832	86.705	0.149	0.676	0.244	[[1342 205]	[15 28]]
9DFM	0.858	86.941	0.143	0.664	0.235	[[1327 220]	[13 30]]
6DM	0.793	79.957	0.090	0.627	0.158	[[1231 316]	[13 30]]
6DF	0.867	88.855	0.149	0.668	0.244	[[1349 198]	[15 28]]
6DFM	0.895	88.581	0.122	0.699	0.208	[[1327 220]	[13 30]]
3DM	0.799	86.117	0.149	0.604	0.239	[[1339 208]	[16 27]]
3DF	0.845	88.229	0.122	0.691	0.207	[[1324 223]	[14 29]]
3DFM	0.884	89.112	0.171	0.794	0.282	[[1383 164]	[9 34]]

Table 5.5: ICL Class-Ones Results on New Customers

Individual Consumer Loan						
Class All						
Dataset	AUC	Accuracy	Precision	Recall	F1	Confusion Matrix Sample
D	0.681	62.421	0.527	0.632	0.575	[[881 772] [487 860]]
11DM	0.835	79.077	0.721	0.724	0.723	[[1365 288] [427 930]]
11DF	0.940	86.040	0.834	0.802	0.817	[[1460 195] [297 1060]]
11DFM	0.942	87.199	0.860	0.786	0.821	[[1470 183] [247 1100]]
9DM	0.821	80.246	0.787	0.618	0.693	[[1493 160] [534 813]]
9DF	0.928	82.872	0.858	0.777	0.815	[[1379 274] [314 1033]]
9DFM	0.938	86.906	0.869	0.814	0.840	[[1450 203] [231 1116]]
6DM	0.853	79.747	0.749	0.699	0.723	[[1394 259] [444 903]]
6DF	0.914	85.467	0.827	0.805	0.816	[[1452 201] [245 1102]]
6DFM	0.955	85.635	0.853	0.844	0.848	[[1454 199] [218 1129]]
3DM	0.893	79.488	0.751	0.705	0.727	[[1423 230] [426 921]]
3DF	0.916	86.072	0.845	0.802	0.823	[[1490 163] [286 1061]]
3DFM	0.955	85.571	0.881	0.845	0.863	[[1458 195] [249 1098]]

