Spatio-Temporal Expenditure Forecasting for Bank Customers using Transactional Data

by

Kaan Telciler

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

KOÇ
UNIVERSITY

September 28, 2017

Spatio-Temporal Expenditure Forecasting for Bank Customers using Transactional Data

Koç University

Graduate School of Sciences and Engineering

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

Kaan Telciler

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

Making an accurate "next place and time" prediction for an individual bank customer's credit card expenditure opens up profit opportunities for a bank. Working together with retailers, a bank can offer targeted campaigns to customers, potentially resulting in higher profits and better customer satisfaction. We propose a data mining approach to predict whether a bank customer will make a credit card expenditure in a given geographical area and time interval. We analyze a one-year dataset of a commercial bank. Besides features related to demographic, financial and product usage information of the customer, we include behavioral features with respect to location, time and purchase category. In addition, we introduce proximity features that measure the distance between an input parameter and the past transactions of the customer in terms of location and time. By testing six data mining algorithms with respect to five performance measures, we predict the expenditures with an accuracy of 92.81% and 40.5% f1 score in datasets generated with different locations and time intervals using a sample of $10,000$ customers. We present the effects of the features in the prediction performance and observe that spatial features play the most critical role in the prediction, followed by temporal features such as time between transactions. We also conduct a sensitivity analysis on prediction radius and time interval and observe significant changes in prediction performance and feature effectiveness.

ÖZETÇE

Bir banka müşterisinin kredi kartı harcaması için "konum ve zaman" tahmini yapabilmek bankalar için büyük bir kar fırsatı yaratabilir. Bankalar, perakende satıcılar ile birlikte kişiye özel kampanyalar sunarak müşteri memnuniyetini arttırırken aynı zamanda daha yüksek karlar elde edebilir. Bu tez çalışmasında bir banka müşterisinin belirlenen konum, zaman aralığı ve günde harcama yapıp yapmayacağını tahmin eden bir yaklaşım geliştirilmiştir. Bir bankanın bir yıllık verisi analiz edilerek, demografik, finansal ve ürün kullanımı özniteliklerinin yanısıra konum, zaman ve ürün alım kategorisine göre tanımlanan davranışsal öznitelikler tahmin sürecinin içerisine entegre edilmiştir. Verilen girdi ile müşterinin geçmiş işlemleri arasındaki ilişkiyi rakamsal olarak ifade edebilmek için "yakınlık" özniteliği tanımlanmıştır. 6 farklı veri madenciliği algoritması test edilmiş ve farklı konum, zaman aralığı ve günler içeren $10,000$ müşterilik veri setlerinde $92,81\%$ doğruluk oranı ve 40.5% f1 skoru ile harcama tahmini yapılmıştır. Tahmin performansında özniteliklerin etkileri incelenmiş ve konum bazlı özniteliklerin en önemli etkiye sahip olduğu gözlemlenmiştir. Önem sırasında konum bazlı öznitelikleri harcamalar arasındaki ortalama süre gibi zaman bazlı öznitelikler takip etmektedir. Ayrıca yaklaşımımızın tahmin çapı ve zaman aralığı değişimine göre performans değişimini gözlemlemek için duyarlılık analizleri yapılmış, farklı senaryolarda tahmin performansı ve özniteliklerin önemlerinde önemli değişiklikler tespit edilmiştir.

ACKNOWLEDGMENTS

I would like to thank my advisor F. Sibel Salman first for her guidance and help in both my undergraduate and graduate studies. Besides her sharing of knowledge and experience, she contributed and supported me in all phases of my career path and provided opportunities to be involved in several studies. I would like to thank Prof. Burçin Bozkaya for his mentoring and help both in my thesis and other research studies that we conducted together. It was a great opportunity to work with him. I also thank my commitee member Assoc. Prof. Özden Gür Ali for accepting to be in my thesis committee and for her valuable suggestions and feedbacks. I would like to thank all the professors that I have met at Koc University for all their contributions.

I met with many valuable people in my university years and I would like to thank all of them for their frienship and support. Firstly, special thanks to Çağan for his perfect friendship and efforts in all the studies we worked together. It is a great chance to have a friend like him. I want to thank my oldest friends Kemal and Ömer for supporting me all the time and making my graduate years memorable. I also want to thank Ezgi and Göksu for their valuable friendship. I want to name and thank all the great people that I met in my university years: Berkay, Sezen, Fazıl, Barış, Radun, Ozan, Ilker, Deniz, Side, Hazal and many others that I can not list. I also want to thank all my friends in the office, It was great to meet them.

Last but not least, I want to thank and dedicate my thesis to my family for their endless support in all stages of my life, especially to my brother Mert.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

NOMENCLATURE

AUC: Area Under Curve

LBO: Location Based Offer

PCA: Principal Component Analysis

VIF: Variance Influation Factor

CFS: Correlation Based Feature Selection

ROC: Reciever Operating Characteristic

DT: Decision Tree

RFT: Random Forest Tree

SVM: Support Vector Machine

NN: Neural Network

NB: Naive Bayesian Classifier

LG: Logistic Regression

Chapter 1

INTRODUCTION

Availability of voluminous amounts of data creates major opportunities for organizations to mine and analyze the data for business gains. Commercial banks are one of the forerunners of organizations that value and implement business analytics practices. Banks accumulate an enormous amount of data over time. According to the report of Consumer Financial Protection Bureau in the United States, only in the first six months of 2015, there were 14.5 billion credit card transactions, which account for more than $ 1.4 trillion (CFP, 2015). This huge volume of transactions is potentiallya key profit source for the banks. Although the market is huge, competition is a challenging reality for banks. In order to achieve a more advantageous position in the market, banks try to use the data accumulated in their databases. There are many applications of data mining in banks such as fraud detection, default prediction and credit scoring.

Location information of customers accumulated in bank databases is a potential profit source for banks. Providing Location Based Offers (LBOs) to customers creates many opportunities for banks to increase their profits by uncovering cross-selling opportunities, increasing customer loyalty and providing new services to retailer's (Cog, 2015). Some examples of using the location and time information of customers are providing higher limits for given locations and times, providing extra bonuses by cooperating with retailers or making special events on predicted locations at predicted times. In order to offer LBOs or other related location based services, banks need to understand and use the locations of customers for predictions. Sources of location information include credit card transactions, ATM transactions and branch visits of

a customer. A new location source, mobile applications of banks, is introduced in the recent years which enables banks to track the location of customers. By using these sources, banks offer real-time LBOs to their customers. Predicting the locations, times and days of the expenditures of customers accurately provides an opportunity to better organize the offers and develop strategic plans for the future offers.

In this study, we offer an approach to predict whether a bank customer will make a credit card expenditure in a given geographic area (specified by radius), time interval and day. We work with a dataset consisting of ten thousand customers sampled from the database of a large bank, according to the number of transactions that contain location information. The first step of the study is analyzing the data and generating meaningful features for spatio-temporal expenditure forecasting. We define two types of features; constant and input dependent. Constant features are calculated based on customer characteristics or transactions regardless of any external condition or input, whereas input dependent features are calculated for some given input and this explains customer behavior in relation to the given input. Besides the features that show the demographic, financial and product usage information of the customer, we include behavioral features that indicate diversity, loyalty and regularity of customer behavior with respect to location, time and purchase category. We also introduce proximity features which measure the "distance" between the input and all transactions of the customer in location, time and day dimensions. We use 63 features in our forecasting process, where 51 of them are constant and 12 of them are input dependent features.

We offer a system design to include spatio-temporal expenditure forecasting in real-world applications. To develop a model that performs well for different inputs, we generate 63 datasets containing different locations, time intervals and days. For model training and analysis processes, we work with samples from these datasets. In order to obtain the best performance, we try 6 data mining algorithms: decision trees, random forest trees, support vector machines, neural networks, naive bayesian classifiers and logistic regression. We compare these algorithms with respect to five performance measures, which are accuracy, recall, precision, F1 score, area under

curve (AUC).

We also investigate the relative effect of features on the prediction performance to understand the main drivers behind an expenditure in a given location, time interval and day. We provide a detailed analysis of proximity and behavioral features, and change of algorithm performances with respect to the number of features. Finally, we provide a sensitivity analysis of our approach for different radius and time ranges.

This thesis is organized as follows. In Chapter 2, we review the literature on feature selection and extraction methods on data mining, data mining applications involving bank data, mobility features and spending behavior, next place prediction and spatio-temporal forecasting. In Chapter 3, we explain the data we have worked with and describe the features used in forecasting. In Chapter 4, we show the details of our implementation and experimental setup. In Chapter 5, we present the results we obtained in the computational studies and discuss the inferences from these computational studies. Finally, in Chapter 6, we provide our concluding remarks and discuss possible extensions of the study.

Chapter 2

LITERATURE REVIEW

In this section we present a review of the related studies in the literature under five topics:

- Feature Extraction, Generation and Selection
- Data Mining Applications with Bank Customer and Transaction Data
- Mobility Features and Spending Behavior
- Next Place Prediction
- Spatio-Temporal Forecasting

The literature review section is concluded with a discussion of our contributions to the literature.

2.1 Feature Extraction, Generation and Selection

A typical data mining study follows the order: (1) feature extraction, generation or selection, (2) implementation of data mining algorithms with the chosen features, and (3) performance analysis. The first and one of the most critical steps of data mining is feature extraction, generation and selection. The main idea behind this step is to create meaningful features from the raw data and choose the most effective ones to optimize the performance and accuracy of the model. In this section, common feature selection and extraction methods, and the features related to our research topic will be explained.

Bank databases contain huge amount of data of their customers and transactions. In order to obtain higher performances from the models used in these studies, appropriate features should be generated from this mass of data. Nie et al. (2011) list many features which can be used in bank data studies. They implement a feature selection approach to include 95 out of 135 features in their classification model. These features are grouped into four: personal variables, card basic variables, risk related variables and transaction variables. In order to eliminate features, they use a simple approach that calculates the multicollinearity with the help of variance influation factor (VIF) and eliminate features with large VIF values.

Feature extraction methods are used to extract new features from given features in order to reduce the dimension of the problem. A widely used feature extraction algorithm is Principal Component Analysis (PCA). PCA is a method to extract the dominant patterns in a data matrix (Wold et al., 1987).

Feature selection methods are used to choose the most related or effective ones from a set of features. Chandrashekar and Sahin state that feature selection methods provides us to reduce the computation time, improve the prediction accuracy and a better understanding of data in machine learning. They further state that more information is not always good for data mining approaches, since this may lead to a decrease in algorithm prediction performance. There are many feature selection algorithms in the literature and there are performance, simplicity, computational requirement and many other differences between these algorithms, but applying feature selection algorithms is always beneficial (Chandrashekar and Sahin, 2014).

There are three types of feature selection methods: filter methods, wrapper methods and embedded methods. Filter methods rank the features according to a criterion and chooses the features which are above a treshold. Wrapper methods basically try to find the best subset of features according to the classification performance. Since evaluating all subsets is an combinatorial problem, different approaches are used to determine these subsets. Embedded methods interact with the classifiers for feature selection (Chandrashekar and Sahin, 2014).

There are several well performing and widely used feature selection algorithms. One of the most common feature selection approaches is proposed by Hall (1999). In this study, a feature selection approach for supervised machine learning problems on the basis of correlation between features is developed. The proposed Correlation based Feature Selection (CFS) algorithm couples a feature evaluation formula with an appropriate correlation measure and a heuristic search strategy. In most cases, it is observed that CFS eliminates approximately half of the features while increasing the prediction accuracy. CFS is basically an improved filter method.

Another approach to feature selection is formulating the problem as a multi-objective optimization problem. Wang and Huang (2009) propose new criteria for single and multi-objective evalutionary feature selection algorithms. They observe that the new criteria improve the performance and accuracy of feature selection algorithms.

2.2 Data Mining Applications with Bank Customer and Transaction Data

Valuable information can be derived from bank transaction data using data mining techniques. Banks use data mining for many purposes such as churn prediction, credit scoring, default prediction, fraud detection and customer segmentation. Various data mining techniques are applied in order to obtain the best performance. In this section we give a brief overview of the development of data mining studies using bank data and since it is a very large research area, our focus is mainly on benchmark studies to present commonly used data mining algorithms and their performances.

The studies in this area are mostly about classification of bank customers with the essential idea of developing a model that aims to predict the class of a customer with the highest accuracy. The simplest approach for classifying customers is to classify them as "good" or "bad" and Orgler made this classification by using regression, which appears to be the first data mining algorithm implemented bank data (Orgler, 1970). He reported the most related features which makes a customer good or bad and observed a prediction accuracy of more than 80% for bad customers. The author also

concluded that the factors influencing the prediction are the characteristic features of customers, not the features of the banks, which shows the importance of the analysis of an individual's data.

Over time, with the improvement of data mining algorithms and computational capabilities, the techniques used for analyzing bank data have diversified. Logistic regression for classification of bank customers is implemented by Wiginton (1980), which is an extension of linear regression. However, logistic regresion did not a bring significant improvement over linear regression in this study. Henley (1994) also claims that logistic regression does not provide any improvement compared to linear regression. Another approach for bank customer classification is by means of linear programming (LP). The flexibility of the LP approach makes it potentially a desirable alternative (Fred and Glover, 1981). The definition of the objective function can be changed and specific constraints of the problem can be involved in the formulation easier than other data mining methods. An example of the most common approach involves defining the objective as minimization of misclassification costs (Fred and Glover, 1981). By giving different objective criteria, different classification tasks can be casted within a mathematical programming framework (Hand and Henley, 1997).

