



REPUBLIC OF TÜRKİYE

ALTINBAŞ UNIVERSITY

Institute of Graduate Studies

Data Analytics

**RETAIL SALES ANALYSIS AND FORECASTING
IN TEXTILE INDUSTRY**

Hüseyin YILDIRIM

Master's Thesis

Supervisor

Asst. Prof Dr. Abdullahi Abdu IBRAHIM

İstanbul, 2023

RETAIL SALES ANALYSIS AND FORECASTING IN TEXTILE INDUSTRY

Hüseyin YILDIRIM

Data Analytics

Master's Thesis

ALTINBAŞ UNIVERSITY

2023

The thesis titled RETAIL SALES ANALYSIS AND FORECASTING IN TEXTILE INDUSTRY prepared by HÜSEYİN YILDIRIM and submitted on 18.12.2023 has been **accepted unanimously** for the degree of Master of Science in Data Analytics Department.

Asst. Prof. Dr. Abdullahi Abdu
IBRAHIM

Thesis Defense Committee Members:

Assoc. Prof. Dr. Sefer KURNAZ	Department of Computer Science, Altınbaş University	_____
Asst. Prof. Dr. Abdullahi Abdu IBRAHIM	Department of Computer Science, Altınbaş University	_____
Asst. Prof. Dr. Tareq Abed MOHAMMED	Department of Computer Science, Kirkuk University	_____

I hereby declare that this thesis meets all format and submission requirements of a Master's thesis.

I hereby declare that all information/data presented in this graduation project has been obtained in full accordance with academic rules and ethical conduct. I also declare all unoriginal materials and conclusions have been cited in the text and all references mentioned in the Reference List have been cited in the text, and vice versa as required by the abovementioned rules and conduct.

Hüseyin YILDIRIM

Signature

ABSTRACT

RETAIL SALES ANALYSIS AND FORECASTING IN TEXTILE INDUSTRY

YILDIRIM, Hüseyin

M.Sc., Data Analytics Department, Altınbaş University,

Supervisor: Asst. Prof. Dr. Abdullahi Abdu IBRAHIM

Date: 12/2023

Pages: 33

This thesis addresses sales forecasting in the retail domain using machine learning methods. Nowadays, the retail sector emerges as a rapidly changing and exponentially growing field. Therefore, obtaining accurate and reliable predictions in sales forecasting holds critical importance for businesses to sustain their competitive advantage. Machine learning is recognized as an effective tool with data analysis and pattern recognition capabilities to address complex problems like sales forecasting.

This study aims to tackle the sales forecasting problem in the retail sector and investigate how machine learning methods can be utilized to solve this issue. Firstly, an exploration of the existing sales forecasting methods in the literature and machine learning algorithms is conducted. Subsequently, the effectiveness and performance of various machine learning algorithms (such as LGBM, LSTM, XGBoost) in sales forecasting are compared.

This thesis is supported by experimental studies conducted on real-world datasets. The datasets encompass sales data from the textile retail sector and comprise data obtained over a specific time frame. Analyses conducted on these datasets reveal how machine learning algorithms can offer an advantage in sales forecasting compared to traditional methods.

The results demonstrate that machine learning algorithms constitute an effective and accurate tool for sales forecasting in the retail sector. This thesis provides valuable insights

into the LightGBM method specifically chosen and can aid businesses in enhancing strategic decisions, such as demand management, inventory optimization, and stock planning.

Keywords: Retail, Machine Learning, Python, Sql, Forecast.



TABLE OF CONTENTS

	<u>Pages</u>
ABSTRACT	v
LIST OF TABLES.....	ix
ABBREVIATIONS.....	x
1.INTRODUCTION AND PURPOSE.....	1
2.GENERAL INFORMATION.....	3
3.METHODOLOGY	5
3.1 RETAIL	5
3.2 TEXTILE	5
3.3 PRODUCT RANGES.....	6
3.4 PREDICTION.....	7
3.5 REGRESSION.....	8
3.6 DATA ANALYSIS.....	9
3.7 PYTHON	9
3.8 STRUCTURED QUERY LANGUAGE	10
3.9 ARTIFICIAL NEURAL NETWORKS.....	11
3.10 BOOSTING METHODS	12
3.11 MACHINE LEARNING	13
3.12 DATASET	14
3.13 NORMALIZATION	14
3.14 ENCODING.....	15
3.15 LONG SHORT-TERM MEMORY	17
3.16 LIGHTGBM	18
3.17 ROOT MEAN SQUARE ERROR	19
4.IMPLEMENTATION	20

4.1 DATA COLLECTION	20
4.2 STORE INFORMATION.....	20
4.3 PRODUCT INFORMATION.....	20
4.4 SALES INFORMATION	20
4.5 INVENTORY INFORMATION	20
4.6 CAPACITY INFORMATION	20
4.7 DATA PREPROCESSING.....	21
4.7.1 Data Transformation	21
4.7.2 Enhancing Data Significance	21
4.7.3 Handling Missing Data	21
4.8 APPLICATION OF MACHINE LEARNING WITH LIGHTGBM	21
5.RESULTS	23
6.DISCUSSION AND CONCLUSIONS.....	24
6.1 DATA QUALITY ENHANCEMENT	24
6.2 FURTHER FEATURE ENGINEERING	24
6.3 MODEL OPTIMIZATION.....	24
6.4 HYBRID MODELS.....	24
6.5 DATA COLLECTION AND MONITORING.....	24
6.6 LITERATURE REVIEW	25
6.7 SEPARATE MODELS FOR OUTLYING BEHAVIOR.....	25
6.8 UTILIZE BOOSTING METHODS IN NON-PERIODIC DATA.....	25
REFERENCES	24