Table 5.6: ICL Class-All Results on New Customers

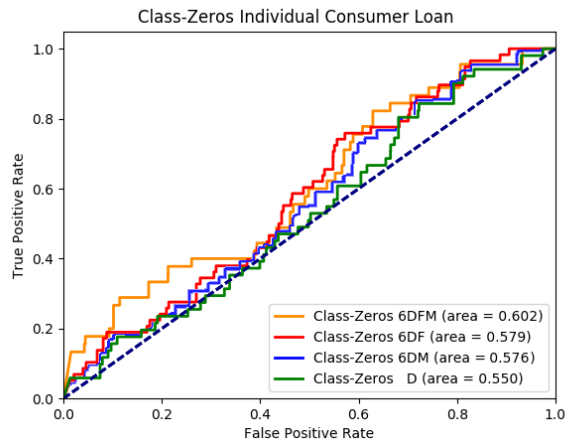


Figure 5.1: ICL Class-Zero 6 Month

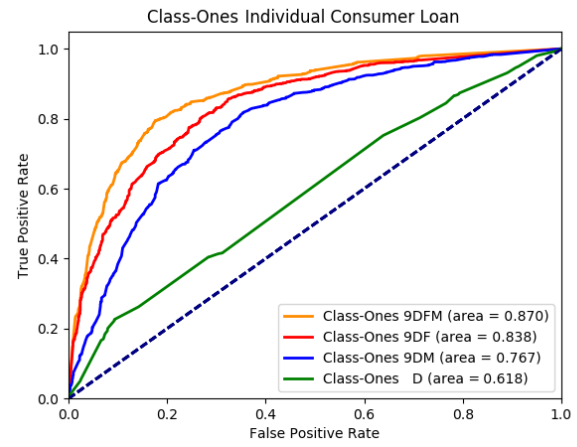


Figure 5.2: ICL Class-Ones 9 Month



Figure 5.3: ICL Class-All 3 Month

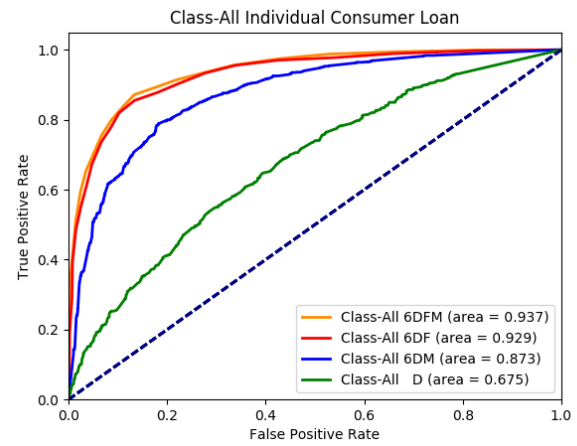


Figure 5.4: ICL Class-All 6 Month

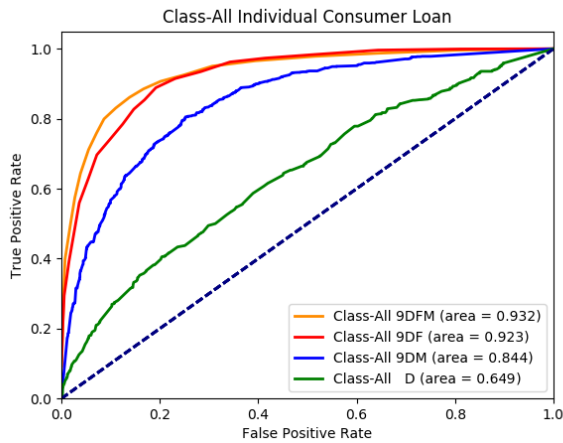


Figure 5.5: ICL Class-All 9 Month

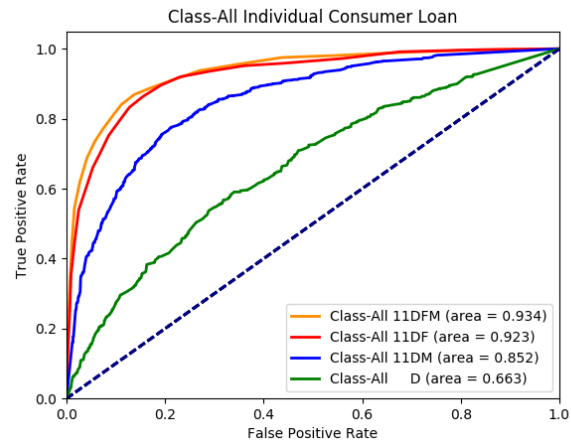


Figure 5.6: ICL Class-All 11 Month

We calculate a special measure for each of the features from our model called "variable importance". The variable importance of a feature is the increase in the prediction error when that specific feature is extracted from the model and all the other features are kept the same. Liaw and Wiener (2002) use this approach in their study and try to understand the importance of each feature independent from others on their random forest classifier. In order to obtain a feature importances table, we extract all the errors for each of the datasets and then sum them all. Then, we take the top 10 largest numbers out of that list and we divide the summation of these variables into the corresponding value to obtain a percentage value. The higher percentage means the higher error when removing that variable, so the percentage is proportional to the importance of the feature. For instance, if the error of the age feature is 75 and that of the income feature is 5, the effect of the age feature is more important than income. Let us suppose that we simplify the process and have only two features. If there exists two customers with the same age, but with unlike incomes, they are likely to be treated the same in the decision tree path. On the other hand, let us suppose we have two customers with ages 20 and 50 with the same income levels. They are expected to be treated differently, because their age difference is large and the age feature has a high importance in the decision tree structure.

We perform this analysis for only Class-All datasets, because we believe that it represents the practice more than the other scenarios. The feature importance measures are illustrated in Table 5.7 for ICL. As it can be seen from the table, number of active products, mean of the credit scores and cluster+d-l-r features are the ones the classifiers give most importance to. The effect column in the figure indicates the proportionality of the features with the ICL usage. In other words, positive effect means that the higher the feature value, the more likely is ICL usage. On the other hand, negative effect means that the lower the feature value, the more likely is ICL usage. Those effects and importances are also shown in Fig. 5.7, where the bars going up (down) means positive (negative) effect.

Individual Consumer Loan		
Feature Name	Importance	Effect
Active Products	25.74%	Positive
Mean	13.25%	Positive
cluster+d-l-r	9.86%	Negative
Income	9.84%	Positive
radial+d-l-r	9.18%	Negative
spatial_cluster_regularity	8.46%	Positive
Average Balance	7.59%	Positive
spatial_radial_regularity	5.86%	Positive
weekly_regularity	5.71%	Positive
week+d+l+r	4.51%	Positive

Table 5.7: ICL Feature Importances

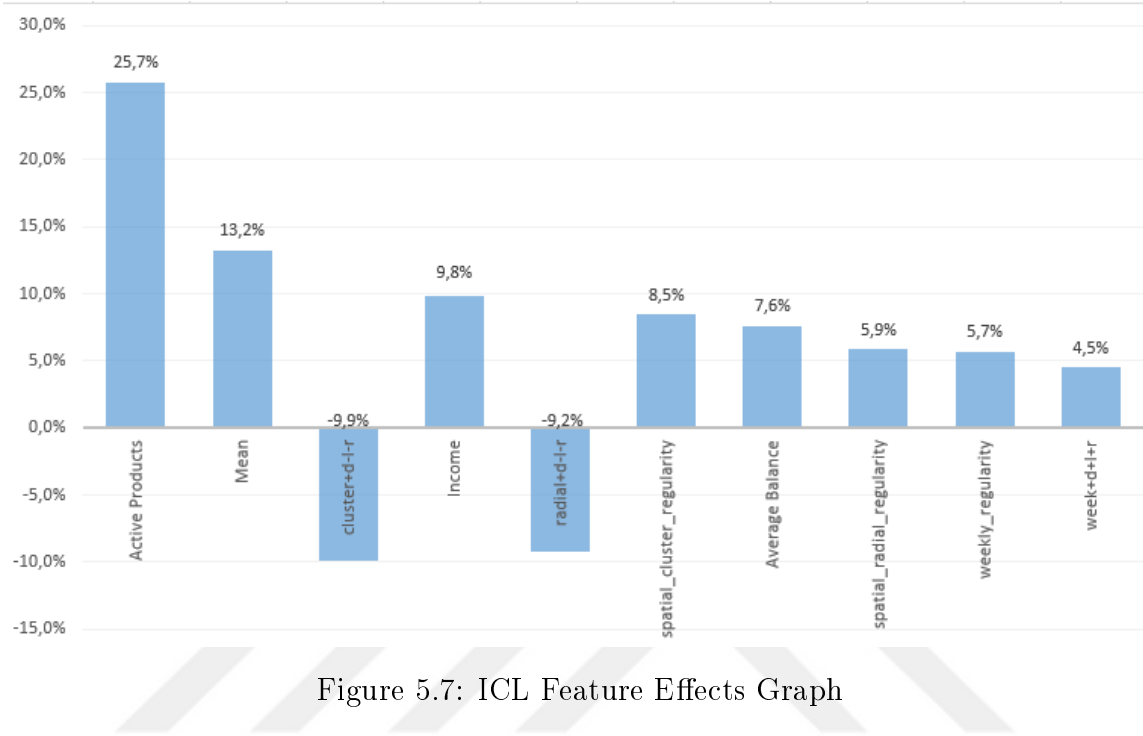


Figure 5.7: ICL Feature Effects Graph

5.1.2 Inferences about ICL

At this point, we obtain all the results including performance measures and feature effects for ICL product. After we examine the results in Tables 5.1, 5.2 and 5.3 in detail, we discover an interesting pattern. No matter the class type is, as the monthly buckets decrease, the overall performance of the classifier increases and the difference between classes DFM and DF increases. In other words, $3DFM \geq 6DFM \geq 9DFM \geq 11DFM$, $3DF \geq 6DF \geq 9DF \geq 11DF$ and $3DM \geq 6DM \geq 9DM \geq 11DM$. Moreover, $(3DFM - 6DFM) \geq (6DFM - 9DFM) \geq (9DFM - 11DFM)$. This pattern brings up the conclusion that as the mobility behavior is more up to date, the classifier performs better for ICL. When we go deeply into this inference, it makes sense, because the mobility pattern for a customer for 11 months is a very long time period and it is very difficult to reflect the behavior of the customer in the 12th month. On the contrary, the mobility pattern of the last 3 months can reflect the user's behavior in the 12th month better.

As we go deeper into the analysis, we also conclude that the product usage of

ICL is determined 34% by mobility behavior, and 66% by financial behavior, as illustrated in Fig. 5.9. Since one third of the result is dependent on mobility behavior, it is reasonable to claim that: As the mobility pattern is up to date, it can significantly increase the classifier performance. Furthermore, as we examine the feature importances in Table 5.7, we first see that the number of active products has the highest importance for ICL. This relation is logical, because an active bank products user is more likely to use an ICL. Other important features are mean, income and average balance. To examine ICL further, we have to keep in mind that, all bank customers can apply for ICL usage, but the bank approves the ones who can use. The bank analyzes the customer's financial status and if they believe that the customer can pay the credit back, then they approve the request. In our dataset, if a customer uses an ICL, this means that the customer has already passed the approval phase and took the credit from the bank. Therefore, the critical importance values of mean, income and average balance features completely make sense, because these are the criteria that the bank evaluates the customers when approving the ICL product. We can also see that all these features have a positive effect, which means that as the mean credit score, income and average balance is larger, the customer is more likely to use the ICL.