In what follows, we review, from a methodological point of view, some survey articles that focus on different decision making applications.

Credit and behavioral scoring helps decision makers in credit approval decisions. Thomas (2000) surveys various credit and behavioral scoring methods. Both statistical and Operations Research approaches for classification are presented in the article. Overall linear regression and recursive partitioning algorithms (RPA) are the best performer's compared to logistic regression, linear programming, neural networks and genetic algorithms according to their classification accuracies. The success of ongoing studies for forecasting financial risk is pointed out in this paper which increased the attention to data mining studies for banks.

Different from classification problems, a benchmark study for forecasting the total volume of a bank's transaction data is provided (Maass et al., 2003). It includes

Yule-Walker AR(L), wavelet method, ensemble methods and two standart time series analysis tools. In order to show the improvement in forecast quality, the authors compare the results of their study with a test forecast named "mean" which is the simplest approach that takes the mean value of previous months as the forecast of the next month. Their forecasts are much better than the intuitive "mean" forecast. The study concludes that scientifically motivated forecasts give better results against intuitive ones.

Baesens et al. (2003) show a detailed benchmark of various data mining algorithms used for credit scoring. The study uses eight real-life credit scoring datasets. According to their results, Neural Networks and RBF LS-SVM algorithms perform best for the benchmark instances with respect to classification accuracy and area under the curve performance measures. However, results in (Baesens et al., 2003) show that the difference between the performance of applied algorithms is not significant.

Default is a risk for banks and predicting the defaulting probability of a customer can be very beneficial. (Yeh and Lien, 2009) compare the performance of 6 algorithms for predicting default probabilty of credit card clients and report that artificial neural network is the best performing algorithm . They compare the algorithms according to two different measures. The first is the classification accuracy, where neural networks perform best, and followed by classification trees. The second comparison is in the prediction of real default probability, and in this case, neural networks has the highest explanatory ability. A novel approach to estimate this probability named "Sorting Smoothing Method" is proposed.

Fraud detection is critical for banks in order to sustain customer satisfaction and to prevent losses. Data mining approaches are typically used to predict fraud. A detailed review of fraud detection studies is provided by (Ngai et al., 2011). The author list the main data mining techniques used to detect and classify fraud as logistic models, neural networks, the Bayesian belief network and decision trees. Another study (Bhattacharyya et al., 2011) provides a comparison between performances of support vector machines (SVM), random forests (RF) and logistic regression by using

a real-life dataset. According to top-K performance measure, RF is the best performing one. Top-K performance which gives the proportion of fraud among top-K cases that are ranked by scores came from the model. According to prediction accuracy measure, three algorithms performs similarly.

Data mining using bank data is a wide research area and there is still room for improvement. There is no single method that outperforms other methods and the performance changes with data and problem. A brief summary of problems and algorithms used are represented in Table 2.1. In order to obtain the best performance, many alternative approaches should be tested for a problem.

2.3 Mobility Features and Spending Behaviour

Locations associated with data points provide an opportunity for mobility analysis and location prediction. Studies mainly aim to understand mobility patterns from the location data, analyze their relation with other features and predict the next-location of users. Three main data sources are used in mobility studies; Telecom data, social media data and bank transaction data. The common point of these data sources are that all of them record the location of the actions of users.

Mobility studies began with the studies that analyzed and tried to explain human trajectory patterns. The study of (Gonzalez et al., 2008) showed that human trajectories show a high degree of temporal and spatial regularity by investigating 100,000 mobile phone users for 6 month. They also stated that individual travel patterns can be collapsed a single spatial distribution. The outcome of this study provided a base for many studies about human mobility. Another important question is that what is the limits of the predictability of human mobility. The study of (Song et al., 2010) investigated the limits of the predictability of human mobility. This study showed that human mobility has a prediction potential of 93%. Their study is based on a 3-month-long mobile phone data of 50,000 individuals. Both of these studies pointed out that developing models to predict human mobility will have a positive impact on the well-being of society.

Article	**Problem**	**Methods Used**
Orgler1971	Credit Scoring	Linear Regression
Wiginton1980	Credit Scoring	Logistic Regression
Henley1995	Credit Scoring	Logistic Regression
Freed1981b	Credit Scoring	Linear Programming
Hand1981	Credit Scoring	Linear Programming
Thomas2000	Credit and Behavioral Scoring	Linear Regression Logistic Regression Linear Programming Classification Trees Neural Networks Genetic Algorithm
Maass2003	Transaction Amount Forecasting	Yule_Walker AR(L) Wavelet Method Ensemble Method Time Series Analysis Tools
Baesens2003	Credit Scoring	Support Vector Machines Neural Networks Logistic Regression Linear and Quadratic Discriminant Analysis Linear Programming Bayesian Network K-Nearest Neighbour Decision Tree
Yeh2009	Default Prediction	K-Nearest Neighbour Logistic Regression Discriminant Analysis Naive Bayesian Neural Networks Classification Tree
Bhattacharyya2011	Fraud Detection	Support Vector Machines Random Forests Logistic Regression

Table 2.1: Data Mining Methods in Articles

After it is understood that the human mobility contains valuable information, studies which aim to interpret the mobility behaviours have been conducted. The raw mobility data is accumulated in some features which works as a meaningful information about all the mobility data of an individual. One of the most used measures is the diversity measure which aims to explain how diversified the behaviours of an individual is. The diversity measure is defined in (Eagle et al., 2010), in order to indicate the diversity of the relationships and the mobility of an individual. The diversity formula in this study uses Shannon's entropy which is one of the most important concepts of information theory, used to calculate the expected value of an information in a message (Shannon, 2001). This formula is used to define the diversity indicator, which basically measures the proportion of a type of action in all actions and sum them up with entropy approach. They used mobile phone data in order the obtain the relationships and mobility of individuals.

In addition to diversity, loyalty measure is used in (Singh et al., 2013). Loyalty shows how loyal an individual is to his top vendors in this study. They used a dataset introduced in (Aharony et al., 2011), containing 52 adults (26 couples) which has a wide range of data and features. Also overspending is introduced as a feature in (Singh et al., 2013). By using social behaviours observed from mobile phones, they showed that the dissimilarity in a couple's social behaviours is correlated with spending diversity, loyalty and overspending using Naive Bayes method. The authors also found that using couple level features can increase the predictive power and provide interesting insights about the spending behaviours.

In another study (Singh et al., 2015), beside the diversity and loyalty features, regularity feature is introduced as a mobility feature and 30% to 49% improvement in financial outcome prediction is observed. Regularity measures the similarity between an individual's behaviour over short and long terms. Four types of bins are defined by spatial and temporal dimensions, which enables incorporating both dimensions seperately in the model.

Another study using credit card transactions data showed that mobility of the for-

eigners demonstrates a strong dependence on the distance to their origin (Sobolevsky et al., 2014). They partitioned the dataset into regions in order to analyze the mobility level in regional level. Further, they introuced a novel probabilistic approach to define a network between pairs of locations across country to indicate the relation between two locations in a more appropriate way.

As explained above, utilizing various mobility features in data mining is an expanding research area which has a strong relation with spending behaviour of individuals. According to these studies we can infer that involving mobility behaviour in models may provide a better predictive and explanatory power.

2.4 Next Place Prediction

The ability to track the locations of individuals and studying mobility data and mobility features provides an opportunity to predict the next place of individuals. Next place prediction is an expanding research area which basically aims to predict the next location of an individual by using previous locations and other information about the individual. There are two main approaches for next place prediction: Markov chains and machine learning algorithms.

POIs (Point of Interest) are introduced by (Ashbrook and Starner, 2003) in order to learn significant locations of a user. POIs are detected by clustering user locations with the k-means clustering algorithm. The location data is obtained from GPS. Markov chains are used in this study to predict the next locations. In (Gambs et al., 2012) n-MMC (Mobility Markov Chain) is used to predict the next location using previous n locations and 70% to 95% prediction accuracy is obtained. In this study, the locations are also clustered into POIs. The main difference between these two studies is the clustering approach they used. However, in (Krumme et al., 2013), it is stated that even perfectly observed transition probabilities do not perform better than the assumption that the individual will go to the place where he visits most. One outcome of this study is that shopping behaviours are highly predictable in longer time scales.

In addition to Markov chain approaches, machine learning algoritms are used for next place prediction. It is found in (Noulas et al., 2012) that supervised methodology based on the combination of multiple features offers the highest level of prediction accuracy. The authors used M5 model trees to rank the potential next locations according to their visit probabilities. They introduced several new mobility features to the literature and they categorized them with three main titles: user mobility features, global mobility features and temporal features. They used Foursquare check-in data in their study.

In order to understand the importance of mobility data in next place prediction, (Sadilek et al., 2012) reported an increase of approximately 30% in prediction accuracy after adding mobility features to the model. Another important outcome of this study is that, with the increase of friends' mobility in the model, the prediction accuracy increases significantly. In this study, they compared several Bayesian Models for prediction and observed that Dynamic Bayesian Networks performed best.

Hoeffding Trees are used to predict the next location of a user in (Gomes et al., 2013). In this study NextLocation framework is developed to apply different machine learning algorithms in a structured way for next place prediction. Also a benchmark of several algorithms is provided. It is reported that ANN and decision trees performed best in the case defined in the paper. They also reported that the regularity in user movements causes a great variety in prediction accuracy across users.

To the best of our knowledge, studies in the literature is limited to predicting the next location of a user given a number of previous locations. We have found no studies that predict the timing of being at the next location.

2.5 Spatio-Temporal Forecasting

Since some databases contain time and location information, data mining studies extended to time and location domains. Temporal data mining studies' purpose is analyzing the events ordered by one or more dimensions of time and spatial data mining is the multi-dimensional equivalent of temporal data mining based on location

domain (Roddick and Spiliopoulou, 1999). Main output patterns of spatial data mining are predictive models, spatial outliers, spatial co-location rules and clusters (Shekhar and Xiong, 2008). For the temporal data mining, most common operations are discovery of association rules, classification, clustering and prediction (Antunes and Oliveira, 2001). An overview of both research areas are prepresented in (Shekhar and Xiong, 2008) and (Antunes and Oliveira, 2001).

Spatio-temporal data mining is the extraction of knowledge, spatial and temporal relationships or patterns from spatio-temporal databases (Yao, 2003). These relationships and patterns can be used to predict the future by considering location and time basis of data, which is named spatio-temporal forecasting. There are many studies about spatio-temporal forecasting using different approaches for various usecases. In order to incorporate spatio-temporal domains in Association Rule Mining(ARM), Verhein and Chawla proposed $spatio-temporal\,association\,rules(STARS)$, which are used to describe the moves of objects between regions over time (Verhein and Chawla, 2006). They also provided spatio-temporal patterns that can be used to explain different conditions in spatio-temporal analysis. These rules provide a basis for analyzing the patterns in spatial and temporal dimensions.

The study of Cheng and Wang proposes an approach to predict the forest fires in target locations and provides a comparison of three approaches (Cheng and Wang, 2006). In this study, they tried to predict whether there will be a fire in the target location by using spatially-correlated locations. They benefit from ARIMA and neural networks in their forecasts and proposed a method named ISTIFF. ISTIFF combines ARIMA and neural networks in spatio-temporal forecasting process. They reported that, ISTIFF performs better than STIFF and ARIMA. In (Nourani et al., 2008), using neural networks in spatio-temporal groundwater level forecasting is investigated. They analyzed performance of 6 different types of neural network architectures. They used the waterlevels in 16 locations, rainfall data, mean temperature and average discharge of a river close to forecasting location as inputs to forecast the waterlevels. In the study, the best neural network and data configuration is explained for spatio-

temporal groundwater level forecasting.

Another popular usecase for spatio-temporal forecasting is traffic level forecasting in urban areas. Spatial and temporal volume characterictics are incorporated in short-term traffic forecasting process in (Vlahogianni et al., 2007), which uses a modified version of neural networks for forecasting. In this study, neural networks are fed by predictive information from multiple locations in a road network. Results of this study show that proposed approach performs better than statistical methodologies or static form of neural networks. In (Min and Wynter, 2011), a model to predict road traffic by using spatio-temporal correlations is proposed. The claim is that the level of traffic in a road is affected by the past traffic conditions of some neighbor links in the road network with a time lag function. They developed a spatial-temporal autoregressive (MSTAR) model to predict the traffic condition and obtained an accuracy that exceeds other published works in short-term.

Spatio-temporal forecasting studies in the literature focuses on systems that contain correlated locations and sequential events, and uses the information from correlated locations to forecast the condition in the target location. Two main approaches are used for forecasts; machine learning approaches and time-series forecasting algorithms. Although spatial and temporal dimensions of forecasting are investigated in many studies, there is not a study that forecasts nonsequential events for different locations and time horizons.