LIST OF TABLES

	<u>Pages</u>
Table 4.1: Important Particles.	22
Table 5.1: Example Data.	23



ABBREVIATIONS

AI	:	Artificial Intelligence
LGBM	:	Light Gradient Boosting Machine
ANN	:	Artificial Neural Network
SQL	:	Structured Query Language
RNN	:	Recurrent Neural Network



1. INTRODUCTION AND PURPOSE

Retail involves the direct provision of goods or services to the ultimate consumer through various channels such as physical stores, store chains, and online platforms. Retail businesses engage in purchasing products wholesale and then selling them to customers for profit. As the final stage reaching the end consumer, retail enterprises can serve a broad consumer base.

One of the significant challenges faced by companies in the retail sector today is their inability to accurately forecast demand, resulting in outcomes like stockouts or excessive stockpiling, which significantly impact profit margins.

Demand projection is a process that utilizes past data to determine future customer needs. Due to its future-oriented nature, this projection carries substantial uncertainty, often making it difficult to precisely predict future demand. Nevertheless, estimations with a small margin of error play a crucial role in decision-making processes.

The rapid changes in fashion, seasons, and desired products underscore the importance of demand forecasting in the contemporary landscape. Consequently, demand projection methods have evolved from traditional statistical techniques to more complex machine learning models, owing to the swiftly changing nature of demand, fashion, and seasonality.

This study aims to develop a system for predicting the sales volume of products in retail stores for future periods, considering historical sales data from a retail company specializing in textile products, which holds a substantial sales volume in the global textile industry.

The second part of the project delves into an examination of existing literature on sales forecasting using machine learning methods. The third section provides theoretical background information on the deep learning algorithm and the decision tree algorithm, specifically LightGBM, which will be compared in the study.

Subsequently, the performance evaluation criteria for the deep learning and decision tree algorithms are established and elaborated upon. The selected criteria include the Mean Absolute Error (MAE) and the square root of the Root Mean Square Error (RMSE), both used to measure prediction performance.

The fourth section of the project outlines the detailed work conducted within the scope of the project. This section first introduces the problem, methodology, and the software program used. It then proceeds to introduce the dataset and discusses the data preprocessing steps.

The project consists of three main stages. In the first stage, data preprocessing is performed, followed by the selection of the optimal method based on regression using artificial neural networks and the LightGBM method. The method with the lowest error is chosen, and subsequently, sales predictions are made.

The results of the two methods are then compared. The final section of the study evaluates the obtained results from the practical implementation.

2. GENERAL INFORMATION

In today's dynamic business landscape, the retail sector has undergone significant transformation, becoming increasingly intricate, and has gained substantial importance, particularly within the textile industry. Businesses in this sector are compelled to make informed decisions in areas such as understanding consumer behaviors, grasping market trends, analyzing retail sales, and generating forecasts for the future. In this context, this literature review focuses on the analysis of retail sales and the development of forecasting methodologies within the textile industry. The study aims to explore contemporary approaches within the literature to enhance the analytical and predictive capabilities of the textile retail sector, contributing to more effective decision-making processes.

In recent years, the dynamic and complex nature encountered by businesses in the retail sector has brought to light the inadequacy of traditional methods. It is evident that traditional approaches have achieved limited success in the face of rapidly changing consumer behaviors, market trends, and the new challenges posed by digital transformation. In this context, machine learning methods such as LightGBM stand out for their ability to analyze the vast and intricate datasets offered by the retail sector more effectively. LightGBM's robust feature extraction capabilities and high learning speed hold significant potential for achieving more accurate and precise outcomes, particularly in critical areas like sales forecasting. Consequently, in the realm of the retail industry, traditional methods are being replaced by machine learning approaches that offer more customized and data-focused results, with innovative approaches like LightGBM taking on a prominent role as a promising direction for the future of retail analytics. Subsequently, the conducted literature reviews have been elucidated.

Through conducted experiments, it has been discovered that in the ready-to-wear retail industry, the artificial neural network method known as nonlinear autoregressive exogenous (NARX) modeling stands out as the most effective approach. [1].

Upon applying various machine learning methods, it has been concluded that selecting a suitable machine learning model for the dataset and, as a recommendation, utilizing ensemble methods have been achieved. [2].

In the context of promoting within the retail sector using machine learning methods, it has been deduced that increasing the number of features leads to a corresponding enhancement in accuracy. [3].

The demand forecasting module combined the practice of using similar products to predict the demand of a new product with a new store or customer cluster data aggregation scheme.[4].

In the fast-paced apparel industry, predicting sales, often referred to as demand or sales forecasting, has been a focal point of research in the last two decades. Numerous methods and measures have been explored to enhance the accuracy of demand predictions, recognizing the inherent challenge of achieving a completely error-free forecast.[5].

Taking into account material lead time and warehouse capacity, the company's production planners define specific safety stock levels for different items to prevent unexpected delays. While demand forecasting was not crucial in the past, the company's recent rapid growth and a substantial increase in product demand have led suppliers to inquire about anticipated material purchases in the near future. [6].