Besides those features, negative `cluster+d-l-r` and positive `spatial_cluster_regularity`, which are the summation of cluster diversity, minus cluster loyalty, minus cluster regularity and following the same spatial cluster pattern through time period, respectively. Before going deeper into the analysis, we should first take a look at our cluster seeds throughout Turkey to better interpret the results, illustrated in Fig. 5.8. As we can see from the figure, most of the cluster seeds are spread almost equitably across Turkey except Istanbul seeds. In fact, 8 of 50 seeds are located in Istanbul. This is not a surprising result, because 19% of Turkey's population lives in Istanbul, although there are 80 other provinces.

After acknowledging the cluster seeds, for better understanding, let us first convert the effect of `cluster+d-l-r` into positive by multiplying the summation with -1 ,

which is minus diversity, positive loyalty and positive regularity (cluster-d+l+r with positive effect). This feature implies that as the customer spends his/her transactions not equitably among his/her clusters and as the customer performs most of his transactions in his/her favorite clusters and finally as the customer follows this pattern during his/her time period, then the customer is more likely to use ICL. Let us visualize a customer following this pattern in Fig. 5.10. Suppose that he/she lives in Istanbul (Sariyer), which will lead to an intensity of transactions in Istanbul. Also suppose that he/she also spends money in both Izmir, Istanbul (Besiktas) and Antalya clusters. Izmir and Antalya expenses can be because of any reason like holiday, family visit, business trip etc., and transactions within Istanbul are probably because of daily expenses. Keeping in mind that the locations of the cluster seeds are mostly provinces in Turkey, it is expected that the expense pattern is not equitably spread, highly loyal and regular in terms of clusters. This kind of pattern results in a standard life, with a regular ongoing job and most of the expenses are gathered in same places. In other words, these two features point out the people with regular life cycles in terms of mobility.



Figure 5.8: Cluster Centers Throughout Turkey

Other features to mention are radial+d-l-r with negative effect and spatial_radial_regularity with positive effect, which are the summation of radial diversity, minus radial loyalty,

minus radial regularity and following the same spatial radial patterns through the time period, respectively. Let us also convert the radial+d-l-r as we did before for better understanding. The converted radial-d+l+r feature implies that as the customer spends his transactions not equitably among his radials and as the customer spends most of his transactions in his top three radials and finally as the customer follows this pattern during his time period, then the customer is more likely to use ICL. For instance, supposing that we only have 1, 3, 5 and 15 kilometers as our radial radius, as illustrated in Fig. 5.11 as an example. In this example, the customer spends 75% of his transactions in one radial among 4 radials, so the diversity is low because the expenses are not equitably spread, and the top preferred radial has most of the expenses, which lead to high loyalty measure. Let us also consider that the customer always follows this radial pattern during his expenses for almost every month, which will lead to a high radial regularity. This type of customer is very predictable in terms of radial expense patterns, meaning that this customer lives a regular life, probably with the same job and same places to spend his money and follows this life pattern regularly through his life. As a result, negative radial+d-l-r and positive spatial_radial_regularity means that, the customer has a steady life, goes to the same places and follows the same life pattern. Moreover, this kind of customers are more likely to apply and get approval from banks for the ICL product, which totally makes sense.

The last features to be examined are positive week+d+l+r and positive weekly_regularity, which are the summation of diversity, loyalty and regularity and following the same weekly patterns over the time period, respectively. As we go deeper into the analysis of these features, high week+d+l+r and weekly_regularity mean that the customer spends his/her transactions equitably among different places, visits most of the time top three places and follows the same pattern through weekly cycles. In other words, this is also another sign that the customer has a weekly steady life, follows the same patterns every week and goes to the same places throughout his life.

To further understand the important mobility behaviors for ICL, let us think about

a customer having a full time well paid office job for 5 years and she lives in Istanbul. In her regular weekly cycle, she works in the office from 8 a.m. to 5 p.m. five days a week. She probably has her breakfast/coffee from the same bakeries and lunches in her favorite restaurants recurrently. Besides, considering that she works in the same company for 5 years, she is highly predictable in terms of mobility and if she will not quit the job, she will most probably follow the same pattern for the next years. She will spend her money from her favorite places equitably, loyally and recursively, which results in high diversity, loyalty and regularity in terms of weekly patterns, but since she mostly spends her money in Istanbul, the diversity will be low, loyalty and regularity will be high in terms of cluster and radial structures.

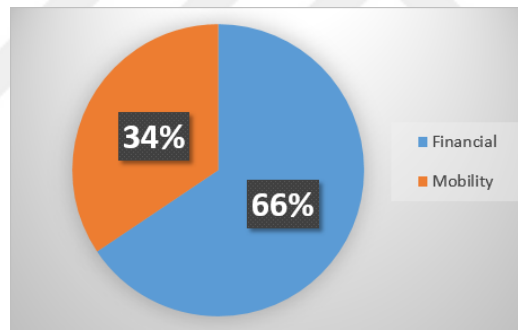


Figure 5.9: ICL Feature Effects Pie Chart

cluster+d-l-r			
Istanbul (Sariyer)	Istanbul (Besiktas)	Izmir	Antalya
156	59	8	15

Figure 5.10: ICL cluster+d-l-r Example

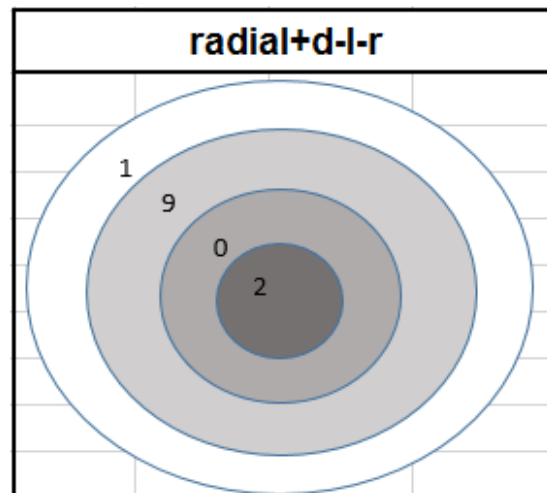


Figure 5.11: ICL radial+d-l-r Example

The summary of the inferences from ICL results is as follows:

- The performance ranking is always $DFM \geq DF \geq DM \geq D$ regardless of the dataset type.
- As the mobility features are more up to date, the classifier performance increases, $3DFM \geq 6DFM \geq 9DFM \geq 11DFM$, $3DF \geq 6DF \geq 9DF \geq 11DF$ and $3DM \geq 6DM \geq 9DM \geq 11DM$.
- The effect of the mobility increases as the mobility data is up to date, $(3DFM - 6DFM) \geq (6DFM - 9DFM) \geq (9DFM - 11DFM)$.
- All the customers are able to apply for ICL, but customers with higher credit score, mean and average balance are more likely get approval from the bank.
- Customers with more steady life cycles are more likely to get approval from the bank for ICL usage.
- Two thirds of the recommendation decision is determined by financial features.
- One third of the recommendation decision is determined by mobility features.