2.6 Our Contribution to the Literature

In this study, we propose a spatio-temporal expenditure forecasting approach for bank customers by analyzing financial transaction data. In our literature search, we could not find any study that uses bank data for spatio-temporal forecasting of customer spending. We aim to predict whether a customer will make a credit card expenditure in a given location, time interval and day. For this purpose, we generate two types of features: constant and input dependent. Constant features are calculated independent of the given input, whereas input dependent features are calculated according to

the given input and explain the relation between the customer behavior and the given input. To the best of our knowledge, the concept of input dependent feature generation is introduced in our study. Besides the known features in the literature, we introduced "proximity" feature to measure the "distance" between the given input and all transactions of the customer in location, time and day dimensions. The relation of past transactions and given input can be expressed by using proximity features, which belong to the input dependent feature type. In our data analysis, we see that proximity features play a significant role in prediction.

In the literature, spatio-temporal forecasting and next-place prediction studies feed their algorithms with sequential or related events to predict the next event. In our study, expenditure predictions are made by using constant and input dependent features. With the approach that we propose, it is possible to make a prediction for different locations, time intervals and days. We compare the performances of six data mining algorithms to obtain the best performance. As an outcome, we present the effect of the features in the prediction performance, which explains the main drivers behind the expenditure decisions of bank customers. Finally, we present the results of our sensitivity analysis.

Chapter 3

DATA AND FEATURE GENERATION

In this chapter, we explain properties of the data that we worked on and our sampling approach. We also describe the features that we used in expenditure prediction, focusing on the behavioral and proximity features which utilize customer mobility data.

3.1 Data Properties and Sampling

We worked with a data set of ten thousand customers sampled from the database of a large financial institution which contains hundred of thousands customers. Before sampling a subset, we removed the customers having missing fields in the database. In the sampling process, we choose top $10,000$ customers having highest number of location data in their transactions.

The data we worked on is a one-year data with dates ranging between 01-07-2014 and 30-06-2015. The sampled customers have $1,684,495$ transactions containing location data in total, which consist of $681,890$ ATM, $927,328$ credit card expenditures and $75,277$ branch visit data.

The data contains various types of information about the customers and we used the following attributes which provide us an opportunity to generate numerous features. Information about the customers in the database can be summarized as below.

- Customer age, gender, job, marital status, education level and estimated income
- Customer home, work and branch location (X and Y coordinates)
- Timestamp (date, hour and minute), amount and location (X and Y coordinates) of ATM transactions

- Timestamp (date, hour and minute), amount, location (X and Y coordinates) and category of credit card transaction
- Timestamp (date, hour and minute) of call center calls
- Number and type of active and owned products, limits for credit cards
- Number and amount of auto payments made by credit card
- Monthly balance of credit cards and risk scores determined by bank

3.2 Features

Because of the nature of the problem and the system design we proposed, features fall into two main categories; constant and input dependent. Constant features include customer features that do not depend on the input and are the same for any given input. Input dependent features are calculated according to the given input for every customer before forecasting. Under the two main categories we designed, features can be further grouped into 6 categories; Demographic, Financial, Product Usage, Behavioral, Proximity and Input Related features. The structure of the feature categories is shown in Figure 3.1. The number inside the paranthesis shows the number of features in that category. We calculated these features by using the data of the first 9 months and tried to forecast for the 10^{th} month.

Table 3.1: List of Features

FeatureCode	Feature Name	Category
X1	Age	Demographic
X2	Education Status	Demographic
X3	Gender	Demographic
		Continued on the next page

Table 3.1 – continued from the previous page

FeatureCode	Feature Name	Category
X4	Marital Status	Demographic
X5	Job	Demographic
X6	Income	Financial
X7	Bank Age	Financial
X8	Variance in Risk Scores	Financial
X9	Average of Risk Scores	Financial
X10	Last Risk Score	Financial
X11	Number of Active Products	Product Usage
X12	Number of Products Owned	Product Usage
X13	Total Number of Call Center Calls	Product Usage
X14	Number of Call Center Calls Last Month	Product Usage
X15	Number of Total Credit Cards	Product Usage
X16	Number of Other Banks' Products	Product Usage
X17	Bank Credit Card Limit	Financial
X18	Total Credit Card Limit	Financial
X19	Average Interval Between Transactions	Product Usage
X20	Variance of Intervals Between Transactions	Product Usage
X21	Average Balance of the Account	Financial
X22	Last Known Balance of the Account	Financial
X23	Average Amount of Auto Payments	Financial

Continued on the next page

Table 3.1 – continued from the previous page

FeatureCode	Feature Name	Category
X24	Days Remaining to Last Payment Date	Product Usage
X25-X33	Monthly Average Expenses	Financial
X34	Variance of Expenses	Financial
X35	hourly_diversity	Behavioral
X36	hourly_loyalty	Behavioral
X37	hourly_regularity	Behavioral
X38	daily_diversity	Behavioral
X39	daily_loyalty	Behavioral
X40	daily_regularity	Behavioral
X41	spatial_radial_diversity	Behavioral
X42	spatial_radial_loyalty	Behavioral
X43	spatial_radial_regularity	Behavioral
X44	spatial_cluster_diversity	Behavioral
X45	spatial_cluster_loyalty	Behavioral
X46	spatial_cluster_regularity	Behavioral
X47	category_diversity	Behavioral
X48	category_loyalty	Behavioral
X49	category_regularity	Behavioral
X50	Distance to Home	Input Related
X51	Distance to Work	Input Related

Continued on the next page

Table 3.1 – continued from the previous page

FeatureCode	Feature Name	Category
X52	Distance to Branch	Input Related
X53	Radial_credit	Proximity
X54	Hourly_credit	Proximity
X55	Daily_credit	Proximity
X56	Radial_atm	Proximity
X57	Hourly_atm	Proximity
X58	Daily_atm	Proximity
X59	total_credit_card_transactions	Financial
X60	total_atm_transactions	Financial
X61	r	Input Related
X62	TW	Input Related
X63	d	Input Related
Y	Prediction Class	Input Dependent

A detailed list of the features and their categories are given in Table 3.1. X_1 is the customer's age, X_2 is a categorical feature that indicates the education status of the customer, X_3 shows whether the customer is male or female, X_4 is a categorical variable that shows the marital status of the customer, X_5 shows the customer's job type. X_6 indicates the income of the customer estimated by the bank, X_7 shows the number of years that the customer is contracted to the bank, X_8 and X_9 denote the variance and the mean of the 9 risk scores of the customer determined by the bank at the end of every month. X_{10} shows the risk score at the end of the 9^{th} month. X_{11}

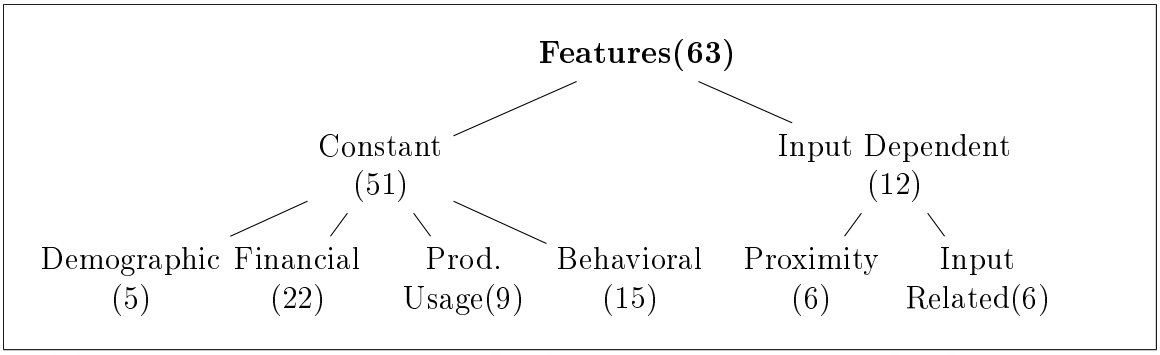

Figure 3.1: Feature Types

and X_{12} show the number of active products and all products, respectively. X_{13} is the total number of call center calls and X_{14} is the number of call center calls in the last month. X_{15} denotes the number of credit cards of the customer belonging to the bank and X_{16} indicates the number of other banks' products of the customer. X_{17} and X_{18} are the total limit of credit cards of the bank and all banks, respectively. X_{19} and X_{20} measure the credit card usage frequency of a customer by indicating the mean and the variance of the intervals between the transactions of the customer. X_{21} and X_{22} show the average and the last known balances of the bank customer. X_{23} is the average amount of auto payments of the customer. In order to get the number of remaining days until the last payment date from the end of the 9^{th} month, X_{24} is defined. X_{25} to X_{33} show the average credit card expenses in months 1 to 9 respectively and X_{34} is the variance of the expense amounts. From X_{35} to X_{49} behavioral features are defined. These features are explained in Section 3.2.1 in detail. X_{50}, X_{51} and X_{52} are input related features, which show the distance between input location and home, work and branch location as the real world distance. From X_{53} to X_{58}, proximity features are defined. Proximity features are calculated according to the given input as explained in Section 3.2.2. X_{59} and X_{60} show the total number of credit cards and ATM transactions, respectively. X_{61}, X_{62} and X_{63} are determined by the input and indicate the input radius, hours between the start time and end time of the input, and an integer that represents the day of the week, respectively. Y is the output, which

is the class that we want to predict. It is a binary feature that shows whether the customer made an expenditure in the given input radius, time range and the day in the 10^{th} month. Feature Y is also dependent to input. All of the features X_1 through X_{63} are used in the models to forecast Y.

3.2.1 Behavioral Features

Behavioral features are calculated from the transactions of the customer and defined to generate meaningful values from a mass of transactions. Behavioral features mainly explain three types of customer behaviour: diversity, loyalty and regularity. We investigate each of these characteristics in 5 different dimensions; hourly, daily, categorical, radius based spatial and cluster based spatial. This leads to 15 behavioral features (X_{35} to X_{49}).

- **Diversity:** Diversity is used to measure the variability of customer transactions over time, location and category. Diversity is calculated for various "bins", which are explained below in detail. For a given bin definition, the fraction of customer i's transactions occuring in bin j is calculated and denoted with p_{ij}. Then the diversity value of customer i is calculated by the formula below, where N is the number of bins and M is the number of non-empty bins.

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

  Diversity is a value between 0 and 1 and larger values show a higher diversity. As the values get closer to 1, the transactions of the customer spread across different locations, time or category.

- **Loyalty:** Loyalty is the percentage of the transactions that occur in the top three bins of a customer. We used three for loyalty calculation, but this number can be changed for different cases. Loyalty score of customer i is calculated by the formula below, where f_i is the fraction of transactions in the top three bins

to transactions in all bins.

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

Loyalty is also between 0 and 1. Higher loyalty means that the customers' transactions occured more in the top three bins.

- **Regularity:** Regularity measures how regular are the characteristics of transactions belonging to the customer over shorter and longer periods. We used 3 months as a short period and 9 months as a long period. The regularity formula uses the Euclidean distance between the vectors that contain the diversity and loyalty values of short and long periods. Regularity of customer i is denoted by R_i and calculated as:

$$R_i = 1 - \frac{\sqrt{(D_i^3 - D_i^9)^2 + (L_i^3 - L_i^9)^2}}{\sqrt{2}} \tag{3.3}$$

In Formula 3.3, D_i^3 and D_i^9 are the diversity scores of customer i for the first 3 and 9 months, respectively. L_i^3 and L_i^9 are the loyalty scores of customer i for the same time ranges. Higher R_i means that customer i's transaction behaviors are similar for shorter and longer periods.

Bin Definitions We used five bin definitions to calculate the behavioral features. Bins are defined on a location, time and category basis.

- Temporal-Hourly: There are 24 hourly bins, where each represents a hour in a day.

- Temporal-Daily: Each day in a week is a bin, so there are 7 daily bins.

- Spatial-Radial: Radial bins are circles with different radiuses and same predefined centers. We used the center of gravity of the coordinates of all transactions of a customer as the center and $1, 3, 5, 15, 50, 150$ and 500 km as radii for these circles.

- Spatial-Cluster: Because the locations in data spread to different regions, we grouped them into 50 clusters and every cluster behaves like a bin. Because the transactions are not spread homogeneously and the data we used covers a large area, cluster based bin is a more appropriate approach than grid based approach. We used k-means clustering algorithm in Python Scikit library. Cluster centers are shown in Figure 7.1.

- Categorical: There are 140 expenditure categories and each of them is defined as a bin.

3.2.2 Proximity Features

Transactions of a customer have different locations, times and dates which should be included in forecasting process. We defined proximity features to explain the relation between input and transaction locations, times and dates, and to include many transactions in a small number of features. Proximity measures the proximity of all transactions of a customer to a given input in 3 dimensions; radial distance, time and day. Proximity measure sums up the difference between the location, time and day of the transaction and given input.

All credit card and ATM transactions carry location, time and date information. Proximity (P_{ij}) of the transactions of customer i to input j is the fraction of the sum of the scores s_{jt} of each transaction t according to input j in the set of all transactions of customer i, T_i, to the number of transactions M. Proximity is calculated as shown in 3.4.

$$P_{ij} = \frac{\sum_{t \in T_i} s_{jt}}{M} \tag{3.4}$$

We used six different proximity features. Radial, hourly and daily proximity of credit card transactions and radial, hourly and daily proximity of ATM transactions. The score of each transaction is calculated differently for each of radial distance, time and day separately. The common point of all scores is that, the score of a transaction increases proportional to how far the transaction is from the input. The calculation

approaches are dependent on the input. For a given input (x, y, r, TS, TE, d), scores for each dimension are calculated as explained below:

- **Radial Distance Score:** Radial distance score is calculated using the distance between the input coordinates (x_j, y_j) and transaction coordinates (x_t, y_t). Radial distance is the lowest integer n that satisfies $d_{jt} < r.2^{(n-1)}$, where d_{jt} is the distance between the input and the transaction, and r is the input radius.

