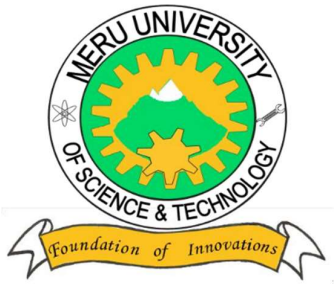

MERU UNIVERSITY OF SCIENCE AND TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE

BACHELOR OF SCIENCE IN DATA SCIENCE

Retail Demand Prediction Using Random Forest Model

CT204/103868/20
SEGEWA SUNNY KAGAME

A Research Proposal Submitted in Partial Fulfillment of the Requirements of the Bachelor of Science in Data Science of Meru University of Science and Technology

December, 2023

DECLARATION

This research proposal is my original work prepared with no other than the indicated sources and support, and has not been presented elsewhere for a different or similar assignment.

CT204/10386/20. _____

Segewa Sunny Kagame 04/12/2023

TABLE OF CONTENTS

TABLE OF FIGURES

CHAPTER ONE

INTRODUCTION

1.1 Background of study

Demand forecasting in supply chain management can be described as the process of predicting the level of interest customers will have in current products or services, in an attempt to deliver the right products in the right amounts to meet the needs of customers without creating surplus inventory.

Historically, demand prediction in retail relied on traditional statistical methods. However, advancements in data analytics and machine learning have transformed the landscape. Modern retailers harness past sales data, market trends, weather patterns, and even social media sentiment analysis to refine their demand prediction models (Seyedan & Mafakheri, 2020). Industry leaders like Amazon and Walmart employ sophisticated algorithms to predict customer demands, optimize inventory, and enable same-day deliveries, setting benchmarks for the sector.

The supply chain data is characterized by its multi-dimensional nature, originating from various points of the supply chain network and serving diverse purposes such as tracking products, assessing supplier capacities, managing orders, monitoring shipments, and understanding customer and retailer interactions. The abundance of suppliers, products, and customers results in large volumes of data, and the continuous processing of numerous transactions within supply chain networks contributes to its high velocity. Given these intricacies, there has been a shift away from traditional (statistical) demand forecasting methods that rely on identifying statistically significant trends, typically defined by mean and variance attributes within historical data. Instead, the trend is moving towards

intelligent forecasting approaches that can adapt and evolve based on learning from historical data, enabling them to effectively anticipate the dynamic and ever-changing demand patterns in supply chains. (Seyedan & Mafakheri, 2020)

A crucial element to consider in supply chain management is reducing the Bullwhip Effect. The bullwhip effect in supply chain management describes the amplification of demand fluctuations as you move upstream in the supply chain. Small changes in consumer demand can result in exaggerated variations in orders and inventory levels at different points in the supply chain. Technologies like real-time data analytics and machine learning can be used to mitigate this factor. Hence, it is essential to create dependable models for forecasting demand in order to enhance the precision and quality of predictions as described by (Aamer et al., 2021).

Forecasting demand is crucial within the retail sector, given the widespread adoption of ecommerce services and evolving consumer tastes, accurate predictions are essential in order to meet customer expectations. Retailers depend on demand forecasting to fine-tune inventory management, cut carrying expenses, prevent stock shortages, and customize marketing approaches for maximum effectiveness. Inaccurate predictions lead to dissatisfied customers, lost sales, and increased operational inefficiencies. Hence, mastering demand prediction is central to providing seamless customer experiences and ensuring the long-term sustainability of retail businesses.

1.2 Motivation for study

The motivation for conducting the research on demand prediction in retail stems from the transformative potential it holds for businesses in the face of modern market challenges.

Retailers grapple with the critical task of keeping track of products to ensure they are available when and where customers desire them. By leveraging advanced technologies, including big data analytics and machine learning algorithms, to unravel intricate patterns within vast datasets, retailers can move beyond traditional forecasting methods, embracing predictive models that adapt to evolving consumer behavior. Therefore, it paves the way for data-driven decision-making and innovation within the retail sector.

1.3 The statement of the Problem

Retailers face a challenge in maintaining a balance between tying up cash due to excess inventory and running out of items resulting in stock outs or unfulfilled orders. This is due to sudden shifts in customer preferences. Accurate forecasting of consumer demand is pivotal for optimizing inventory levels, helping retailers make better informed decisions, and estimating total anticipated sales and revenue.