Forecasts are often inaccurate, leading to skepticism in some companies, with over 30% considering them unimportant (Silverman, 2017). Despite this, choosing the right forecasting method and implementation can positively impact a company's profitability and customer satisfaction. Barriers to effective forecasting include a lack of data, tools, and knowledge. Understanding the pros and cons of each method is crucial for successful implementation.[7].

3. METHODOLOGY

3.1 RETAIL

Retail, a crucial element in the economic framework, involves the processes of directly selling goods or services to end consumers. It serves as a vital link connecting manufacturers or suppliers to the ultimate users. Retail operates through various channels like physical stores and online platforms, managing aspects such as inventory, marketing, sales, customer engagement, and logistics. The industry's dynamism arises from its intricate interactions with consumer behavior, market trends, and economic changes, necessitating adaptive strategies for competitiveness. Successfully navigating this complex landscape requires a deep understanding of consumer preferences, market dynamics, and strategic decision-making to enhance business performance. The internet has given rise to a prevalent trend known as online retailing, where goods and services are sold through digital platforms. The increasing ubiquity of the internet has fueled a surge in consumer preference for e-retailing.[8].

Retail entails the sale of consumer goods or services to customers, utilizing various distribution channels with the aim of generating a profit. The process involves reaching customers through diverse means to offer products and services in exchange for financial gain.[9].

3.2 TEXTILE

Playing a pivotal role in the industrial landscape, the textile industry involves complex processes of designing, manufacturing, and distributing various fibers, fabrics, and finished textile products. Rooted in both tradition and innovation, this sector fulfills diverse human needs, ranging from clothing and home furnishings to industrial applications. The journey of textile production begins with raw materials, such as natural or synthetic fibers, and progresses through stages like spinning, weaving, knitting, dyeing, printing, and finishing. These intricate processes not only showcase historical craftsmanship but also integrate modern technological advancements to enhance textile quality, durability, and functional attributes. Beyond economic contributions, the global significance of the textile industry

extends to environmental and sustainability considerations, reflecting an increased awareness of responsible production practices and circular economies.

Textile stands as the fifth architectural material, alongside timber, stone, metal, and glass. Numerous articles in recent professional journals and magazines highlight projects that leverage the benefits of incorporating textiles alongside conventional construction materials. In certain instances, textiles are emerging as viable replacements for these traditional materials. The internet is a vast repository of information encompassing manufacturers, designers, engineers, constructors, and even new educational fields dedicated to textiles.[10].

3.3 PRODUCT RANGES

Apparel sizing ranges refer to the set of predefined sizes that a particular garment is produced and marketed in. These sizing ranges encompass a range of body measurements and dimensions, typically including attributes such as chest circumference, waist circumference, hip circumference, sleeve length, and inseam length. The purpose of apparel sizing ranges is to accommodate the diversity of human body shapes and sizes, providing consumers with options that fit their individual proportions.

Sizing ranges are a critical aspect of the apparel industry, influencing various stages of the production and distribution process. Designers and manufacturers utilize sizing ranges as a framework to create patterns and prototypes for different sizes. These ranges help ensure that garments maintain consistent proportions and aesthetics across various sizes, while also considering factors such as fit, comfort, and style.

Retailers and marketers play a key role in selecting and communicating sizing ranges to consumers. By offering a range of sizes that cater to a wide spectrum of body types, retailers aim to enhance customer satisfaction and inclusivity. Sizing ranges are often displayed on clothing labels, online product descriptions, and marketing materials, allowing consumers to make informed choices based on their measurements.

The determination of sizing ranges involves a combination of anthropometric data, statistical analysis, market research, and industry standards. Anthropometric measurements of target demographics are collected to understand the distribution of body sizes within a population.

Statistical methods are applied to analyze this data and define the breakpoints for different size categories.

In conclusion, apparel sizing ranges play a pivotal role in ensuring that clothing items are available in a variety of sizes to cater to the diverse range of human body shapes and proportions. By considering various measurements and dimensions, sizing ranges contribute to the creation of well-fitting and aesthetically pleasing garments, promoting consumer satisfaction and accessibility.

3.4 PREDICTION

Within the context of research and analysis, refers to the process of estimating future outcomes, events, or values based on existing data, patterns, and trends. In the realm of scientific inquiry, prediction serves as a valuable tool for generating insights and making informed decisions. It entails the application of various statistical, mathematical, or computational methods to extrapolate and anticipate forthcoming developments. Predictive modeling, a prominent approach within prediction, involves the creation of mathematical representations that capture relationships between variables, allowing researchers to forecast outcomes under different scenarios.

Prediction operates on the premise that historical data, when subjected to appropriate methodologies, can yield valuable foresight into potential future scenarios. By identifying patterns, correlations, and causal relationships within data, researchers can construct models that elucidate possible trajectories and probabilities. These models aid in identifying trends, guiding strategic planning, and minimizing uncertainty in decision-making processes. The robustness and accuracy of predictions rely on the quality and relevance of the underlying data, the appropriateness of the chosen methodologies, and the consideration of relevant contextual factors.

In the pursuit of academic research, prediction serves as a powerful tool to support hypothesis testing, policy formulation, and risk assessment. It enables researchers to explore the implications of their findings, make projections, and contribute to the advancement of knowledge within their respective fields.