- **Hourly Score:** Hourly score is calculated using the transaction time TT and the start and end of the time interval in the input, TS and TE, respectively. For a given TS and TE the length of the time interval $TE - TS$, the hourly score returns how many units of the time interval length the transaction time far from the input time interval as an integer. For instance, let $TS = 4, TE = 6$ and $TT = 9$, length of time interval is $TE - TS = 2$ and the equation $6 + 1 \times 2 < 9 < 6 + 2 \times 2$ holds. That means that the transaction occurred two time intervals far from the input time interval. So the hourly score of the transaction is 2.

- **Daily Score:** Daily score is the number of days between the day input d and the day of transaction. Daily score can be between $[0, 3]$. For instance, if the input day d is Monday and transaction day TD is Wednesday, then its 2 day away from the input day.

In order to observe whether there is a difference between the effect of older and more recent transactions, a time adjustment parameter α can be defined. If the transaction occurred n months before the month that we want to forecast for, then the score is multiplied by α^n to change the contribution of the transaction to the proximity measure. Proximity with time adjustment is shown in 3.5.

$$P_{ij} = \frac{\sum_{t \in T_i} s_{it}(\alpha^n)}{M} \tag{3.5}$$

An increase in α value increases the effect of the previous transactions. Clearly, α can be adjusted for different cases.

Chapter 4

IMPLEMENTATION AND EXPERIMENTAL SETUP

In order to analyze the performance of various algorithms and the effect of features in prediction performance, we conduct experiments and a sensitivity analysis. We execute all of the experiments with a sample of 10,000 customers.

In this chapter, an overview of this process is represented. We also explain our dataset generation approach and the way we handle minority classes in our dataset. Finally we provide brief explanations of the algorithms we use in this study.

4.1 Overview of Our Forecasting Approach

In this study we propose an approach to make spatio-temporal expenditure forecasting for bank customers. A flowchart of this approach is shown in Figure 4.1. Input is given to the system by the user. For the given input, input dependent features are calculated using the input and past transactions data of customers obtained from the data warehouse. Calculated input dependent features and constant features for customers constitutes the dataset. This dataset is used by the trained forecasting model to obtain forecasts.

An input is a vector with six elements: (x, y, r, TS, TE, d) where x and y are the longitude and the latitude of the location; radius r is the distance range to predict; TS and TE are the bounds of the time interval; d is the day that we want to predict for. An input $(29, 41, 10, 12, 16, 3)$ means that we want to predict whether a customer will make a credit card expenditure in a range of 10 kms from location with coordinates $(29, 41)$ between $12:00$ and $16:00$ in day 3, which is Wednesday. The forecasting model tries to predict whether a customer will make an expenditure inside the input.

Our computational platform has the following properties. We used PostgreSQL for

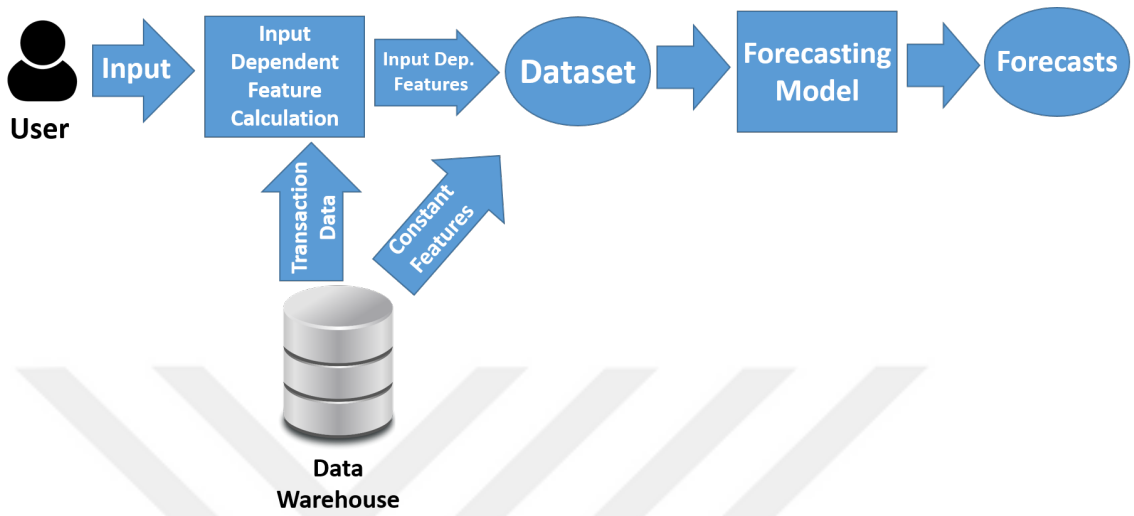

Figure 4.1: Flow of the Spatio-Temporal Forecasting Approach

data management, Python language for programming and (Scikit) library to implement the algorithms. Our computational experiments were performed on a computer with 64-bit Intel(R) Core (TM) i7 CPU X 980 @ 3.3 GHz and 12GB of memory.

4.2 Dataset Generation

We used the 63 features described in Section 3.2 to predict whether a customer made an expenditure according to the given input in the 10^{th} month of our data. As explained before, some features are constant and some features are input dependent. Input dependent features have to be calculated for a given input. Our aim is developing a model which performs well for different inputs. For dataset generation, we choose 7 central locations and define 3 radius and 3 time categories which are shown in Tables 4.1 and 4.2. For each location, we generate $3x3 = 9$ datasets which consists of all combinations of radius and time categories. This generation approach leads to 63 datasets in total which means a huge and homogeneus dataset containing 630.000 rows. Dataset generation process is shown in Figure 4.2.

We generated 9 datasets for each of the 7 locations. For sensitivity analysis, we used the datasets generated with analyzed category.

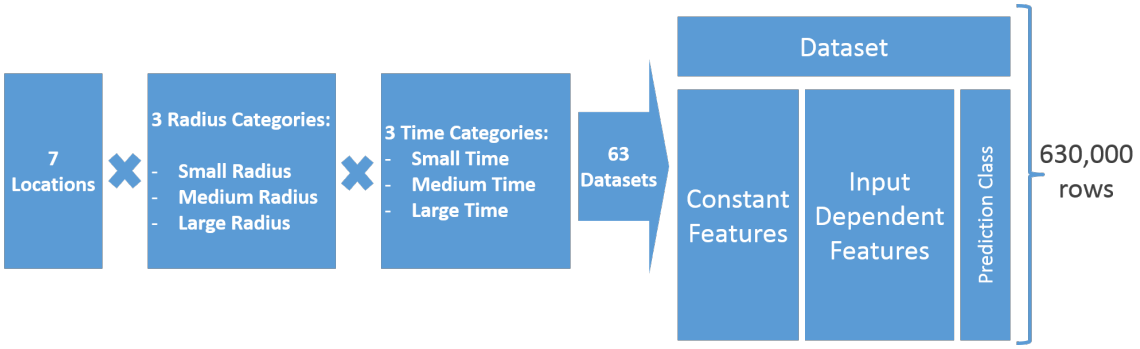

Figure 4.2: Dataset Generation Process

Small Radius	0-4 km
Medium Radius	4-10 km
Large Radius	10+ km

Table 4.1: Radius Categories

Small Time Range	0-4 hours
Medium Time Range	4-8 hours
Large Time Range	8-24 hours

Table 4.2: Time Range Categories

Descriptive statistics for the dataset used in this study are given in Table 7.1. Columns in the datasets are normalized between 0 and 1 before the model training phase. An x value in a given column C, $min(C)$ and $max(C)$ are the minimum and maximum values in the column C, respectively. Each data point is normalized and normalized value $x^{'}$ are used in model. $x^{'}$ is calculated by Equation 4.1

$$x^{'} = \frac{x - \min(C)}{\max(C) - \min(C)} \tag{4.1}$$

4.3 Handling of the Minority Class

One of the main issues in data mining is handling the minority prediction class in the solution approach. If there is an imbalance in the distribution of prediction classes in the dataset, it has to be considered for a better solution. For our problem, the percentage of ones in the prediction class is relativelty small for a given input. Decreasing the radius of prediction location, choosing less populated locations, decreasing the time range of prediction or other conditions that make the input more specific and narrow

causes having less number of ones in the prediction class. Our priority is detecting customers who will make expenditure in the given input, which are the ones in the prediction class column. Without any extra work for considering the minority class, models can not differentiate between zero and ones. In order to handle this issue, we tried three approaches.

- **SMOTE:** Instead of over-sampling by replacing the data points, SMOTE generates "synthetic" data points from the dataset to over-sample the minority class (Chawla et al., 2002).

- **Under Sampling:** Under sampling is a simple approach which removes some data points from the majority class to balance the dataset.

- **Weight Adjustment:** Data mining algorithms minimize the error function to obtain the best performance. For the imbalanced case, increasing the weight of errors for the minority classes in the error function causes the minority class to be more effective.

Approach	**Acc**	**Recall**	**F1**	**Precision**	**AUC**
Under Sampling	0.9279	**0.8333**	**0.4046**	0.2671	**0.9564**
SMOTE	0.9539	0.2686	0.2815	0.29508	0.9172
Weight Adjustment	**0.9703**	0.098	0.166	**0.5454**	0.8985

Table 4.3: Results of the Approaches for Handling Minority Class

A comparison of different minority class handling approaches is provided in Table 4.3. RFT is used in these comparisons as classifier. Our priority is predicting customers who will make an expenditure and accuracy is not a proper measure to explain the success with this priority. To explain the prediction power of the algorithms, recall and F1 score are more proper performance measures. Under sampling is the best performing approach in 3 of the performance measures and most successful

in predicting ones in the prediction class. For our problem, under sampling in the dataset generation is the best approach for datasets with relatively small number of ones.

We implement 3 fold cross-validation by dividing the dataset into 3 and use 2 of them for sampling train set and one of them for sampling test set. For the under-sampling, we choose 25% as the ratio for the minority class and used 20000 rows for training and 10000 rows for test phases.

4.4 Algorithms Tested

In order to obtain the best prediction performance, we tested six different algorithms.

Decision Trees: Decision tree is a classification model widely used in data mining, proposed in (Breiman et al., 1984). A decision tree consists of nodes and edges, where each node denotes an input variable, and an edge connects a node to its child according to the input values. A given input follows the nodes according to the input values until it reaches to the prediction class.

Random Forest Trees: Random forests are a combination of tree predictors, and use averaging to improve prediction performance and prevent over-fitting. Each tree in the forest makes a classification and the highest chosen class is the result of the random forest tree. The working principle of random forests is provided in (Breiman, 2001).

Support Vector Machines: Support vector machines (SVM) is a supervised learning algorithm which changes the dimension of data by representing the features in a different space. The representation is made to obtain the best separation of data points to make classification (Suykens and Vandewalle, 1999).

Neural Networks: Neural networks mimic human neural system, composed of interconnected neurons and nodes. Each node recieves an input signal and processes it by a function and transfers it to other nodes or external output (Zhang et al., 1998). Number of nodes and layers are the parameters of neural network which effects the performance of neural network classifiers.

Naive Bayesian Classifier: Naive bayesian classifier considers each feature separately, which means the value of a feature is not dependent on other features. Each feature inpedendently affects the probability of the data point belonging to a class. Contributions of all features determines the prediction of the classifier (Zhang, 2004).

Logistic Regression: Logistic regression is a special case of linear regression which has a binary prediction class where the outcome is determined by a logistic function that shows the probability of being a "success" rather than a "failure". If there are more than two labels in the prediction class, then multinomial logistic regression is used. A brief explanation of how logistic regression works is provided in (Yeh and Lien, 2009).

Chapter 5

RESULTS AND DISCUSSION

5.1 Performance Measures

We used performance measures widely used in data mining literature to analyze the success of our models. In our problem, we have a prediction class which can be either 0 or 1, where 1 means the customer made an expenditure in given input in 10^{th} and 0 means customer don't have any expenditure in the given input in month 10.

Confusion Matrix is a matrix that visualizes the prediction performance of an algorithm. Confusion matrix is shown below:

$$\begin{bmatrix} TN & FP \\ FN & TP \end{bmatrix}$$

- TN is the number of *True Negatives*, which are zeros predicted correctly.
- FP is the number of *False Positives*, which are zeros predicted as one.
- FN is the number of *False Negatives*, which are ones predicted as zero.
- FP is the number of *True Positives*, which are ones predicted correctly.

The following performance measures are used to analyze the performance of our models. In order to assess different aspects, we used five performance measures:

- Accuracy: Accuracy is the ratio of true predictions to all predictions. This measure gives a general idea about the performance, however it does not include

the information of prediction of classes separately. Accuracy formula is below, where N is the number of instances.

$$Accuracy = \frac{TP + TN}{N} \tag{5.1}$$

- Recall: Recall shows what percentage of ones are predicted correctly. Recall can be calculated as:

$$Recall = \frac{TP}{TP + FN} \tag{5.2}$$

- Precision: Precision is the ratio of true positive predictions to all predicted ones. The formula of precision is:

$$Precision = \frac{TP}{TP + FP} \tag{5.3}$$

- F1 Score: Although precision and recall can be used to explain the power of predicting ones, using only one of them may mislead to wrong inferences. F1 score includes both precision and recall measures in a single score and is formulated as:

$$F_1 = 2 \times \frac{1}{\frac{1}{Recall} + \frac{1}{Precision}} \tag{5.4}$$

- Area Under Curve (AUC): AUC is the area under the receiver operating characteristic (ROC) curve which is the plot of true positive rate against false positive rate for various treshold values. It is a common measure to describe and compare the performance of algorithms (Hanley and McNeil, 1983).