A lack of proactive product demand anticipation not only leads to excess inventory and lost sales due to stock-outs but also may lead to frequent price changes or promotions to clear excess inventory. This can erode profit margins and diminish the perceived value of products. To address this challenge, my research proposes the development and implementation of a machine learning and time series-based model capable of predicting product demand in the retail sector. The model will empower retailers with the knowledge and insights needed to implement data-driven product demand prediction strategies, ensuring appropriate inventory levels. In turn, space utilization in the warehouse is optimized so that the fastest moving products are the easiest to fulfill. With proper demand forecasting, retailers can better position themselves to meet customer needs, thereby maximizing profitability, decreasing overhead costs, and increasing customer satisfaction.

1.4 Research objectives

1.4.1 General objectives

To estimate product demand by utilizing machine learning techniques and historical sales data to create a reliable and accurate demand prediction model.

1.4.2 Specific objectives

The objectives of this research project are:

i. Develop a predictive model to forecast product sales in order to anticipate their demand at a particular time
ii. To determine the key features that affect a product's sales volume.
iii. Evaluate the developed demand prediction model. Assess its accuracy, reliability, and adaptability to different product categories.
iv. Develop an interactive web application to forecast future sales of the products available.

1.5 Significance of the study

The contribution of this project would be of interest to retailers by offering accurate demand prediction whereby retailers can leverage optimal strategies to manage their inventory. This optimization leads to reduced stock-outs and improved product availability. Retailers can make data-driven decisions on procurement, pricing, and marketing, leading to increased operational efficiency.

1.6 Scope of the study

This research project will have a focus on demand forecasting for a tech-gadget e-commerce retailer. Specifically, it aims to leverage machine learning algorithms to predict future sales based on historical data. The primary dataset for the study includes weekly sales data for an e-commerce retailer spanning 100 weeks, from October 2016 to September 2018. This dataset contains 44 unique stock-keeping units (SKUs), each of which represents a distinct item sold by the retailer, as well as a variety of SKU-related information. The first 70 weeks of data will be used as the training dataset, while the last 30 weeks of data will serve as the testing dataset for evaluating the performance of the model.

The study is not limited to tech-gadget e-commerce companies; rather, its goal is to provide insights into enhancing demand forecasting accuracy for retailers generally. On a side note, the business environment in question might require a slight modification of the techniques and feature set used for predicting demand.

1.7 Assumptions in the study

The research makes several key assumptions to establish a foundation for its methodology and findings. Firstly, it assumes the availability and accuracy of historical sales data to be comprehensive and reliable. The analysis and predictions heavily rely on the quality and completeness of this dataset.

The study also presupposes uniformity in data collection methods across various retail sources. Any variations or biases in these methods may influence the accuracy of the predictions. Additionally, the research assumes the ongoing relevance and consistency of

external factors, such as market trends, natural events, and seasonal patterns, throughout the research period. Significant changes in these elements might have the potential to influence the precision of forecasts regarding demand.

Lastly, the study assumes a certain level of consistency in customer behavior patterns during the research period. Changes in customer preferences, buying habits, or socio-economic factors have the capability to impact the accuracy of demand forecasts. These assumptions collectively establish a foundation on the methodology employed in the study and the subsequent interpretation of findings.

1.8 Limitations of study

The research project, despite its comprehensive approach, is subject to certain limitations that warrant consideration. Firstly, the temporal scope of the study, covering a specific period from October 2016 to September 2018, implies a potential limitation in capturing the most recent dynamics in external factors, market trends, and consumer behaviors. Any changes or developments post-2018 may not be fully reflected in the study, affecting the model's adaptability to current conditions.

Another noteworthy limitation is inherent in the assumption of uniformity in the methodology used in collecting the data across different retail sources. This may face challenges if variations exist in how data is collected or recorded, potentially impacting the integration and analysis of information. The research acknowledges the dynamic nature of external factors, but the assumption that these factors remained consistent during the research period might be limiting if significant changes occurred.

Furthermore, the study's reliance on the assumption of consistent customer behavior patterns may be constrained by unforeseen shifts in consumer preferences, purchasing habits, or socio-economic factors not accounted for in the dataset. The limited feature set available in the dataset could also pose a limitation, as certain relevant variables may have been omitted, affecting the model's predictive power. Lastly, the models developed during the training phase may face challenges in generalizing well to unseen data, impacting their performance on the testing dataset.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

This chapter will provide an insight into the existing body of knowledge and scholarly work that forms the foundation for the study. The review serves as a critical exploration of relevant academic and industry literature, offering insights into the theoretical frameworks, methodologies, and findings that shape the research landscape on demand forecasting in the retail sector. This chapter aims to provide a comprehensive overview of key concepts, models, and trends related to demand forecasting. By synthesizing existing literature, the research seeks to identify gaps, challenges, and opportunities that will inform in developing a robust and contextually relevant machine learning model for the prediction of product demand in the specified tech-gadget e-commerce retail setting.