3.5 REGRESSION

In the field of statistics and data analysis, regression is a powerful analytical technique employed to model the relationship between one or more independent variables and a dependent variable. It is widely utilized for predictive modeling, aiming to make informed estimations or forecasts based on observed data points.

Logistic regression is frequently employed to establish a connection between the probability of a categorical outcome, such as detection or non-detection, and a set of explanatory variables.[11].

In situations where identifying the subset of explanatory variables ($x_1, \dots, x_n$) with the most significant collective impact on the response variable's variation is unclear, a variable selection scheme becomes essential. Stepwise regression is a commonly employed method for this purpose, consisting of three fundamental forms: forward selection, backward elimination, and stepwise elimination. This allows for a systematic approach to determine the most influential variables affecting the response variable.[12].

Regression analysis entails constructing a model that optimally fits data, accounting for inherent variability. This model defines the functional connection between independent variables and the dependent variable, often visualized as a straight line (in simple linear regression) or a more intricate curve (in multiple linear or polynomial regression).

In the realm of academic research, the application of regression methods involves several pivotal steps. Initially, the researcher must formulate the research question and pinpoint relevant variables, categorizing them as independent (predictor) and dependent (response). Following this, data collection ensues, ensuring the availability of precise and representative data points.

The regression model is then created utilizing a chosen algorithm, such as ordinary least squares (OLS) for linear regression. The model undergoes refinement through parameter estimation, where the algorithm minimizes the disparity between observed and predicted values. This optimization process aims to identify regression coefficients that effectively capture the relationship between variables.

3.6 DATA ANALYSIS

Data analysis, within the realm of research and scholarly inquiry, refers to the systematic process of examining, interpreting, and deriving meaningful insights from collected data sets. It plays a pivotal role in transforming raw data into valuable knowledge, contributing to informed decision-making and the advancement of scientific understanding. Data analysis involves a structured sequence of activities, encompassing data cleansing, transformation, exploration, modeling, and interpretation.

The fundamental objective of data analysis is to uncover patterns, relationships, trends, and anomalies within the data, thereby facilitating the extraction of meaningful information. Researchers employ a variety of statistical, mathematical, and computational techniques to process and analyze data, allowing them to draw evidence-based conclusions and make informed predictions. By harnessing the power of data analysis, researchers can identify correlations, establish causality, test hypotheses, and generate models that encapsulate the underlying mechanisms governing observed phenomena.

The effectiveness of data analysis hinges on the rigor of the chosen methodologies, the quality of the data collected, and the appropriate selection of tools and techniques. Through the application of data analysis, researchers gain the ability to illuminate hidden insights, validate or challenge existing theories, and contribute to the advancement of knowledge in their respective fields.

3.7 PYTHON

Python, the programming language, stands as a versatile and widely embraced tool in the realms of computational science, data analysis, and software development. Recognized for its simplicity, readability, and robust library support, Python has become a primary choice for professionals in various fields, including researchers, engineers, and practitioners.

Distinguished by its clean syntax and object-oriented design, Python facilitates the creation of concise and expressive code, promoting efficient algorithmic development. As an open-source platform, it encourages collaboration, leading to the development of numerous libraries, frameworks, and modules that enhance functionality and broaden its application range.

Python finds application across various domains, serving purposes such as data manipulation, statistical analysis, machine learning, artificial intelligence, and web development. Its robust ecosystem enables seamless integration with other programming languages and platforms, further expanding its utility.

Python stands out as a versatile language suitable for both desktop and web-based programming. It excels in user-friendliness, particularly for beginners in basic programming. This is attributed to the simplicity, tidiness, and conciseness of Python's language commands. Recognized as one of the most potent languages, Python is widely employed and valued in the scientific community for its versatility and practical applications.[13].

Within scientific research and scholarly pursuits, Python's adaptability and accessibility empower researchers to construct intricate computational models, conduct sophisticated data analyses, and create innovative software solutions. The thriving Python community and comprehensive documentation ensure that individuals, whether new or experienced in programming, can harness its capabilities to advance their research objectives.

3.8 STRUCTURED QUERY LANGUAGE

Structured Query Language (SQL) serves as a specialized programming language integral to the domain of relational database management systems (RDBMS). Tailored to interact with and manipulate structured data within databases, SQL allows users to define, retrieve, modify, and manage data using a declarative syntax and a command set designed for diverse database operations.

Playing a crucial role in the management and organization of information within relational databases, SQL operates on structured tables storing data in rows and columns. It offers a standardized and user-friendly approach to querying databases, empowering users to formulate intricate queries for extracting specific data subsets, performing aggregations, and conducting calculations. Through SQL, users can create, modify, and delete tables, establish relationships between tables, and enforce data integrity constraints, ensuring consistency and reliability.

Additionally, SQL supports various tasks such as data insertion, updating, and deletion, along with the creation and management of indexes and views. With a rich set of data

manipulation and definition capabilities, SQL serves as a fundamental tool for database administrators, developers, analysts, and researchers.

In the contemporary landscape of data-driven applications and analytics, SQL remains a cornerstone for interacting with databases, facilitating data exploration, decision-making, and extracting insightful information. Its widespread adoption across industries underscores its significance as a standard language for managing and harnessing the potential of structured data.

3.9 ARTIFICIAL NEURAL NETWORKS

Artificial Neural Networks (ANNs) belong to a category of computational models inspired by the intricate structure and functions of biological neural networks found in the human brain. ANNs find extensive applications in machine learning, artificial intelligence, and data analysis due to their proficiency in effectively capturing and learning complex patterns from data.