5.2 Algorithm Performances

In order to decide on the best performing algorithm, we tested six algorithms with the same dataset. The algorithms we implemented are: Decision Tree (DT), Logistic Regression (LG), Naive Bayesian Model (NB), Neural Network (NN), Random Forest

Tree (RFT) and Support Vector Machine (SVM). These algorithms are explained briefly in Section 4.4. To prevent over-fitting, we used three fold cross-validation in the experiments, with the same seeds for all algorithms. Grid search is made to obtain the best performing parameters for each algorithm. Table 5.1 shows the average scores of 3 folds with best performing parameters.

Alg	**Accuracy**	**Recall**	**F1**	**Precision**	**AUC**
RFT	**92.81%**	83.33%	**40.50%**	**26.77%**	**95.46%**
SVM	88.84%	82.65%	30.30%	18.57%	93.59%
DT	90.04%	67.68%	28.55%	18.09%	79.20%
LG	83.52%	**90.47%**	24.51%	14.17%	93.33%
NB	80.39%	87.55%	20.83%	11.81%	89.90%
NN	90.97%	74.14%	32.56%	20.86%	93.50%

Table 5.1: Performance of the Algorithms

In Table 5.1 a comparison of the algorithms with respect to five performance measures are shown. Random Forest Tree algorithm is the best performing algorithm according to four of the measures. Although Logistic Regression model has the highest recall, it does not perform well in terms of other performance measures. According to these results, RFT is the most suitable algorithm for our problem. SVM, NN and LG also performed well in the experiments in term of AUC scores, which can be alternative models for the problem. DT and NB are not as successful as to other algorithms. It is observed that the expenditure of a customer in a given location, time window and day can be predicted with an accuracy of 92.82% and with an F1 score of 40.5% which shows the ability of predicting ones in the model. Also we can detect 83.33% of the people who made expenditure in the given input.

5.3 Effects of Features

In this section, we investigate the critical features in our spatio-temporal expenditure prediction approach. The models are trained with datasets containing 63 features, however we need further investigation and experiments to understand the effect of the features on the prediction performance. At the first step, we looked for the effect of constant and input dependent features. Figure 5.1 shows different model performances that aims to predict the same classes, which are model with only constant features and model with both constant and input dependent features. Results of the best performing algorithm and the parameters are shown in the Figure 5.1.

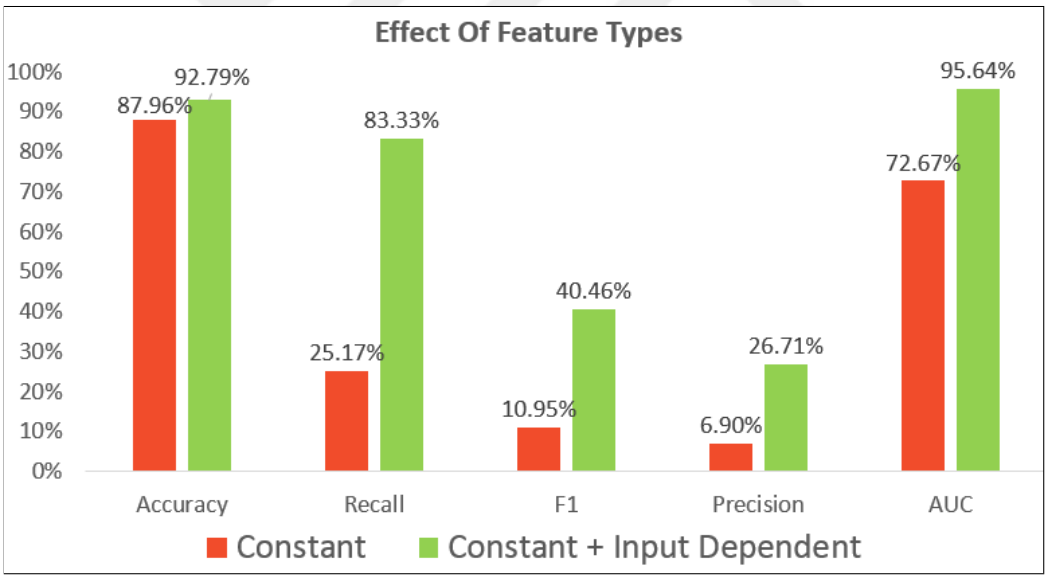

Figure 5.1: Effect of Input Dependent Features

We observe significant increase in the model performance in terms of all performance measures after adding the input dependent features. According to AUC scores, adding the input dependent features increases the prediction power of the model by 32%. This result shows that in order to establish an effective prediction system, adding the input dependent features is essential.

For a more detailed investigation, we used the *feature_importances* function defined in Scikit library. This function calculates the "gini-importance" of the features

(Breiman et al., 1984). Most important 15 features and their gini-importance values for RFT are shown in Figure 5.2. Negative values mean negative correlation between the feature and the prediction class.

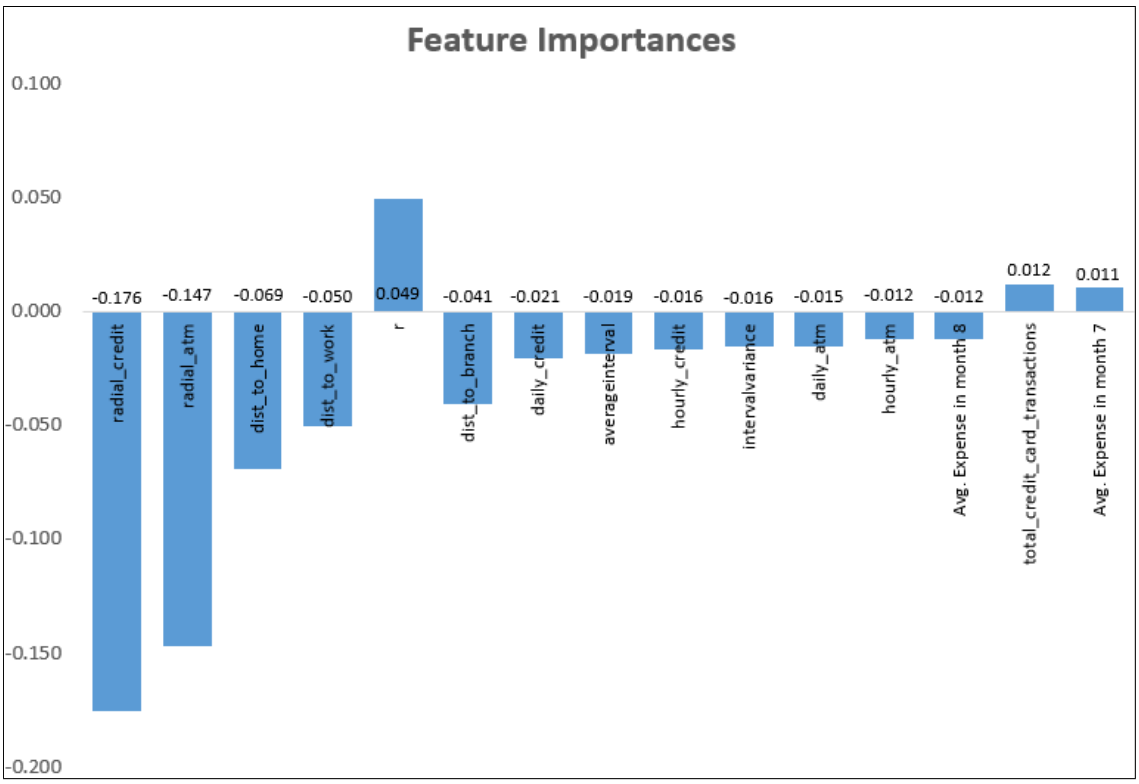

Figure 5.2: Feature Importances

The main inferences from these results are:

- The six important features are spatial features and related to the spatial dimension of the problem. Spatial features play the most critical role in the prediction.

- Customers have more tendency to spend in locations close to their home, work or their bank branch.

- Hour and day related proximity features have close effects on the prediction performance, but not as significant as spatial features.

- Intervals between transactions are effective in the prediction power. Customers with lower average interval between transactions and lower transaction interval variance tend to make an expenditure inside the given input.

Among the input dependent features, there are two feature types: Proximity and Input Related features. As shown in Figure 5.1, input dependent features have a significant effect in the prediction power and the proximity features that we proposed in this study are such feature types.

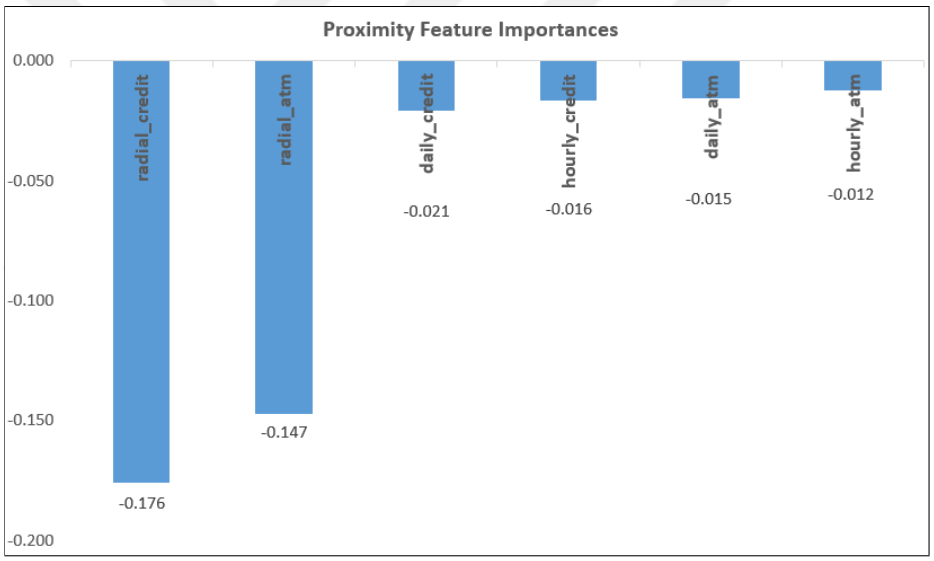

Figure 5.3: Proximity Feature Importances

Figure 5.3 shows the comparison of these proximity features. According to these results:

- All of the proximity features are negatively correlated with the prediction class, which means customers with lower proximity score are more likely to make an expenditure inside the given input.

- Radial proximity scores are the most important features and negatively correlated with the prediction class. Customers with lower radial proximity scores are more likely to make an expense in the given input.

- Proximity scores calculated from credit card transactions perform better than the proximity scores calculated from ATM transactions.

- Day related proximity features have an effect on the prediction performance slighty higher than hour related proximity features. However, both day and hour related proximity features are much less effective in the prediction power compared to spatial related features.

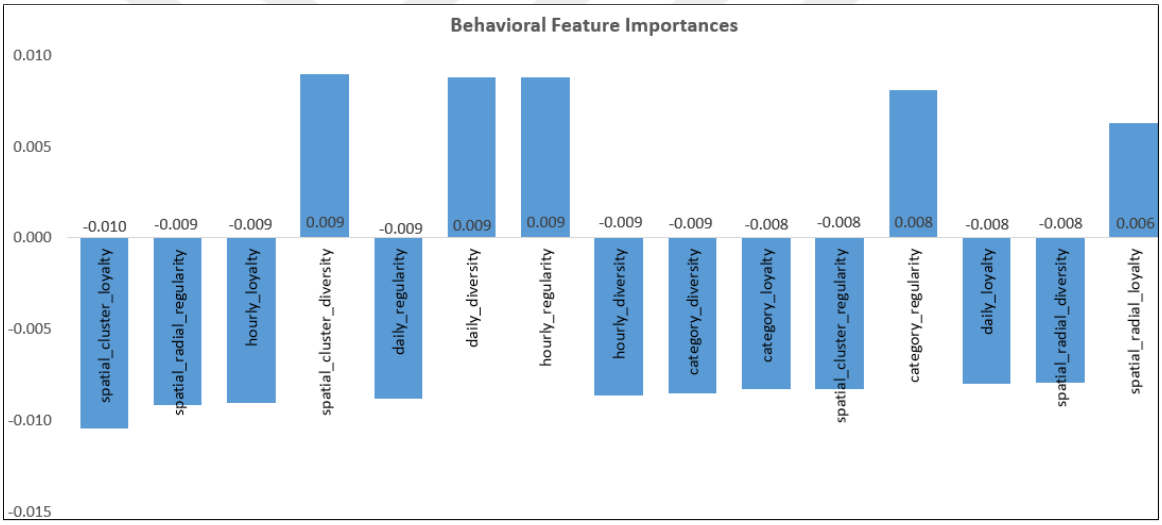

Figure 5.4: Behavioral Feature Importances

Behavioral features accumulate many transactions of a customer into meaningful scores that explain some behavioral characteristics. They have a relatively small effect on the prediction power compared to the proximity features. We see that using bins generated with clustering is more useful for spatio-temporal expenditure forecasting than radial bins in spatial aspect. There is not a significant difference in the importances between the behavioral features.