2.2 Theoretical Literature

Supply chain management involves overseeing the movement of products from their origins to customers through a connected network of entities and activities. This includes the flow of goods from raw materials passing through the suppliers to the manufacturer and the distribution of the finished goods by the distributor to the retailers who ensures it reaches to the consumers.

There exists a number of studies of supply chain management and demand forecasting applications. However, this research will be focusing on demand prediction in the retail sector in the supply chain network, it will employ the technology of hybrid models in forecasting product demand.

Recent developments in computing and information technology offer enormous potential for demand prediction. Collecting transactional data and developing strategies to utilize this data for an extensive selection of retailers, ranging from electronics to fast fashion to food delivery services, has become routine.

The approaches for anticipating demand have evolved dramatically over time. In the past, conventional statistical techniques such as time series analysis and causal models were extensively used. More complex ways have emerged however, with the introduction of modern technologies and the availability of large data. These consist of machine learning models, hybrid models that blend machine learning and conventional statistical methodologies, and deep learning models.

Despite the developments, there are conflicts among various sources about the efficacy of these approaches. Whereas some research indicates better performance of machine learning models to demand forecasting than classic statistical models, others maintain that traditional techniques can still be useful if utilized appropriately. Moreover, there exists a research gap regarding the optimal methods for particular sectors or situations.

The focus of this literature review is to address these gaps and conflicts. The goal is to evaluate the efficiency of various demand forecasting techniques and identify the approaches that work best in certain sectors or situations. This review aims to give a thorough grasp of the state of knowledge on demand forecasting techniques.

2.3 Big Data Analytics and Algorithms in the Supply Chain

Big data analytics as used in supply chain involves the collection, processing, and analysis of large amounts of data to gain insights, optimize operations, and make informed

decisions in inventory management. The term "big data" refers to the massive volume, variety, and velocity of data generated in the modern business environment.

All phases of supply chains, including manufacturing, sales management, warehousing, logistics, and transportation, have used Big Data Analysis. Prescriptive, predictive, and descriptive analytics makeup big data analytics. Prescriptive, predictive, and descriptive analytics makeup big data analytics. Descriptive analysis involves the depiction and classification of past events. Predictive analytics, on the other hand, aim to forecast future occurrences and identify predictive trends in data. While prescriptive analytics utilize data and mathematical algorithms to inform decision-making. In their study, (Nguyen et al., 2018) examined the types of Big Data Analytics models utilized in supply chain management and the Big Data Analytic methodologies employed in their development.

2.3.1 Linear Regression and Neural Networks

(Huang et al., n.d.) conducted an assessment using linear regression analysis and a backpropagation (BP) neural network to predict e-logistics demand in both rural and urban areas of China. The comparison of mean absolute error (MAE) and average relative errors between the backpropagation neural network and linear regression revealed that the former exhibited greater accuracy, indicating smaller disparities between predicted and actual data. This heightened accuracy is attributed to the utilization of a sigmoid function in the hidden layer of backpropagation, which, unlike linear regression suitable for linear problems, is differentiable and more effective in addressing nonlinear situations as observed in their case study.

While deep learning methods have the potential to generate highly precise predictions in specific scenarios, their effectiveness may be hindered in other cases by a lack of data. The performance of these techniques is typically tied to the amount of data accessible. Artificial neural networks and transformers, which are deep learning approaches, can prove beneficial for extensive retail establishments characterized by substantial transaction volumes and velocity.

2.3.2 Time series

Time series analysis involves the examination of sequential data, which comprises extended sequences of numerical information recorded at consistent time intervals, such as per minute, per hour, or per day. Time series methods are employed to identify trends and seasonal patterns within the data, enabling the detection of hidden and emerging patterns in product demand. Some of the technologies developed to deal with time series data include Autoregressive Integrated Moving Average (ARIMA), Seasonal Autoregressive Integrated Moving Average (SARIMA), Prophet and Holt-Winters (HW) among others.

A variety of perishable dairy products' demand was predicted by (Veiga et al., 2014) by comparing the Holt-Winters (HW) and Autoregressive Integrated Moving Average (ARIMA) models. The Teil inequality index (U-Teil) and mean absolute percentage error (MAPE) metrics were used in comparing the accuracy of both algorithms on the data. The Holt-Winters model exhibited a superior fit according to both performance metrics. It is characterized by its simplicity and ease of use. The study noted that it is important that the data horizon should not exceed a seasonal cycle; exceeding this limit would result in a significant decrease in forecast accuracy. This is due to the fact that a Holt-Winters model's

inputs are inherently projected values that could contain longer-term uncertainties and mistakes.