Comprising interconnected nodes, often termed neurons, organized into layers, an ANN consists of three primary types of layers: the input layer, one or more hidden layers, and the output layer. Each neuron is associated with learnable parameters, including weights and biases, defining the interconnections between neurons.

The functioning of an ANN involves a forward propagation process, where input data is initially presented to the input layer and successively transformed through hidden layers using weighted connections and activation functions. These activation functions introduce non-linearities, enabling the network to grasp intricate data relationships, ultimately generating the final transformed output from the output layer.

Emerging from the domain of machine learning, an artificial neural network represents a black box model capable of processing data, grasping its dynamics, and ultimately delivering insightful outputs when appropriately designed.[14].

Artificial neural networks (ANNs) have been effectively utilized in tackling intricate challenges, including the identification of cardiac abnormalities based on physiological measures and the detection of breast cancer through mammograms. A proficiently designed

ANN has even demonstrated the potential to outperform physicians in diagnosing patients.[15].

Training an ANN is a crucial aspect of its application. During training, the network adjusts its parameters by iteratively comparing predictions to actual target values, typically facilitated by optimization algorithms minimizing a predefined loss or cost function. This optimization process refines weights and biases, enhancing the network's predictive accuracy.

ANNs showcase impressive capabilities in tasks like classification, regression, image recognition, and natural language processing. They can generalize from patterns in training data to make predictions on new, unseen data. The network's architecture, including the number of layers and neurons, significantly influences its performance, making tuning these aspects crucial for achieving desired outcomes.

3.10 BOOSTING METHODS

Boosting methods stand as a category of ensemble learning techniques in machine learning, designed to enhance the predictive performance of weak base models by iteratively amalgamating them into a robust predictive model. These techniques have garnered substantial attention for their efficacy in handling intricate and high-dimensional data, making them indispensable tools in diverse research and practical applications.

The fundamental concept behind boosting involves sequentially training a series of weak learners, like decision trees with limited depth or simple linear models. In each iteration, the boosting algorithm assigns higher weights to instances that were misclassified or exhibited higher prediction errors in the preceding iteration. This directs subsequent weak learners towards instances that pose greater challenges in accurate classification.

As the iterations progress, the weak learners are amalgamated through a weighted sum or a weighted majority vote, where their individual predictions contribute to the final prediction of the boosted model. The weights assigned to the learners in this combination are determined based on their performance during the training process, with consistently well-performing weak learners exerting more influence on the final model.

AdaBoost (Adaptive Boosting) and Gradient Boosting stand out as two renowned boosting algorithms. AdaBoost adjusts the weights of misclassified instances, enabling subsequent models to focus on these challenging cases. Conversely, Gradient Boosting optimizes subsequent models to minimize errors in the predictions of the previous model, refining the model over iterations.

Boosting methods also provide mechanisms to control overfitting. Techniques such as early stopping and regularization can be applied to prevent the boosting process from overfitting to the training data excessively, preserving its ability to generalize to new data.

3.11 MACHINE LEARNING

Stands as a pivotal branch of artificial intelligence (AI), encompassing the study and development of algorithms and models that enable computers to learn from and make predictions or decisions based on data. Rooted in the concept of pattern recognition and computational statistics, machine learning empowers computers to improve their performance over time through exposure to data without being explicitly programmed.

At its core, machine learning operates by training models on historical data to identify patterns, relationships, and dependencies within the data. These models, often referred to as learners or predictors, generalize from past observations and then apply the learned insights to new, unseen data for making informed predictions or classifications.

Machine learning encompasses a spectrum of techniques, ranging from supervised learning, where models are trained on labeled data, to unsupervised learning, which discovers patterns in unlabeled data, and semi-supervised and reinforcement learning, which bridge the gap between the two. Furthermore, machine learning leverages various algorithms such as decision trees, support vector machines, neural networks, and ensemble methods, each tailored to different types of problems and data.

In contemporary research and applications, machine learning finds wide-ranging utility in fields like natural language processing, computer vision, finance, healthcare, and more. It serves as a cornerstone in data-driven decision-making, automation, and predictive modeling. The growth and advancement of machine learning continue to shape the landscape

of AI, enabling computers to extract meaningful insights from data and make informed judgments, thus contributing to the development of intelligent and autonomous systems.

3.12 DATASET

A dataset refers to a structured collection of raw or processed data points that are systematically organized and grouped together for analysis, exploration, or evaluation. In the context of scientific research, a dataset constitutes a fundamental component in empirical investigations, serving as the foundation upon which various analyses and interpretations are built.

Datasets can encompass diverse types of information, ranging from numerical measurements and categorical attributes to textual content and multimedia resources. These data points are often acquired through observation, experimentation, surveys, simulations, or other data collection methods, depending on the nature of the study. The data within a dataset may be cross-sectional, capturing information at a specific point in time, or longitudinal, tracking changes and variations over time.

A dataset typically comprises multiple variables, where each variable corresponds to a specific characteristic or attribute being measured or recorded. The relationships and interactions among variables can be further explored and analyzed to reveal patterns, trends, and insights within the data.

Datasets play a crucial role in the advancement of scientific knowledge, enabling researchers to test hypotheses, validate theories, and draw evidence-based conclusions. The process of preparing and curating a dataset involves data preprocessing, cleaning, transformation, and validation to ensure its quality, reliability, and suitability for analysis.