We also analyzed the change of the performance measures with the change of the number of features used. We used the "select K-Best" method in Scikit library, which uses chi-squared statistics between features and prediction class for feature selection. We used RFT as the classifier and varied K, which is represented in the

X axis in Figure 5.5, to see the performance changes. Analysis results are shown in Figure 5.5. There is an increase trend in all of the performance measures, with the increase in the number of features. Another important inference from these results is that by using only one feature, we can obtain approximately 90% percent of the maximum prediction performance in terms of AUC. In terms of precision and recall, we observe higher improvements with changing number of features. Using more features in prediction increases the ability of algorithm to predict ones.

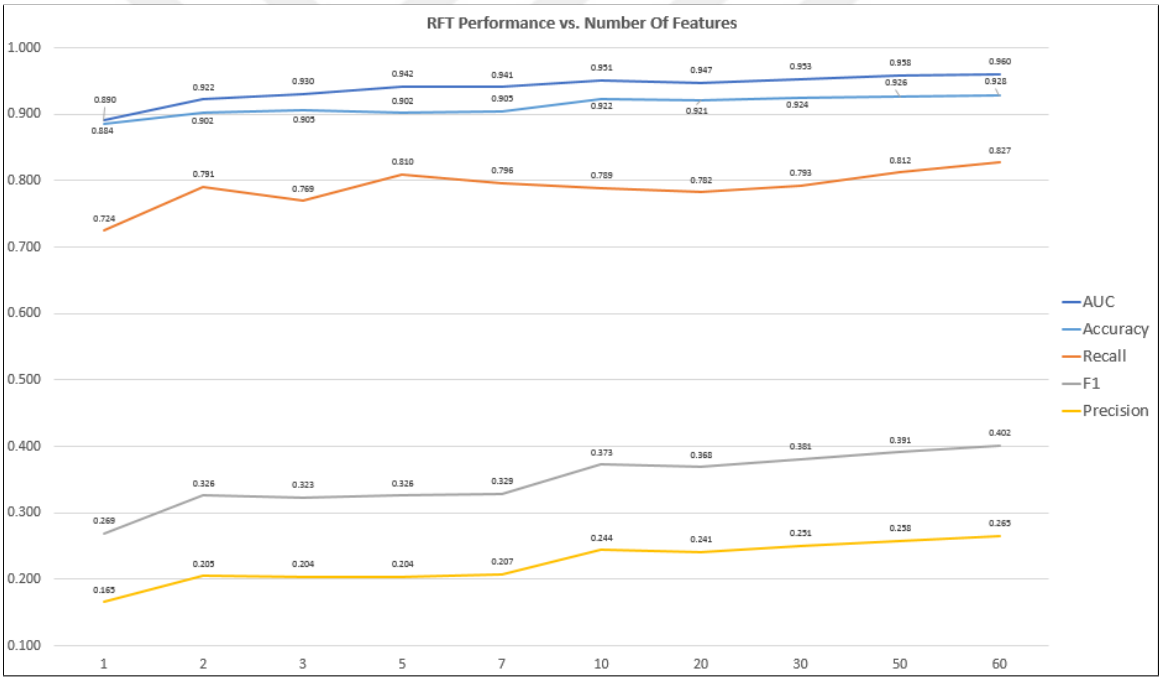

Figure 5.5: Performance vs. Number of Features

The list of features selected by the feature selector is shown in Table 5.3. Only with *radial_credit* feature, we can obtain 88.9% AUC score. The first five most effective features are related to the spatial dimension of the problem, which indicates the importance of spatial features in prediction of expenditures. For the first seven features, feature selector chooses input dependent features in all folds of cross validation. Only one constant feature, average interval between transactions of the customer is selected in the first ten features.

# Of Features	Selected Features
1	radial_credit
2	radial_credit, radial_atm
3	dist_to_home, radial_credit, radial_atm
5	dist_to_home, dist_to_work, radial_credit, radial_atm, r
7	dist_to_home, dist_to_work, dist_to_branch radial_credit, radial_atm, r, hourly_credit
10	dist_to_home, dist_to_work, dist_to_branch hourly_atm, radial_credit, hourly_credit radial_atm, averageinterval, daily_credit, r

Table 5.2: Selected Features in Feature Number Analysis

5.4 Sensitivity Analysis

Our aim is developing an approach that performs well for different location, radius, time interval and days. In order to test the flexibility of our approach and observe the changes in different cases, we implement a sensitivity analysis. The ranges for radius and time values are shown in Tables 4.1 and 4.2. We first analyze the prediction performance for all cases, which are shown in Figure 5.6 and Figure 5.7. The values in the paranthesis near the categories represent the ratio of ones in the test set.

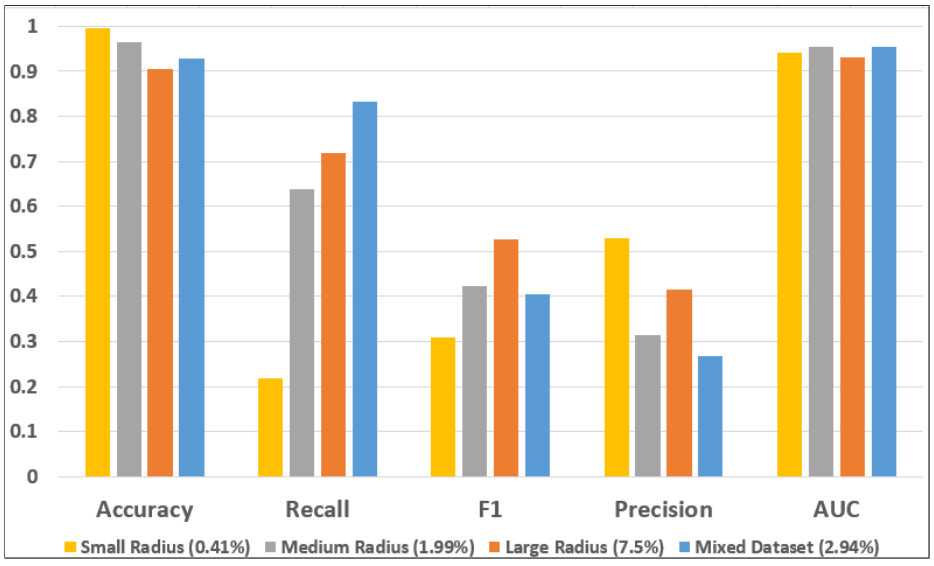

Figure 5.6: Radius Sensitivity Performances

The outcomes of radius sensitivity analysis are:

- We observe changes in prediction performances with changing prediction radius.
- The prediction performance decreases while input radius increases according to accuracy. With the increase in radius, recall and F1 values increase. For larger radiuses, ones can be predicted better.
- Datasets including smaller number of zeroes perform better in terms of accuracy, however have poor performance according to recall and f1 measures.

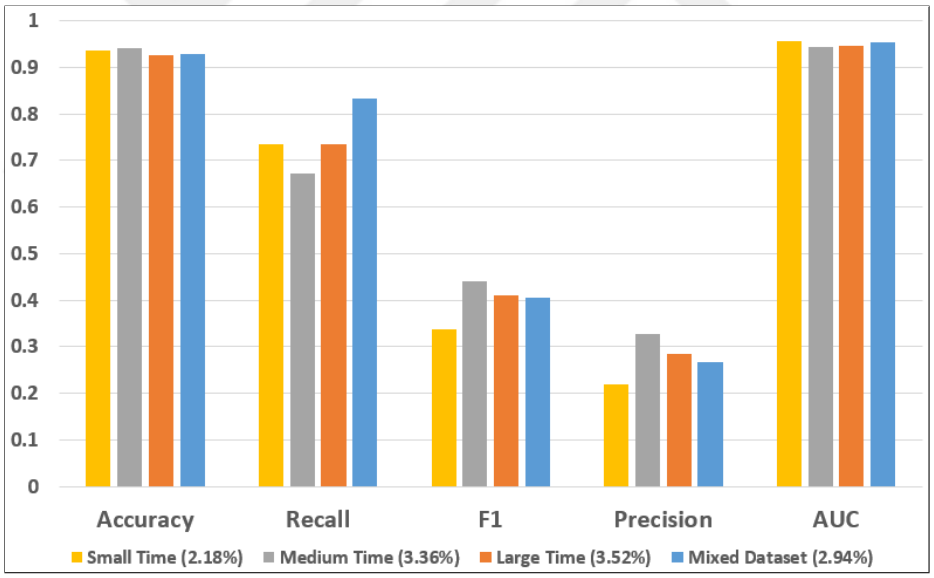

Figure 5.7: Time Interval Sensitivity Performances

The outcomes of time interval sensitivity analysis are:

- Changing prediction time interval affects the prediction performance less compared to radius sensitivity.
- We can not observe significant changes in the prediction performance according to accuracy and AUC measures. In terms of recall, precision and AUC, changing time range affects the prediction performance more.

- In terms of accuracy, F1 and precision best performance is obtained in the medium time range.

After analyzing the performances, we investigated the feature importances in all cases of sensitivity analysis. In Figure 5.8, the importance of input dependent features in radius sensitivity is presented. Main inferences from the results are:

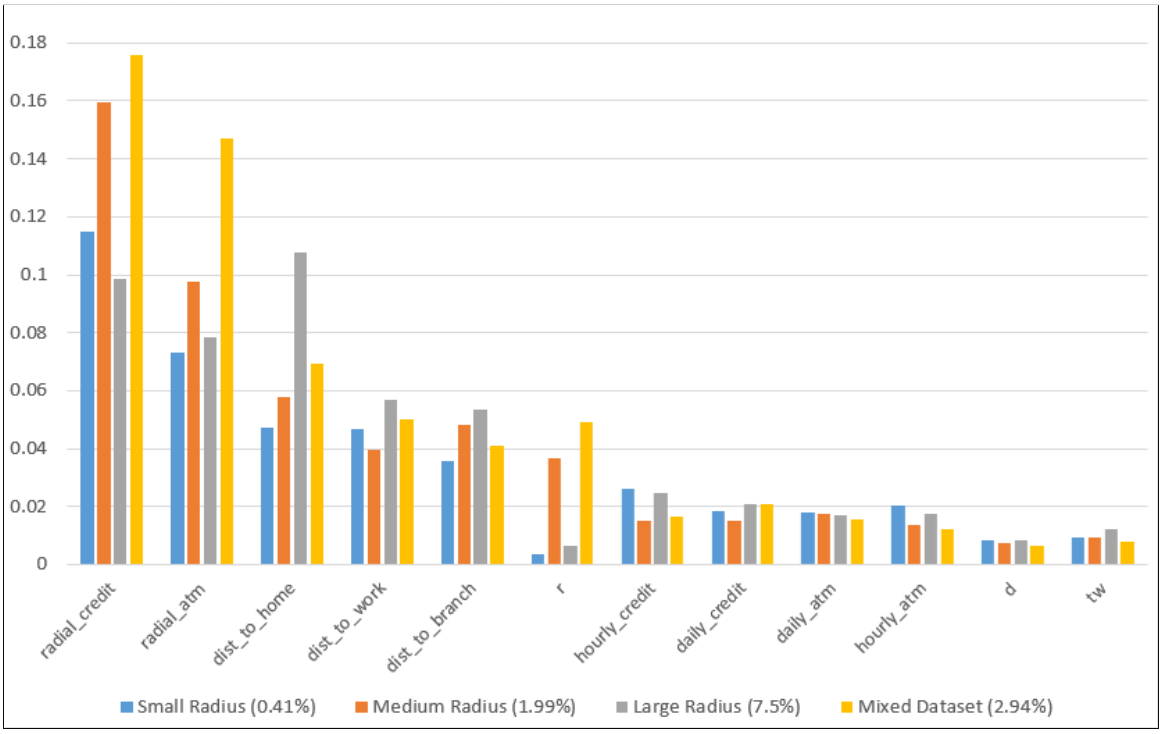

Figure 5.8: Input Dependent Feature Importances in Radius Sensitivity

- Radial proximity scores are the most important features in the expenditure prediction; followed by distance to home, work, branch features among the input dependent features. Location based features are the most effective features in all radius sensitivity cases.

- Time and day based input dependent features are less effective in prediction compared to location based ones. However, we cannot see a pattern in the importances with changing radius values.

- For the dataset with mixed radius, radial proximity scores and input radius (r) have higher importances than the importances of these features in other cases.