5.2 Retirement Savings

In this section, we will examine the results of RS in detail and we will interpret the results in terms of demographic, financial and mobility behaviors.

5.2.1 Computational Results of RS

At the end of performing all the preprocessing and analysis for RS, we finally obtain the results for all scenarios with corresponding optimum assignments. The average AUC, accuracy, precision, recall, F1 scores and a confusion matrix sample are shown in Tables 5.8, 5.9 and 5.10 for each scenario. Besides, the sample AUC graphs of RS

are shown in Figures 5.12 to 5.17. Afterwards, to evaluate the performance of the model on a completely different customer set, we test our model on the next 3,000 customers which we have never trained our model with. Those results for the RS product are shown in Tables 5.11, 5.12 and 5.13. After comparing these results with the ones that we have before, we observe that our model is not over fitting and there is not a significant difference between the performances.

For Class-Zero, we train and test the model with customers who have not used RS for the last 3 months. The ratio of the customers who start using the product in the 12th month is extremely low with only 1%, so the dataset is highly imbalanced. It is challenging to accurately find those 1% of the customers while achieving a great performance. The AUC, accuracy and recall measures are quite good, 79%, 82% and 60%, respectively. As we examine the confusion matrix example in Table 5.8, we observe that the model accurately predicts 85% of the customers who will not use the product. Moreover, it also succeeds to predict accurately the 60% of the customers who will use the product, which actually corresponds to approximately 1% of all customers. Thus, the model performs quite well on this scenario. The AUC graphs of this scenario is represented in Fig. 5.12. As it can be observed from both the graph and the table, the performance ranking is $DFM \geq DF \geq DM \geq D$.

For Class-Ones scenario, we train the model with mixed customers and test the model with customers who have not used RS for the last 3 months. Since this dataset is also a highly imbalanced one, the critical measures to evaluate are AUC, accuracy and recall, which are 0.86%, 96% and 0.50%, respectively. Although the recall measure is weak, the AUC and the accuracy is high enough to claim that the model performs well on this scenario. As we examine the confusion matrices in Table 5.9, we observe that the accuracy is extremely high and the model accurately predicts 98% of the customers who will not use RS in the 12th month. On the other hand, the recall is weak with 50% but it is extremely challenging to find 1% of the customers accurately. In this scenario, like in the ICL product, since only 1% of the customers start using the product in the 12th month, it is almost impossible to achieve a high precision value.

As we discussed in Sec. 5.1.1, considering the precision measure as a performance indicator for this dataset is not appropriate. The AUC graph for one of the Class-Ones dataset is illustrated in Fig. 5.13. It cannot be directly seen from the graph, but for all the datasets, the performance ranking is $DFM \geq DF \geq DM \geq D$.

Retirement Savings							
Class-Zero							
Dataset	AUC	Accuracy	Precision	Recall	F1	Confusion Sample	Matrix
D	0.598	59.482	0.021	0.591	0.041	[[1690 1273]	[15 22]]
11DM	0.690	78.667	0.032	0.421	0.059	[[2342 621]	[21 16]]
11DF	0.842	86.991	0.055	0.601	0.100	[[2588 375]	[15 22]]
11DFM	0.843	86.054	0.051	0.635	0.095	[[2560 403]	[14 23]]
9DM	0.729	82.267	0.032	0.435	0.059	[[2341 622]	[20 17]]
9DF	0.849	88.841	0.056	0.658	0.102	[[2431 532]	[14 23]]
9DFM	0.851	86.926	0.059	0.680	0.109	[[2499 464]	[13 24]]
6DM	0.719	79.267	0.031	0.430	0.057	[[2333 630]	[18 19]]
6DF	0.859	87.041	0.071	0.680	0.129	[[2640 323]	[11 26]]
6DFM	0.861	87.126	0.071	0.703	0.129	[[2637 326]	[11 26]]
3DM	0.734	72.133	0.030	0.692	0.057	[[2155 808]	[12 25]]
3DF	0.858	88.717	0.064	0.683	0.118	[[2597 366]	[12 25]]
3DFM	0.860	88.570	0.070	0.673	0.127	[[2649 314]	[13 24]]

Table 5.8: RS Class-Zero Results

For Class-All, we train and test the model with mixed customers. The results are shown in Table 5.10. In this scenario, the model performs best in terms of AUC, precision and F1 measures, with values 0.86%, 0.79% and 0.64%, respectively. The accuracy performance is also impressive with 88%, but Class-Ones has a better accuracy. For Class-All scenarios of RS, the dataset is slightly imbalanced such that 17% of the customers use the product in the 12th month, and 83% do not. Despite this, we

still use all the measures to evaluate the performance for this scenario. Despite the imbalance, the performance measures are promising except for the recall measure. We infer that the overall problem for RS product is actually the recall measure regardless of the scenario type: all the scenarios are suffering in terms of the recall metric. In other words, the model struggles while predicting the product usage pattern for RS. The AUC graph of Class-All for each monthly data type, is shown in Figures 5.14 to 5.17. Here it can be seen that the performance ranking is $DFM \geq DF \geq DM \geq D$.