In contemporary research, the availability of large and diverse datasets has been facilitated by technological advancements, contributing to the development of innovative methodologies and techniques across various fields of study.

3.13 NORMALIZATION

Normalization in machine learning refers to a preprocessing technique aimed at transforming feature values within a dataset into a consistent and standardized range. The primary

objective of normalization is to enhance the convergence and performance of machine learning algorithms by mitigating the influence of varying scales and magnitudes of features on the learning process.

In the context of feature normalization, each feature is scaled to ensure that it has a similar magnitude and distribution across the dataset. This process eliminates any dominance exerted by features with larger scales, preventing them from disproportionately affecting the model's training and outcomes. Normalization facilitates efficient optimization of model parameters and aids in avoiding convergence issues that may arise due to disparate feature scales.

A common approach to normalization is Min-Max scaling, which linearly transforms each feature to a predefined range, typically $[0, 1]$. This method involves subtracting the minimum feature value from each observation and then dividing it by the range of the feature values. Another technique is Z-score normalization, which transforms features to have a mean of zero and a standard deviation of one. This approach is particularly useful when the distribution of feature values is not uniform or when outliers are present.

Normalization contributes to improving the stability and interpretability of machine learning models, making them less sensitive to variations in input feature scales. It also aids in enhancing the model's ability to generalize and make accurate predictions on new, unseen data.

In summary, normalization plays a crucial role in creating a consistent and standardized foundation for machine learning algorithms, enabling them to effectively learn patterns and relationships within data without being hindered by feature scale discrepancies.

3.14 ENCODING

Encoding in the context of machine learning refers to the process of converting categorical data, which represents distinct and non-numeric attributes, into numerical values that machine learning algorithms can effectively interpret and utilize. Categorical data, such as nominal or ordinal variables, pose a challenge for many machine learning models that require numerical inputs.

Encoding categorical variables is essential to enable machine learning algorithms to analyze and learn patterns from such data. The process involves mapping each category or level of a categorical variable to a unique numerical representation. This numerical transformation preserves the relationships between different categories while allowing algorithms to perform calculations and comparisons.

There are various methods of encoding categorical variables, each with its own characteristics and applications. One common approach is one-hot encoding, which creates binary columns for each category of a variable, indicating its presence or absence. Another technique is label encoding, where categories are assigned integer values based on their order or frequency. Target encoding involves replacing categories with the mean of the target variable for that category. Ordinal encoding assigns integer values to categories based on their predefined order.

The choice of encoding method depends on the nature of the data, the type of categorical variable, and the machine learning algorithm being used. Proper encoding is crucial to ensure that categorical data are appropriately incorporated into the learning process, preventing biases or misinterpretations. Incorrect encoding can lead to misleading results and hinder the performance of machine learning models.

In summary, encoding is a fundamental preprocessing step in machine learning, facilitating the integration of categorical data into numerical models, thereby enabling effective analysis, pattern recognition, and prediction.

3.15 LONG SHORT-TERM MEMORY

Long Short-Term Memory (LSTM) is a type of RNN architecture that has gained prominence in the field of machine learning, particularly in the domain of sequential data analysis and prediction. LSTMs are designed to address the challenge of capturing long-range dependencies and temporal patterns within sequences, which is often challenging for traditional RNNs.

The fundamental innovation of LSTM lies in its ability to selectively retain or forget information over varying time intervals, making it well-suited for tasks involving sequences with gaps or delays between relevant events. This is achieved through a network of

specialized memory cells that allow LSTMs to remember and utilize relevant information for longer durations.

The LSTM architecture includes three primary components: the cell state, input gate, and output gate. The cell state serves as the 'memory' of the LSTM, maintaining and propagating information over time. The input gate regulates the flow of new information into the cell state, while the output gate controls the flow of information from the cell state to the final output.

LSTMs have demonstrated exceptional performance in various applications, such as natural language processing, speech recognition, time series forecasting, and image captioning. Their ability to capture temporal dependencies and learn from complex sequences makes them particularly effective for tasks involving sequential data.

In summary, Long Short-Term Memory (LSTM) is a specialized recurrent neural network architecture that excels in capturing long-term dependencies and patterns within sequential data. Its unique memory cells and gating mechanisms enable it to effectively model sequences with varying time intervals, contributing to its success in a wide range of applications.

3.16 LIGHTGBM

LightGBM (Light Gradient Boosting Machine) is a state-of-the-art machine learning framework that belongs to the family of gradient boosting algorithms. It has gained significant attention in recent years due to its remarkable performance, efficiency, and ability to handle large and complex datasets. LGBM is particularly well-suited for tasks involving tabular data, structured features, and categorical variables.

The fundamental principle behind LightGBM is boosting, a technique that combines multiple weak learners (typically decision trees) into a strong predictive model. LightGBM enhances the traditional gradient boosting approach by introducing innovative features such as the Gradient-based One-Side Sampling (GOSS) technique and Exclusive Feature Bundling (EFB). These optimizations lead to faster training times, lower memory consumption, and improved accuracy.

A notable feature of LightGBM is its histogram-based approach to data binning, which reduces the memory footprint and accelerates the training process. This is achieved by constructing histograms of feature values and using them to efficiently find the best splits during tree construction. Additionally, LightGBM supports categorical feature handling through techniques like categorical split and categorical shift, enabling the effective use of non-numeric attributes.