In Figure 5.9, feature importances for time sensitivity analysis are presented. Main outcomes of this analysis are:

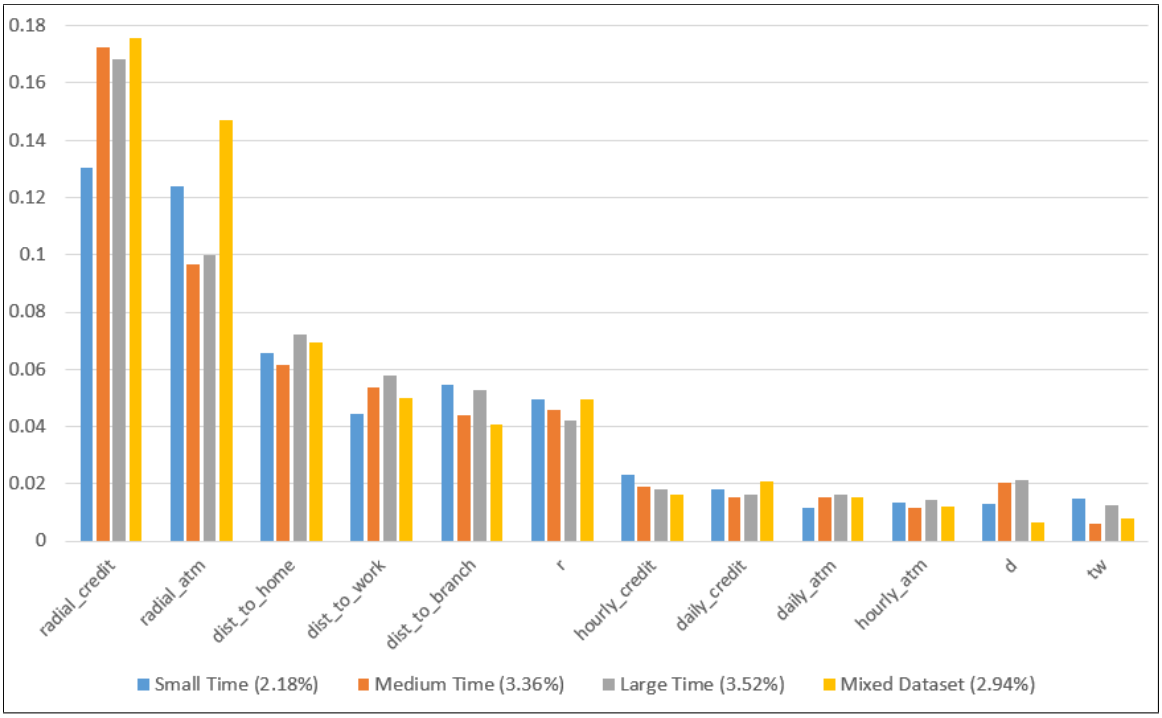

Figure 5.9: Input Dependent Feature Importances in Time Sensitivity

- Radial proximity scores are the most important features for prediction independent of the prediction time interval.

- We cannot see a pattern in the importances of hourly proximity features between cases. The time of activities does not include beneficial information for future expenditures compared to the radial proximity features.

In the calculation of proximity, all transactions have the same weights. In order to assess whether the recency of the transactions is important, we tried several α

values defined in Equation 3.5. By changing the α value in proximity calculation, we investigated the effect of the time coefficient in the prediction performance. The results in Table 5.4 shows that best performance can be obtained by not including the time effect in the prediction process. This results implies that, recent expenditures and old expenditures have the same effect in prediction.

Time Coefficient Selection					
Time Coefficient	**Accuracy**	**Recall**	**F1**	**Precision**	**AUC**
1	0.9279	**0.8333**	**0.4046**	**0.2671**	**0.9564**
1.25	**0.9289**	0.7925	0.3959	0.2638	0.949
1.5	0.927	0.8027	0.3926	0.2599	0.9496
0.75	0.9264	0.8061	0.3917	0.2587	0.953
0.5	0.9296	0.7857	0.3872	0.2569	0.9525

Table 5.3: Time Adjustment Analysis

In order to test the input dependency of our model, we made a sensitivity analysis on locations. For the 7 locations that we determined, we trained our model with the train data from the other 6 locations and take the test data from the chosen location. Main outcomes from this analysis are:

- Algorithm performs well for the locations including more expenditures (ones), in terms of all performance measures except accuracy.

- For the locations where there are few expenditures, algorithm cannot predict the expenditures well. In locations 4,5 and 7 none of the expenditures are predicted.

- Best performance is obtained when all locations included in the train and test datasets.

Location Out	**Ratio of Ones**	**Accuracy**	**Recall**	**F1**	**Precision**	**AUC**
1	0.081	0.843	0.599	0.381	0.279	0.828
2	0.071	0.800	0.805	0.363	0.234	0.881
3	0.058	0.855	0.701	0.359	0.241	0.883
4	0.008	0.992	0.000	0.000	0.000	0.637
5	0.004	0.996	0.000	0.000	0.000	0.512
6	0.002	0.996	0.500	0.321	0.237	0.865
7	0.0004	0.9996	0.000	0.000	0.000	0.356
All Included	0.029	0.928	0.833	0.405	0.268	0.955

Table 5.4: Results of Location Sensivity Analysis

Chapter 6

CONCLUSION

Predicting future expenditures of a bank customer in a given location, time interval and day is beneficial for effective offer decisions for banks. In this study, we propose an approach for making spatio-temporal expenditure forecasts for bank customers. In order to successfully predict the expenditures, we analyzed 63 features that consist of constant and input dependent features. We introduced the concept of input dependent features and generated proximity features to explain the relation between the transactions of customers and given input.

Using generated features, we tested six different data mining algorithms in a dataset containing various input locations, time intervals and days. We compared these algorithms with respect to five different performance measures. For the best performing algorithm, we analyzed the effects of the specific features in prediction performance and investigated proximity and behavioral features in detail. Finally, we made a sensitivity analysis for various radius and time range categories.

In this study, we succeeded to develop a spatio-temporal expenditure forecasting approach with an accuracy of 92.81% and with an F1 score of 40.5%. Another important contribution of this study is the definition of the proximity features which we found to be the most effective features in prediction performance. We also present insights behind the predictability of credit card expenditures.

This study can provide a guideline for spatio-temporal expenditure forecasting and can be extended in several directions such as including different location data sources and predicting the amount or quantity of expenditures.

BIBLIOGRAPHY

(2015). The consumer credit card market. Technical report, Consumer Financial Protection Bureau. Available: http://files.consumerfinance.gov/f/201512_cfpb_report-the-consumer-credit-card-market.pdf, Accessed on June 2017.

(2015). U.s. consumer banks and the potential of location-based offers. Technical report, Cognizant. Available: https://www.cognizant.com/whitepapers/US-Consumer-Banks-and-the-Potential-of-Location-Based-Offers-codex1279.pdf, Accessed on June 2017.

Aharony, N., Pan, W., Ip, C., Khayal, I., and Pentland, A. (2011). The social fmri: measuring, understanding, and designing social mechanisms in the real world. In *Proceedings of the 13th international conference on Ubiquitous computing*, pages 445–454. ACM.

Antunes, C. M. and Oliveira, A. L. (2001). Temporal data mining: An overview. In *KDD workshop on temporal data mining*, volume 1, page 13.

Ashbrook, D. and Starner, T. (2003). Using gps to learn significant locations and predict movement across multiple users. *Personal and Ubiquitous computing*, 7(5):275–286.

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., and Vanthienen, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the operational research society*, 54(6):627–635.

Bhattacharyya, S., Jha, S., Tharakunnel, K., and Westland, J. C. (2011). Data

mining for credit card fraud: A comparative study. *Decision Support Systems*, 50(3):602–613.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Chandrashekar, G. and Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16–28.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Cheng, T. and Wang, J. (2006). Applications of spatio-temporal data mining and knowledge for forest fire. *Remote Sensing: From Pixels to Processes*, pages 148–153.

Eagle, N., Macy, M., and Claxton, R. (2010). Network diversity and economic development. *Science*, 328(5981):1029–1031.

Fred, N. and Glover, F. (1981). Simple but powerful goal programming formulations for the statistical discriminant problem. *European Journal of Operational Research*, 7:44–60.

Gambs, S., Killijian, M.-O., and del Prado Cortez, M. N. (2012). Next place prediction using mobility markov chains. In *Proceedings of the First Workshop on Measurement, Privacy, and Mobility*, page 3. ACM.

Gomes, J. B., Phua, C., and Krishnaswamy, S. (2013). Where will you go? mobile data mining for next place prediction. In *International Conference on Data Warehousing and Knowledge Discovery*, pages 146–158. Springer.

Gonzalez, M. C., Hidalgo, C. A., and Barabasi, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, 453(7196):779–782.

Hand, D. J. and Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 160(3):523–541.

Hanley, J. A. and McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148(3):839–843.

Krumme, C., Llorente, A., Cebrian, M., Moro, E., et al. (2013). The predictability of consumer visitation patterns. *arXiv preprint arXiv:1305.1120*.

Maass, P., Koehler, T., Costa, R., Parlitz, U., Kalden, J., Wichard, J., and Merkwirth, C. (2003). Mathematical methods for forecasting bank transaction data. *Zentrum für Technomathematik*.

Min, W. and Wynter, L. (2011). Real-time road traffic prediction with spatio-temporal correlations. *Transportation Research Part C: Emerging Technologies*, 19(4):606–616.

Ngai, E., Hu, Y., Wong, Y., Chen, Y., and Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, 50(3):559–569.

Noulas, A., Scellato, S., Lathia, N., and Mascolo, C. (2012). Mining user mobility features for next place prediction in location-based services. In *Data mining (ICDM), 2012 IEEE 12th international conference on*, pages 1038–1043. IEEE.

Nourani, V., Mogaddam, A. A., and Nadiri, A. O. (2008). An ann-based model for spatiotemporal groundwater level forecasting. *Hydrological processes*, 22(26):5054–5066.

Orgler, Y. E. (1970). A credit scoring model for commercial loans. *Journal of money, Credit and Banking*, 2(4):435–445.

Roddick, J. F. and Spiliopoulou, M. (1999). A bibliography of temporal, spatial and spatio-temporal data mining research. *ACM SIGKDD Explorations Newsletter*, 1(1):34–38.

Sadilek, A., Kautz, H., and Bigham, J. P. (2012). Finding your friends and following them to where you are. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 723–732. ACM.

Shannon, C. E. (2001). A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55.

Shekhar, S. and Xiong, H. (2008). Spatial data mining. *Encyclopedia of GIS*, pages 1087–1087.

Singh, V. K., Bozkaya, B., and Pentland, A. (2015). Money walks: implicit mobility behavior and financial well-being. *PloS one*, 10(8):e0136628.

Singh, V. K., Freeman, L., Lepri, B., and Pentland, A. S. (2013). Predicting spending behavior using socio-mobile features. In *Social Computing (SocialCom), 2013 International Conference on*, pages 174–179. IEEE.

Sobolevsky, S., Sitko, I., Des Combes, R. T., 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. the case of residents and foreign visitors in spain. In *Big Data (BigData Congress), 2014 IEEE International Congress on*, pages 136–143. IEEE.

Song, C., Qu, Z., Blumm, N., and Barabási, A.-L. (2010). Limits of predictability in human mobility. *Science*, 327(5968):1018–1021.

Suykens, J. A. and Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300.

Verhein, F. and Chawla, S. (2006). Mining spatio-temporal association rules, sources, sinks, stationary regions and thoroughfares in object mobility databases. In *International Conference on Database Systems for Advanced Applications*, pages 187–201. Springer.

Vlahogianni, E. I., Karlaftis, M. G., and Golias, J. C. (2007). Spatio-temporal short-term urban traffic volume forecasting using genetically optimized modular networks. *Computer-Aided Civil and Infrastructure Engineering*, 22(5):317–325.

Wold, S., Esbensen, K., and Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52.

Yao, X. (2003). Research issues in spatio-temporal data mining. In *Workshop on Geospatial Visualization and Knowledge Discovery, University Consortium for Geographic Information Science, Virginia*, pages 1–6.

Yeh, I.-C. and Lien, C.-h. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2):2473–2480.

Zhang, G., Patuwo, B. E., and Hu, M. Y. (1998). Forecasting with artificial neural networks:: The state of the art. *International journal of forecasting*, 14(1):35–62.

Zhang, H. (2004). The optimality of naive bayes. *AA*, 1(2):3.