Retirement Savings							
Class-Ones							
Dataset	AUC	Accuracy	Precision	Recall	F1	Confusion Matrix	Sample
D	0.529	71.524	0.008	0.295	0.016	[[1400 564]	[12 4]]
11DM	0.824	98.908	0.264	0.381	0.312	[[1939 25]	[9 7]]
11DF	0.899	97.891	0.209	0.547	0.302	[[1929 35]	[7 9]]
11DFM	0.902	97.253	0.205	0.558	0.300	[[1925 39]	[7 9]]
9DM	0.874	97.653	0.182	0.499	0.267	[[1932 32]	[8 8]]
9DF	0.913	98.421	0.223	0.604	0.326	[[1935 29]	[7 9]]
9DFM	0.916	98.260	0.264	0.607	0.368	[[1939 25]	[6 10]]
6DM	0.831	98.093	0.171	0.368	0.233	[[1932 32]	[9 7]]
6DF	0.909	97.798	0.223	0.602	0.326	[[1925 39]	[6 10]]
6DFM	0.912	98.804	0.215	0.601	0.316	[[1930 34]	[5 11]]
3DM	0.866	98.723	0.192	0.333	0.244	[[1942 22]	[11 5]]
3DF	0.905	97.915	0.202	0.591	0.301	[[1924 40]	[7 9]]
3DFM	0.908	98.676	0.214	0.595	0.315	[[1931 33]	[7 9]]

Table 5.9: RS Class-Ones Results

Retirement Savings						
Class-All						
Dataset	AUC	Accuracy	Precision	Recall	F1	Confusion Matrix Sample
D	0.582	68.666	0.195	0.287	0.232	[[1897 594] [364 145]]
11DM	0.837	90.100	0.707	0.496	0.583	[[2389 102] [250 259]]
11DF	0.906	90.611	0.881	0.572	0.693	[[2454 37] [220 289]]
11DFM	0.908	91.456	0.889	0.579	0.701	[[2452 39] [218 291]]
9DM	0.834	89.201	0.704	0.466	0.561	[[2387 104] [269 240]]
9DF	0.907	91.344	0.886	0.578	0.699	[[2449 42] [218 291]]
9DFM	0.909	91.556	0.889	0.579	0.701	[[2452 39] [217 292]]
6DM	0.850	89.833	0.828	0.546	0.658	[[2429 62] [228 281]]
6DF	0.911	91.335	0.871	0.636	0.735	[[2434 57] [186 323]]
6DFM	0.912	91.133	0.879	0.637	0.739	[[2433 58] [176 333]]
3DM	0.845	89.265	0.825	0.511	0.631	[[2431 60] [248 261]]
3DF	0.906	90.978	0.839	0.601	0.701	[[2431 60] [200 309]]
3DFM	0.908	90.033	0.821	0.607	0.698	[[2425 66] [197 312]]

Table 5.10: RS Class-All Results

Retirement Savings							
Class Zero							
Dataset	AUC	Accuracy	Precision	Recall	F1	Confusion Matrix	Sample
D	0.603	58.348	0.030	0.572	0.057	[[1685 1272]	[17 26]]
11DM	0.699	78.678	0.024	0.403	0.045	[[2318 639]	[24 19]]
11DF	0.854	86.040	0.061	0.649	0.112	[[2560 397]	[16 27]]
11DFM	0.857	87.006	0.045	0.616	0.083	[[2582 375]	[17 26]]
9DM	0.717	82.278	0.043	0.448	0.078	[[2331 626]	[23 20]]
9DF	0.841	86.939	0.065	0.693	0.119	[[2490 467]	[16 27]]
9DFM	0.864	88.857	0.042	0.671	0.078	[[2421 536]	[17 26]]
6DM	0.701	79.251	0.041	0.412	0.074	[[2325 632]	[22 21]]
6DF	0.875	87.053	0.076	0.666	0.137	[[2640 323]	[15 28]]
6DFM	0.875	87.114	0.067	0.717	0.122	[[2632 325]	[15 28]]
3DM	0.720	72.121	0.024	0.680	0.046	[[2150 807]	[15 28]]
3DF	0.848	88.580	0.054	0.663	0.101	[[2643 314]	[16 27]]
3DFM	0.871	88.705	0.074	0.698	0.134	[[2591 366]	[15 28]]

Table 5.11: RS Class-Zero Results on New Customers

Retirement Savings							
Class Ones							
Dataset	AUC	Accuracy	Precision	Recall	F1	Confusion Matrix	Sample
D	0.519	71.538	0.009	0.305	0.018	[[1399 555]	[17 9]]
11DM	0.839	98.894	0.254	0.395	0.309	[[1929 25]	[14 12]]
11DF	0.890	97.237	0.193	0.574	0.289	[[1915 39]	[12 14]]
11DFM	0.913	97.873	0.193	0.558	0.286	[[1919 35]	[12 14]]
9DM	0.887	97.669	0.197	0.513	0.285	[[1922 32]	[13 13]]
9DF	0.926	98.436	0.234	0.620	0.340	[[1925 29]	[12 14]]
9DFM	0.929	98.246	0.252	0.619	0.359	[[1929 25]	[11 15]]
6DM	0.815	98.075	0.182	0.356	0.241	[[1922 32]	[14 12]]
6DF	0.896	98.794	0.229	0.612	0.333	[[1920 34]	[10 16]]
6DFM	0.922	97.812	0.238	0.584	0.339	[[1915 39]	[11 15]]
3DM	0.856	98.733	0.182	0.349	0.240	[[1932 22]	[16 10]]
3DF	0.893	97.928	0.212	0.575	0.310	[[1914 40]	[12 14]]
3DFM	0.920	98.692	0.226	0.610	0.330	[[1921 33]	[12 14]]

Table 5.12: RS Class-Ones Results on New Customers

Retirement Savings						
Class All						
Dataset	AUC	Accuracy	Precision	Recall	F1	Confusion Matrix Sample
D	0.568	68.656	0.253	0.277	0.264	[[1892 589] [370 149]]
11DM	0.819	90.113	0.691	0.508	0.585	[[2379 102] [260 258]]
11DF	0.890	91.467	0.901	0.593	0.715	[[2452 39] [218 301]]
11DFM	0.896	90.593	0.895	0.558	0.687	[[2454 37] [228 291]]
9DM	0.844	89.217	0.717	0.452	0.555	[[2387 104] [274 245]]
9DF	0.917	91.356	0.899	0.562	0.692	[[2449 42] [224 295]]
9DFM	0.924	91.566	0.899	0.569	0.697	[[2442 39] [222 297]]
6DM	0.864	89.847	0.814	0.534	0.645	[[2419 62] [228 291]]
6DF	0.922	91.149	0.892	0.648	0.751	[[2428 53] [180 339]]
6DFM	0.923	91.349	0.853	0.648	0.736	[[2424 57] [196 323]]
3DM	0.861	89.276	0.809	0.495	0.614	[[2426 55] [258 261]]
3DF	0.917	90.966	0.849	0.589	0.696	[[2426 55] [209 310]]
3DFM	0.924	90.015	0.834	0.593	0.693	[[2420 61] [202 317]]

Table 5.13: RS Class-All Results on New Customers

The other critic measure we obtain from the results is the individual importance measure of each feature. We also run this analysis for Class-All datasets of RS, where all the classifiers are decision tree based; either an extra tree classifier, a decision tree classifier or a random forest tree classifier (Liaw and Wiener (2002)). For the importance measure evaluation, we basically follow the same approach as discussed in Sec. 5.1.1. The importance features of RS are shown in Table 5.14. As it can be seen from the table, the number of active products, average balance and income features are the ones the classifiers pay more attention to. Both the importances and the effects of the features are shown in Fig. 5.18, where the bars going up(down) show positive(negative) effect.