LightGBM also provides advanced regularization techniques, like L1 and L2 regularization, to prevent overfitting and enhance model generalization. The framework offers flexibility in terms of hyperparameter tuning, enabling practitioners to optimize model performance for specific tasks. Its user-friendly interface and compatibility with various programming languages contribute to its popularity in both research and industry applications.

In summary, LightGBM is a cutting-edge gradient boosting framework that offers exceptional performance, efficiency, and versatility for a wide range of machine learning tasks. Its unique optimizations, histogram-based approach, and categorical feature handling make it a valuable tool for researchers and practitioners working with complex and high-dimensional datasets.

3.17 ROOT MEAN SQUARE ERROR

Root Mean Square Error is a widely used metric in the fields of statistical analysis and data mining, commonly employed to assess the quality of a predictive or regression model. RMSE quantifies the magnitude and consistency of the differences between actual observations and predicted values.

The computation of RMSE involves the following steps:

Calculate the difference between the actual values and the predicted values for each observation.

Square these differences.

Compute the average of these squared values.

Finally, take the square root of this average to obtain the RMSE.

One of the advantages of RMSE is that it treats errors of different magnitudes uniformly, meaning that large errors do not outweigh small errors. The closer the RMSE value is to zero, the closer the model's predictions are to the actual values, indicating better model performance.

RMSE is particularly useful for assessing the quality of regression models and for comparing different models. In our thesis, we employ this metric to analyze the predictive ability and performance of our model.



4. IMPLEMENTATION

In this study, the boosting method known as LightGBM, a machine learning technique, has been employed. In this section, the LightGBM model is discussed, and a predictive model will be developed using a two-year dataset encompassing sales, inventory and product attributes from a ready-to-wear retail company operating in the apparel industry.

4.1 DATA COLLECTION

Data pertaining to a retail company engaged in the sale of ready-to-wear apparel has been retrieved from a database using SQL. This data encompasses store information, product details, as well as sales, inventory, and capacity data from the last two years. Subsequently, the acquired data was transferred to the Python environment for further analysis.

4.2 STORE INFORMATION

Details concerning the geographical locations and stylistic attributes of the stores have been procured.

4.3 PRODUCT INFORMATION

Information regarding the nature of the products, their range categorizations, colour specifications, seasonal attributes, and class associations has been collected.

4.4 SALES INFORMATION

Insights into the weekly sales volumes of products across stores during the preceding two years have been acquired.

4.5 INVENTORY INFORMATION

Data regarding the quantities of products held in stores, including storage facilities, has been obtained.

4.6 CAPACITY INFORMATION

Details regarding the spatial occupancy of product categories within stores, encompassing both current and potential maximum occupancy, have been gathered.

4.7 DATA PREPROCESSING

The applied procedures on the data encompass the following:

4.7.1 Data Transformation

In order to standardize the data and ensure that they do not exhibit any inherent advantage over each other, apart from the dependent variable containing numerical data, all data were scaled to a range between 0 and 1 (see normalization).

4.7.2 Enhancing Data Significance

To amplify the significance of the employed data and facilitate accurate prediction, an encoding process was implemented (see encoding).

4.7.3 Handling Missing Data

Additionally, the imputation of missing data using the mean value of their corresponding product family not only enhanced the comprehensiveness of the dataset but also contributed to a more robust and representative analysis. This approach aimed to minimize the impact of missing values on the overall integrity of the research findings.

4.8 APPLICATION OF MACHINE LEARNING WITH LIGHTGBM

In this study, a machine learning application using LightGBM method was conducted. A prediction model was applied on prepared data in accordance with the LightGBM machine learning method. To enhance the accuracy and robustness of the predictions, adjustments were made to the parameters of the method. The modified parameters are provided below.

Table 4.1: Important Parameters.

boosting_type	This parameter specifies the type of boosting algorithm to be used, such as "gbdt" (Gradient Boosting Decision Tree) or "dart."
num_leaves	This parameter determines the maximum number of leaf nodes for each decision tree. Larger values can increase model complexity but also pose a risk of overfitting.
learning_rate	It controls the rate at which weights are updated in each iteration. A lower learning rate can lead to a more stable model, albeit with slower learning.
n_estimators	Specifies the total number of boosting iterations (number of trees). Generally, more trees are required for better results, but this can also increase the risk of overfitting.
max_depth	Limits the maximum depth of each decision tree. Lower values result in a simpler model and mitigate the risk of overfitting.

5. RESULTS

In this study, the results obtained from the time series using the LightGBM model have been analyzed."

Table 5.1: Example Data.

MerchYearWeek	StoreRef	OptionRef	SalesForecastQuantity	SalesQuantity
202336	2634	2694479	2	2
202336	2634	2743710	4	3
202336	2634	2822282	1	1
202336	2634	2822319	1	0
202336	2634	2682213	3	7

In time series models, the consistent flow of data at regular intervals is one of the most critical parameters. Another significant aspect, as observed in this study, is that in the retail sector, forecasting relies not only on recent data but also on historical data over extended periods. Therefore, LightGBM model stands out as a more accurate model compared to other prediction methods for several reasons. It exhibits resilience against gapped data and excels in thoroughly examining all periods within the given time frame.

As depicted in Figure X, it is noticed that the accuracy of the LightGBM model declines during periods of high sales volume. This phenomenon can be attributed to the fact that the general dataset is dominated by low sales volumes, causing the model to perceive high-volume sales as outliers and consequently adjusting the predictions downward.