Chapter 7

APPENDIX

Table 7.1: Descriptive Statistics

Demographic Features

Age	**Education Status**	**Gender**
Mean: 38.5743	College: 4208	Female: 2638
Std: 8.9768	High School: 4311	Male: 7362
Min: 20	Master: 420	
1st Qu: 32	Middle School: 543	
Median: 37	PHD: 33	
3rd Qu: 44	Primary School: 422	
Max: 83	Unknown: 63	

Marital Status	**Type Of Work**
Divorced: 484	Other: 123
Married: 7469	Retired: 770
Single: 1759	Self-Employed:1224
Unknown: 288	Public Servant: 583
	Private Sector Employee: 7300

Financial Features

Income	**Bank Age**	**Mean Risk Score**
Mean: 4550.7653	Mean: 9.5589	Mean: 3.0890
Std: 7895.1318	Std: 4.7274	Std: 0.6502
Min: 550	Min: 1	Min: 0
1st Qu: 1622.75	1st Qu: 6	1st Qu: 2.8888
Median: 2572	Median: 9	Median: 3
3rd Qu: 5000	3rd Qu: 13	3rd Qu: 3.5555
Max: 200000	Max: 33	Max: 5

Continued on next page

Table 7.1 – continued from previous page

Variance of Risk Scores	**Last Known Risk Score**	**Bank Credit Card Limit**
Mean: 0.1828	Mean: 3.0853	Mean: 10980.7702
Std: 0.2579	Std: 0.7935	Std: 11239.466299
Min: 0	Min: 0	Min: 0
1st Qu: 0	1st Qu: 3	1st Qu: 3400
Median: 0.1111	Median: 3	Median: 7200
3rd Qu: 0.25	3rd Qu: 4	3rd Qu: 15000
Max: 3.5277	Max: 5	Max: 150000
Total Credit Card Limit	**Avg. Balance**	**Last Known Balance**
Mean: 24407.6605	Mean: 2361.046265	Mean: 2094.949249
Std: 32531.376341	Std: 3161.290697	Std: 3174.880416
Min: 0	Min: 0	Min: 0
1st Qu: 5800	1st Qu: 660.814722	1st Qu: 448.4
Median: 12800	Median: 1377.946667	Median: 1124.98
3rd Qu: 29700	3rd Qu: 2760.686667	3rd Qu: 2512.4475
Max: 498252	Max: 61315.202222	Max: 45051.51
Avg. Auto Payment Amount	**Avg. Expense in Month 1**	**Avg. Expense in Month 2**
Mean: 14.208489	Mean: 159.28034	Mean: 189.103562
Std: 44.317348	Std: 416.142091	Std: 644.854013
Min: 0	Min: 0	Min: 0
1st Qu: 0	1st Qu: 32.11	1st Qu: 30.5575
Median: 0	Median: 73.496111	Median: 70.500192
3rd Qu: 0	3rd Qu: 152.355808	3rd Qu: 150
Max: 1815	Max: 16200	Max: 17221.5
Avg. Expense in Month 3	**Avg. Expense in Month 4**	**Avg. Expense in Month 5**
Mean: 185.004909	Mean: 179.284181	Mean: 154.693506
Std: 519.185326	Std: 590.99417	Std: 372.775395
Min: 0	Min: 0	Min: 0
1st Qu: 34.345238	1st Qu: 36.469583	1st Qu: 35.213472
Median: 77.1	Median: 75.63	Median: 75
3rd Qu: 167.034286	3rd Qu: 150.880833	3rd Qu: 146.290833
Max: 16965	Max: 16700	Max: 12560.53

Continued on next page

Table 7.1 – continued from previous page

Avg. Expense in Month 6	**Avg. Expense in Month 7**	**Avg. Expense in Month 8**
Mean: 172.673456	Mean: 165.6179	Mean: 153.249862
Std: 488.206699	Std: 373.608044	Std: 345.388716
Min: 0	Min: 0	Min: 0
1st Qu: 36.690741	1st Qu: 37.582143	1st Qu: 35.582455
Median: 75.248514	Median: 77.765	Median: 71.863667
3rd Qu: 155.066667	3rd Qu: 165.765692	3rd Qu: 148.98
Max: 15657.5	Max: 10000	Max: 10000

Avg. Expense in Month 9	**Variance of Monthly Expenses**	**Credit Card Trans.**
Mean: 165.443985	Mean: 184758.8	Mean: 27.7032
Std: 462.329656	Std: 1778218	Std: 39.325719
Min: 0	Min: 0	Min: 0
1st Qu: 37.363697	1st Qu: 874.8357	1st Qu: 9
Median: 74.9705	Median: 4188.88	Median: 20
3rd Qu: 154.860833	3rd Qu: 22281.7	3rd Qu: 36
Max: 15000	Max: 95238770	Max: 1475

Num. Of ATM Transactions

Mean: 16.6527

Std: 14.693293

Min: 0

1st Qu: 6

Median: 14

3rd Qu: 23

Max: 197

Product Usage & Bank Relations

Num. Of Active Products	**Num. Of All Products**	**Total CC Calls**
Mean: 7.0213	Mean: 8.1627	Mean: 81.8228
Std: 2.125205	Std: 2.177224	Std: 132.256606
Min: 0	Min: 0	Min: 0
1st Qu: 6	1st Qu: 7	1st Qu: 18
Median: 7	Median: 8	Median: 48
3rd Qu: 8	3rd Qu: 9	3rd Qu: 101
Max: 20	Max: 21	Max: 5600

Continued on next page

Table 7.1 – continued from previous page

Last Month CC Calls	**Num. Of Credit Cards**	**Other Bank Credit Cards**
Mean: 10.0546	Mean: 2.4356	Mean: 1.4533
Std: 26.45362	Std: 1.56547	Std: 1.559122
Min: 0	Min: 0	Min: 0
1st Qu: 0	1st Qu: 1	1st Qu: 0
Median: 0	Median: 2	Median: 1
3rd Qu: 12	3rd Qu: 3	3rd Qu: 2
Max: 651	Max: 14	Max: 13

Mean Transaction Interval	**Var. Of Intervals**	**Last Payment Date**
Mean: 5.487344	Mean: 803522.3	Mean: 13.620919
Std: 5.958314	Std: 1074579	Std: 7.050448
Min: 0	Min: 0	Min: 0
1st Qu: 2.359649	1st Qu: 298048.1	1st Qu: 7.7
Median: 3.791045	Median: 526140.8	Median: 12.363636
3rd Qu: 6.166667	3rd Qu: 867811	3rd Qu: 18.083333
Max: 99	Max: 19731760	Max: 28.5

Behavioral Features

Hourly Diversity	**Hourly Loyalty**	**Hourly Regularity**
Mean: 0.874652	Mean: 0.423028	Mean: 0.935787
Std: 0.046452	Std: 0.090999	Std: 0.072449
Min: 0.517541	Min: 0.206349	Min: 0
1st Qu: 0.854938	1st Qu: 0.357513	1st Qu: 0.921618
Median: 0.883597	Median: 0.407407	Median: 0.951782
3rd Qu: 0.905308	3rd Qu: 0.472222	3rd Qu: 0.970955
Max: 0.958391	Max: 0.918519	Max: 1

Daily Diversity	**Daily Loyalty**	**Daily Regularity**
Mean: 0.905197	Mean: 0.556722	Mean: 0.940031
Std: 0.028237	Std: 0.054728	Std: 0.072361
Min: 0.432792	Min: 0.441176	Min: 0
1st Qu: 0.896624	1st Qu: 0.517544	1st Qu: 0.923343
Median: 0.912935	Median: 0.548954	Median: 0.955822
3rd Qu: 0.923039	3rd Qu: 0.585106	3rd Qu: 0.977843

Continued on next page

Table 7.1 – continued from previous page

Max: 0.935257	Max: 0.896296	Max: 1
Spatial Radial Diversity	**Spatial Radial Loyalty**	**Spatial Radial Regularity**
Mean: 0.442354	Mean: 0.969306	Mean: 0.890773
Std: 0.174756	Std: 0.037698	Std: 0.095812
Min: 0	Min: 0.705128	Min: 0
1st Qu: 0.306825	1st Qu: 0.956522	1st Qu: 0.846886
Median: 0.458319	Median: 0.981818	Median: 0.916197
3rd Qu: 0.577917	3rd Qu: 0.998287	3rd Qu: 0.958306
Max: 0.864828	Max: 1	Max: 1
Spatial Cluster Diversity	**Spatial Cluster Loyalty**	**Spatial Cluster Regularity**
Mean: 0.624739	Mean: 0.75438	Mean: 0.918573
Std: 0.148378	Std: 0.144477	Std: 0.069558
Min: 0	Min: 0.285714	Min: 0
1st Qu: 0.534788	1st Qu: 0.651163	1st Qu: 0.900942
Median: 0.649	Median: 0.768043	Median: 0.932184
3rd Qu: 0.735663	3rd Qu: 0.871429	3rd Qu: 0.954623
Max: 0.921297	Max: 1	Max: 1
Categorical Diversity	**Categorical Loyalty**	**Categorical Regularity**
Mean: 0.630586	Mean: 0.75728	Mean: 0.843002
Std: 0.189135	Std: 0.148802	Std: 0.227063
Min: 0	Min: 0.285714	Min: 0
1st Qu: 0.528655	1st Qu: 0.646295	1st Qu: 0.860101
Median: 0.672589	Median: 0.761905	Median: 0.919073
3rd Qu: 0.772446	3rd Qu: 0.87931	3rd Qu: 0.948478
Max: 0.95351	Max: 1	Max: 1
Input Related Features		
Distance To Home	**Distance To Work**	**Distance to Branch**
Mean: 52.54355	Mean: 53.288832	Mean: 64.191192
Std: 111.418329	Std: 116.154242	Std: 144.242048
Min: 0.513871	Min: 0.155085	Min: 0.246391
1st Qu: 6.668118	1st Qu: 6.230901	1st Qu: 6.070833
Median: 13.274889	Median: 12.373794	Median: 11.958574

Continued on next page

Table 7.1 – continued from previous page

3rd Qu: 22.398789	3rd Qu: 22.52169	3rd Qu: 21.95362
Max: 1243.071999	Max: 1398.802018	Max: 1449.411924
Input Radius	**Input Time Range**	**Input Day**
Mean: 11.3665	Mean: 6.4585	Mean: 4.1288
Std: 7.683052	Std: 3.699303	Std: 2.113683
Min: 1	Min: 2	Min: 1
1st Qu: 5	1st Qu: 4	1st Qu: 2
Median: 15	Median: 5	Median: 4
3rd Qu: 15	3rd Qu: 10	3rd Qu: 6
Max: 30	Max: 14	Max: 7
Proximity Features		
Radial Proximity Credit	**Hourly Proximity Credit**	**Daily Proximity Credit**
Mean: 2.496051	Mean: 3.99649	Mean: 2.341756
Std: 1.531212	Std: 1.968364	Std: 0.632884
Min: 1	Min: 0	Min: 0
1st Qu: 1.205128	1st Qu: 2.612831	1st Qu: 1.845997
Median: 1.94392	Median: 3.684211	Median: 2.217391
3rd Qu: 3.396928	3rd Qu: 5.036851	3rd Qu: 2.818182
Max: 6	Max: 11.584416	Max: 6
Radial Proximity ATM	**Hourly Proximity ATM**	**Daily Proximity ATM**
Mean: 2.505448	Mean: 4.011223	Mean: 2.289148
Std: 1.58022	Std: 2.063439	Std: 0.711713
Min: 0	Min: 0	Min: 0
1st Qu: 1.122807	1st Qu: 2.638889	1st Qu: 1.75
Median: 1.984848	Median: 3.754852	Median: 2.145455
3rd Qu: 3.529802	3rd Qu: 5.15392	3rd Qu: 2.842209
Max: 6	Max: 14	Max: 6
Prediction Class		
0: 15000		
1: 5000		

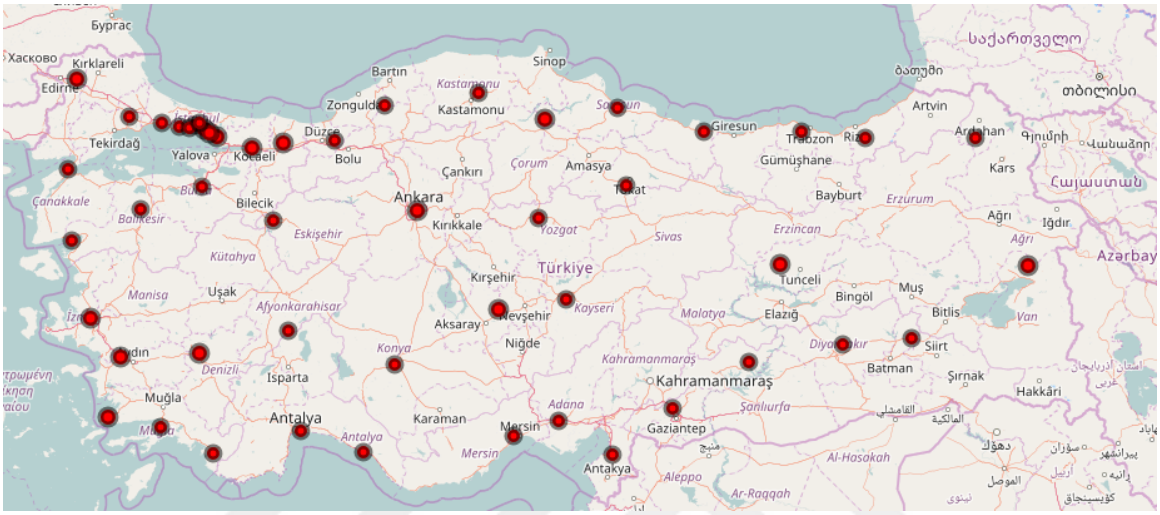

Figure 7.1: Cluster Centers

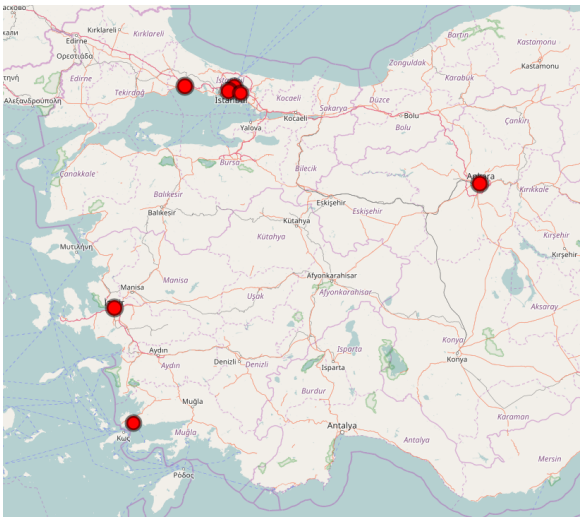

Figure 7.2: Input Locations