Figure 5.12: RS Class-Zero 6 Month



Figure 5.13: RS Class-Ones 3 Month

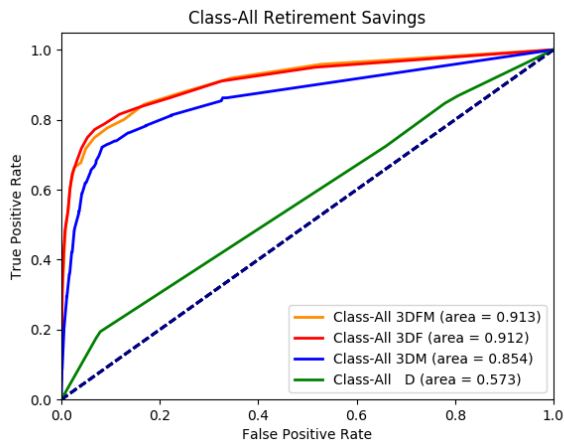


Figure 5.14: RS Class-All 3 Month



Figure 5.15: RS Class-All 6 Month



Figure 5.16: RS Class-All 9 Month

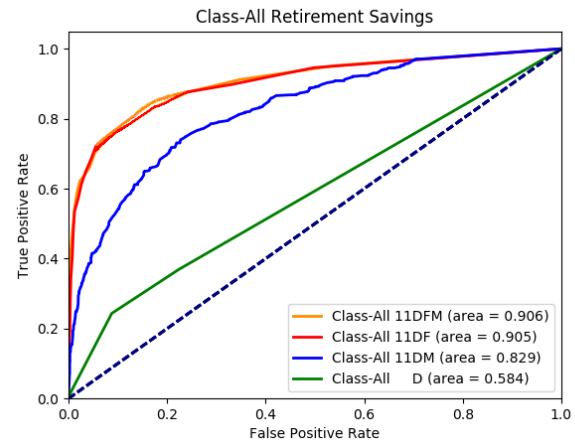


Figure 5.17: RS Class-All 11 Month

Retirement Savings		
Feature Name	Importance	Effect
Active Products	25.49%	Positive
Average Balance	11.54%	Positive
Income	10.83%	Positive
Last Balance	9.21%	Positive
Monthly Total Expense	9.20%	Negative
Total Credit Card Limit	8.84%	Positive
Credit Card Limit	8.38%	Positive
Age	7.60%	Negative
weekly_regularity	4.56%	Negative
week+d+l+r	4.36%	Negative

Table 5.14: RS Feature Importances

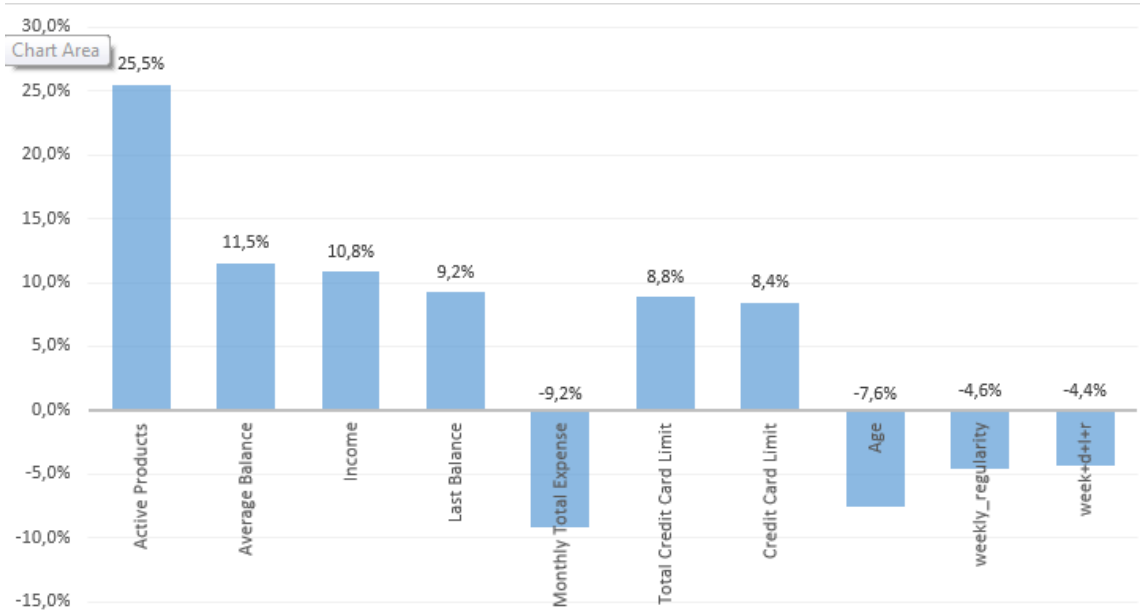


Figure 5.18: RS Feature Effects Graph

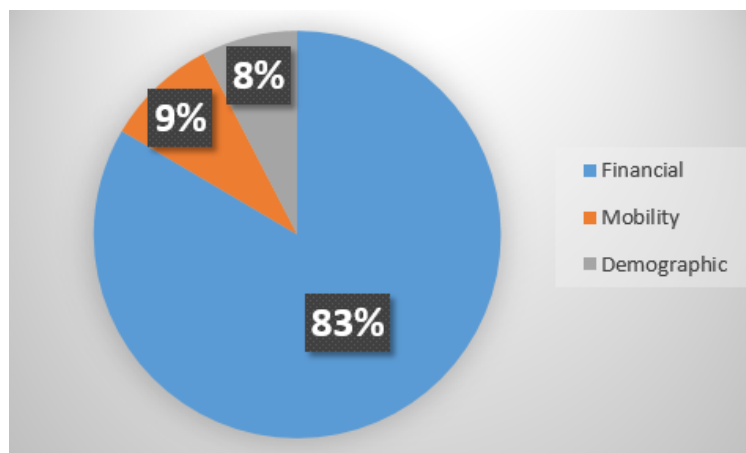


Figure 5.19: RS Feature Effects Pie Chart

5.2.2 Inferences about RS

At this point, we have obtained all the results including the performance measures and feature effects for the RS product. After we examine the results in Tables 5.8, 5.9 and 5.10 in detail, we find out that there is no significant relation between the mobility behavior and RS usage. As it can be observed from the tables, the mobility information slightly increases the overall performance of the model, but there is not a great effect on the performance. As we analyze further, we see that 84% of the decision is determined according to financial behaviors, as illustrated in Fig. 5.19. The mobility and demographic features are only 9% and 8%, respectively. This result supports the previous deduction that, the mobility behavior has no significant role determining the product usage of RS.