When optimization was performed on the LightGBM model parameters, it was observed that more accurate results were achieved.

6. DISCUSSION AND CONCLUSIONS

Based on the findings of this study, the following recommendations are provided to enhance the accuracy and effectiveness of time series forecasting in the retail sector:

6.1 DATA QUALITY ENHANCEMENT

Improving data quality, especially during periods of high sales volume, can enhance the accuracy of predictions. Identifying and rectifying outliers can contribute to the model's sensitivity in such situations.

6.2 FURTHER FEATURE ENGINEERING

Further feature engineering efforts can assist the model in making better predictions. Specifically, new features should be created to capture trends and patterns in specific periods of historical data.

6.3 MODEL OPTIMIZATION

Optimizing the parameters of the LightGBM model can lead to sharper predictions. This can help distinguish periods in which the model performs well from those in which it performs less effectively.

6.4 HYBRID MODELS

To achieve improved predictions during high sales volume periods, consider the use of hybrid models. The combination of multiple models can provide better alignment with different data periods.

6.5 DATA COLLECTION AND MONITORING

Establish continuous data collection and monitoring processes to adapt to changing consumer behaviors and market conditions in the retail sector. This can enable the model to make predictions based on up-to-date data.

6.6 LITERATURE REVIEW

Continue to monitor recent developments and best practices in time series forecasting methods by staying updated with the latest literature. Focus on continually enhancing your model based on the most recent advances in the field.

6.7 SEPARATE MODELS FOR OUTLYING BEHAVIOR

Consider building separate models for groups that exhibit outlier behavior to achieve more accurate results.

6.8 UTILIZE BOOSTING METHODS IN NON-PERIODIC DATA

In cases where data does not exhibit periodic behavior, the use of boosting methods like LightGBM can improve prediction accuracy.

REFERENCES

- [1] S. Pinar, “SALES FORECASTING WITH ARTIFICIAL NEURAL NETWORK AT A APPAREL RETAILING COMPANY,” M.S. thesis, Yıldız Technical Univ., Istanbul, Turkey, 2012.
- [2] S. Ozturk, “FORECASTING OF SALES QUANTITIES BY MACHINE LEARNING METHODS IN FAST FASHION SECTOR,” M.S. thesis, Yıldız Technical Univ., Istanbul, Turkey, 2020.
- [3] G. Surmeli, “FORECASTING THE PROMOTIONAL DEMAND OF RETAIL PRODUCTS AND THE APPLICATION OF A FORECASTING MODEL,” M.S. thesis, Istanbul Technical Univ, Istanbul, Turkey, 2013.
- [4] G. Andres, “NEW PRODUCT DEMAND FORECASTING AND DISTRIBUTION OPTIMIZATION : A CASE STUDY AT ZARA, “ M.S. thesis, Massachusetts Institute of Technology Univ, Massachusetts, USA, 2011.
- [5] M. Jenni, “DEMAND FORECASTING IN THE APPAREL INDUSTRY, “ B.S. thesis, Helsinki Metropolia Univ, Helsinki,Finland,2014.
- [6] D. Vi Phung-Thao “DEMAND FORECASTING FOR BOUGHT-IN MATERIALS USING TIME-SERIES METHODS,” B.S thesis, Haaga Helia Univ, Tampere,Finland,2019.
- [7] Z. Nadezhda “DEMAND FORECASTING OF AN INDUSTRIAL ITEM: A METHODOLOGY STUDY,” B.S thesis, Jamk Univ, Espoo,Finland,2019.
- [8] J. Dikker “BOOSTED TREE LEARNING FOR BALANCED ITEM RECOMMENDATION IN ONLINE RETAIL,” M.S. thesis, Eindhoven Univ, Eindhoven, Holland,2017
- [9] N. MARAGKOS “RETAIL DEMAND FORECASTING WITH ARTIFICIAL NEURAL NETWORKS,” M.S thesis,Aristotle Univ, Thessaloniki,Greece,2020

- [10] K. Terhi Kristiina “TEXTILE IN ARCHITECTURE,” M.S thesis,Tampere Univ,Tampere,Finland,2010
- [11] R. Roberts “MEASURING NAVSPASUR SENSOR PERFORMANCE USING LOGISTIC REGRESSION MODELS,”,M.S thesis,Naval Univ,Washington, USA,1992
- [12] R. Roberts “MEASURING NAVSPASUR SENSOR PERFORMANCE USING LOGISTIC REGRESSION MODELS,”,M.S thesis,Naval Univ,Washington, USA,1992
- [13] S. Suraya “DESIGNING AND IMPLEMENTING A DATABASE FOR THESIS DATA MANAGEMENT BY USING THE PYTHON FLASK FRAMEWORK”, M.S thesis,Akprind Univ, Yogyakarta,Indonesia,2022
- [14] S. Björklund “ARTIFICIAL NEURAL NETWORKS FOR FINANCIAL TIME SERIES PREDICTION AND PORTFOLIO OPTIMIZATION,”,M.S thesis, Linköping Univ,Linköping,Sweden,2017
- [15] S. Björklund “ARTIFICIAL NEURAL NETWORKS FOR FINANCIAL TIME SERIES PREDICTION AND PORTFOLIO OPTIMIZATION,”,M.S thesis, Linköping Univ,Linköping,Sweden,2017