Before we go deeper into the analysis, apart from ICL, RS usage does not require a bank approval, so anyone can use RS from a bank as long as he/she pays the agreed amount of money. From now on, we see that all the top 7 important features are financial, as illustrated in Fig. 5.14. The number of active products with positive effect is again the most important feature for the model. We can claim that as the customer uses more bank products, he/she also has a tendency to use RS. Furthermore, average balance, income, last balance, total credit card limit and credit card limit are other important features with positive effect and monthly total expense has a negative effect. This combination is actually surprising. It means that: Customers with a well financial being, who also spend less money, have more tendency to use RS. In other words, wealthy and thrifty customers are more likely to use RS. If we consider that RS is essentially a way of saving money, which is a characteristic of thrifty persons, this deduction becomes completely logical. Furthermore, the age has a negative effect on RS usage, which is totally sensible, because young customers are more likely to use RS. There is not much benefit to using RS after 50 years old.

The only mobility features that are included in the Table 5.14 are negative weekly_regularity and negative week+d+r+l, which are weekly regular pattern and summation of weekly diversity, loyalty and regularity, respectively. This is another surprising conclusion

since large `weekly_regularity` and `week+d+r+l` mean that the customer follows the same weekly patterns recursively, including visiting same places equitably, loyally and recurrently. Since the sign of these features is negative, it says that: As the customer becomes more unpredictable in terms of weekly spending patterns, he/she is more likely to use RS. In other words, the customers who spend their time in different places without any sign of regularity are more likely to save their money for future purposes.

The summary of the inferences from RS results is as follows:

- The performance ranking is always $DFM \geq DF \geq DM \geq D$ regardless of the dataset type.
- The mobility feature slightly increases the performance of the model.
- Wealthy and thrifty customers are more likely to use RS.
- Customers with less steady life cycles are more likely to use RS for future savings.
- 83% of the decision is determined by financial features.
- 9% and 8% of the decision is determined by mobility and demographic features, respectively.

Chapter 6

CONCLUSION

The results obtained in this study highlight the importance of spatio-temporal mobility features in predicting the financial outcomes of the individuals. The approach we implement in this study is fundamentally different from other studies. To the best of our knowledge, there does not exist any study investigating the direct relation between the financial product usage and individual's mobility patterns. In this manner, we analyze a one year bank transaction data of Turkey's one of the largest banks and obtain several surprising outcomes for both ICL and RS product usage patterns. During the analysis, we use various data preprocessing approaches and then, we divide the data into different characteristics. Afterwards, we obtain the best performing method assignments for each of them. Eventually, after the exhaustive runs that we perform on each product independently, we notice surprising patterns. We observe that since there is a bank approval requirement for ICL, the financial requirements are extremely tight for ICL usage. Moreover, we discover that customers with a more steady life are more likely to use the ICL product. Furthermore, one of the most interesting outcomes that we find out is; as the mobility information becomes up to date, it provides useful information about the user's future predictability. As a result, we find a direct relation between the mobility behavior and ICL usage pattern. As the predictability and regularity increases in terms of mobility, the customer has more tendency to use ICL. In other words, people with regular well-paid jobs and regular life cycles are more likely to use ICL. On the other hand, we could not discover any relation between the mobility behavior and RS usage. Most of the RS usage is explained by financial behaviors. The mobility information increases the classifier performance slightly. Even so, we discover an interesting pattern for RS, too. We

conclude that the customers with more money who spend their money less are more likely to use RS. In other words, wealthy and thrifty customers have a tendency of using RS. Furthermore, in terms of mobility, as the regularity and stability decreases, the customers become more likely to use RS. This interesting outcome means that; people who live their lives more random and unsteadily, are more likely to save money for their future. Since there is not a study in the literature seeking the direct relation between different mobility behaviors and product usage patterns, this study is unique in terms of this topic. The obtained results reveal the deep interconnections between financial outcomes and human mobility data. As a future work, we want to work simultaneously with banks and directly use our product recommendation on their customers. Afterwards, after some amount of time, we will measure the performance of the model according to the successive returns from the customers. This will provide us to see the real performance of how our product recommendations work.

BIBLIOGRAPHY

- Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. *Proceedings of COMPSTAT'2010. Physica-Verlag HD*, pages 177–186.
- Candia, J., Gonzalez, M. C., Wang, P., Schoenharl, T., Madey, G., and Barabas, A. L. (2008). Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical*, 4206(22).
- Chandrashekar, G. and Sahin, F. (2014). A survey on feature selection methods. *Computers and Electrical Engineering*, 40(1):16–28.
- Choi, S. H., Kang, S., and Jeon, Y. J. (2006). Personalized recommendation system based on product specification values. *Expert Systems with Applications*, 31(3):607–616.
- Cohen, J., Dolan, B., Dunlap, M., Hellerstein, J. M., and Welton, C. (2009). Mad skills: New analysis practices for big data. *Proceedings of the VLDB Endowment*, 2(2):1481–1492.
- Cordon, O., Damas, S., and Santamaria, J. (2006). Feature-based image registration by means of the evolutionary algorithm. *Image and Vision Computing*, 24(5):525–533.
- Desir, C., Petitjean, C., Heutte, L., Salaun, M., and Thiberville, L. (2012). Classification of endomicroscopic images of the lung based on random subwindows and extra-trees. *IEEE Transactions on Biomedical Engineering*, 59(9):2677–2683.
- Ding, C. and Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. *Journal of Bioinformatics and Computational Biology*, 3(2):185–205.

- Drummond, C. and Holte, R. C. (2003). Class imbalance and cost sensitivity: Why sampling beats over-sampling. *In workshop on learning from imbalanced datasets II (within ICML)*.
- Freund, Y. and Schapire, R. E. (1996). Experiments with a new boosting algorithm. *Machine Learning: Proceedings of the Thirteenth International Conference*.
- Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., and Herrera, F. (2011). An overview of ensemble methods for binary classifiers in multi-class problems: Experimental study on one-vs-one and one-vs-all schemes. *Pattern Recognition*, 44(8):1761–1776.
- Gardner, M. and Dorling, S. (1998). Artificial neural networks (the multilayer perceptron): A review of applications in the atmospheric sciences. *Atmospheric Environment*, 32(14-15):2627–2636.
- George, J. H. and Langley, P. (1995). Estimating continuous distributions in bayesian classifiers. *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345.
- Gonzalez, M. C., Hidalgo, C. A., and Barabasi, A. L. (2008). Understanding individual human mobility patterns. *Nature*, 453:779–782.
- Isaacman, S., Becker, R., Caceres, R., Kobourov, S., Rowland, J., and Varshavsky, A. (2010). A tale of two cities. *Proceedings of the Eleventh Workshop on Mobile Computing Systems and Applications*, pages 19–24.
- Kim, H., Golub, G. H., and Park, H. (2005). Missing value estimation for dna microarray gene expression data: local least squares imputation. *Bioinformatics*, 21(2):187.
- Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Ijcai*, 14(2):1137–1145.

- Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273–324.
- Krumme, C., Llorente, A., Cebrian, M., Pentland, A., and Moro, E. (2012). The predictability of consumer visitation patterns. *Scientific Reports*, 3(1645).
- Kung, K. S., Greco, K., Sobolevsky, S., and Ratti, C. (2014). Exploring universal patterns in human home-work commuting from mobile phone data. *Plos One*, 9(6).
- Law, M., Figueiredo, M., and Jain, A. (2004). Simultaneous feature selection and clustering using mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1154–1166.
- Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R News*, 2(3):18–22.
- Liu, D. R. and Shih, Y. (2005). Integrating ahp and data mining for product recommendation based on customer lifetime value. *Information and Management*, 42(3):387–400.
- Martin-Fernandez, J., Barcelo-Vidal, C., and Glahn-Pawłowsky, V. (2003). Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Mathematical Geology*, 35(3):253–278.
- Ng, A. Y. and Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in Neural Information Processing Systems 14*, pages 841–848.
- Noulas, A., Scellato, S., Lathia, N., and Mascolo, C. (2012). A random walk around the city: New venue recommendation in location-based social networks. *SOCIALCOM-PASSAT '12 Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust*, pages 144–153.

- Pal, M. (2005). Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, 26(1):217–222.
- Park, Y. J. and Chang, K. N. (2009). Individual and group behavior-based customer profile model for personalized product recommendation. *Expert Systems with Applications*, 36(2):1932–1939.
- Pudil, P., Novovicova, J., and Kittler, J. (1994). Floating search methods in feature selection. *Pattern Recognition Letters*, 15(11):1119–1125.
- Pujol, O., Radeva, P., and Vitria, J. (2006). Discriminant ecoc: a heuristic method for application dependent design of error correcting output codes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(6):1007–1012.
- Reunanen, J. (2003). Overfitting in making comparisons between variable selection methods. *Journal of Machine Learning Research*, 3:1371–1382.
- Riyaz, P. A. and Varghese, S. M. (2016). A scalable product recommendations using collaborative filtering in hadoop for big data. *International Conference on Emerging Trends in Engineering, Science and Technology (ICETEST-2015)*, 24:1393–1399.
- Rodrigues, F. and Ferreira, B. (2016). Product recommendation based on shared customer’s behaviour. *International Conference on Enterprise Information Systems/International Conference on Project Management/International Conference on Health and Social Care Information Systems and Technologies, CENTERIS/ProjMAN / HCist 2016*, 100:136–146.
- Safavian, S. R. and Landgrebe, D. (1991). A survey of decision tree classifier methodology. *IEEE Transactions on Systems, Man, and Cybernetics*, 21(3):660–674.
- Sanchez-Moreno, D., Gonzalez, A. B. G., Vicente, M. D. M., Batista, V. F. L., and Garcia, M. N. M. (2016). A collaborative filtering method for music recommendation using playing coefficients for artists and users. *Expert Systems with Applications*, 66:234–244.

- Schneider, C. M., Belik, V., Couronne, T., Smoreda, Z., and Gonzalez, M. C. (2013). Unravelling daily human mobility motifs. *Journal of the Royal Society Interface*, 10(84).
- Sevilla, J. and Sola, J. (1997). Importance of input data normalization for the application of neural networks to complex industrial problems. *IEEE Transactions on Nuclear Science*, 44(3):1464–1468.
- Shlens, J. (2014). A tutorial on principal component analysis. *CoRR*, 1404.1100.
- Singh, V. K., Bozkaya, B., and Pentland, A. (2015). Money walks: Implicit mobility behavior and financial well-being. *Plos One*, 10(8).
- Singh, V. K., Freeman, L., Lepri, B., and Pentland, A. S. (2013). Classifying spending behavior using socio-mobile data. *Human Journal*, 2('2):99–111.
- Sobolevsky, S., Sitko, I., Combes, R. T. D., Hawelka, B., Arias, J. M., and Ratti, C. (2014). Money on the move: Big data of bank card transactions as the new proxy for human mobility patterns and regional delineation. *2014 IEEE International Congress on Big Data*.
- Sohn, T., Varshavsky, A., LaMarca, A., Chen, M. Y., Choudhury, T., Smith, I., Consolvo, S., Hightower, J., Griswold, W. G., and de Lara, E. (2010). Mobility detection using everyday gsm traces. *Lecture Notes in Computer Science*, pages 212–224.
- Song, C., Qu, Z., Blumm, N., and Barabasi, A. L. (2010). Limits of predictability in human mobility. *Science*, 327(5968):1018–1021.
- Song, L., Kotz, D., Jain, R., and He, X. (2004). Evaluating location predictors with extensive wi-fi mobility data. *IEEE INFOCOM 2004*.
- Sun, Y., Babbs, C. F., and Delp, E. J. (2005). A comparison of feature selection methods for the detection of breast cancers in mammograms. *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, pages 6532–6535.

- Wang, C. M. and Huang, Y. F. (2009). Evolutionary-based feature selection approaches with new criteria for data mining: A case study of credit approval data. *Expert Systems with Applications*, 36(3):5900–5908.
- Webb, G. I. (2000). Multiboosting: A technique for combining boosting and wagging. *Machine Learning*, 40(2):159–196.
- Wei, J., He, J., Chen, K., Zhou, Y., and Tang, Z. (2016). Collaborative filtering and deep learning based recommendation system for cold start items. *Expert Systems with Applications*, 69(1):29–39.
- Weiss, G. and Provost, F. (2001). The effect of class distribution on classifier learning. *Technical Report ML-TR 43*.
- Yewatkar, A., Inamdar, F., Singh, R., Ayushya, and Bandal, A. (2016). Smart cart with automatic billing, product information, product recommendation using rfid and zigbee with anti-theft. *Proceedings of International Conference on Communication, Computing and Virtualization (ICCCV) 2016*, 79:793–800.
- Zweig, M. H. and Campbell, G. (1993). Receiver-operating characteristic plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, 39(4):561–577.