2.3.3 Hybrid Models

In hybrid models for demand forecasting, both time-series and regression (or machine learning methods in recent) are used to model the demand patterns. A hybrid model composed of a combination of the Seasonal Autoregressive Integrated Moving Average (SARIMA) and Random forests is proposed in the research for predicting the demand of products in e-commerce retail.

Seasonal Autoregressive Integrated Moving Average (SARIMA) is a variant of the ARIMA time series-forecasting model that is capable of handling seasonality in the data. SARIMA can be particularly useful for forecasting data with strong seasonal trends. SARIMA models incorporate three main components: seasonal autoregression (SAR), seasonal differencing (D), and seasonal moving average (SMA). The SAR component models the influence of previous observations from the same season, while the D component accounts for seasonality by differencing the data at the seasonal frequency. Finally, the SMA component models the error terms as a weighted sum of past errors at the same seasonal frequency. The parameters for the SARIMA model are typically estimated using maximum likelihood estimation. SARIMA models are useful for forecasting time series data that exhibit a seasonal pattern, such as monthly or quarterly data.

Random Forest is an ensemble learning technique that constructs multiple decision trees during training, it can be used with SARIMA for time series forecasting as it excels at

handling non-linear relationships in the data. It is particularly valuable when demand patterns exhibit complexity and are influenced by various factors beyond traditional time series components. When combined with SARIMA, a time series forecasting method, a hybrid approach emerges, creating a robust and adaptable solution that leverages the strengths of each method.

2.4 Research gaps

A notable gap identified in the research is the minimal investigation of hybrid models for demand forecasting with a focus on the retail industry. While previous theoretical research has explored a range of forecasting methodologies and models, there is a notable lack of thorough studies regarding the effectiveness and implementation of hybrid models that combine both contemporary machine learning techniques like Random Forest and traditional time series methods like SARIMA in the retail sector.

2.5 Summary

The literature review provided a thorough investigation into diverse forecasting methodologies, encompassing big data analytics, traditional time series technologies like Holt-Winters and ARIMA, machine learning approaches including linear regression and neural networks, as well as general time series models. The analysis highlighted the transformative impact big data analytics has provided in supply chain management, and explored the complexity-handling capabilities of machine learning.

A significant observation within the review was the research gap pertaining to the underexplored realm of hybrid models in the retail sector. This identified gap underscores the need for comprehensive investigations into integrating various forecasting techniques,

with hybrid models emerging as a promising choice to enhance forecasting accuracy and adaptability in dynamic retail environments.

CHAPTER THREE

RESEARCH METHODOLOGY

3.1 Introduction

In this chapter, the research delves into the methodology employed in the execution of the project. Methodology serves as the backbone of any research or project endeavor, guiding the process from inception to conclusion. The research aims to provide a clear understanding of the systematic approach undertaken to address the research objectives and achieve the desired outcomes of the project.

3.2 Data Collection

3.2.1 Data Sources:

The data utilized for the research is a secondary data, the data is derived from pre-existing datasets that are publicly accessible. The data is sourced from https://demandprediction.github.io/dataset.html website, and was originally provided by an online electronics retailer.

3.2.2 Sample of the data

The image below shows the first five rows of the dataset:

	week	sku	weekly_sales	feat_main_page	color	price	vendor	functionality
0	2016-10-31	1	135.0	True	black	10.16	6	06.Mobile phone accessories
1	2016-11-07	1	102.0	True	black	9.86	6	06.Mobile phone accessories
2	2016-11-14	1	110.0	True	black	10.24	6	06.Mobile phone accessories
3	2016-11-21	1	127.0	True	black	8.27	6	06.Mobile phone accessories
4	2016-11-28	1	84.0	True	black	8.83	6	06.Mobile phone accessories

Figure 1 Data Sample

3.2.3 Description of the data:

The data can be extracted and is available in CSV format. This dataset reports the weekly sales of a tech-gadget e-commerce retailer over a period of 100 weeks, from October 2016 to September 2018. It includes the weekly sales of 44 items, also called stock-keeping units (SKUs), as well as diverse information on these SKUs.

The dataset comprises of 4400 rows and 8 columns. Each row corresponds to a SKU week pair (44 SKUs for 100 weeks), whereas each column corresponds to a feature.

The features of this dataset are described below:

- Week: The dataset covers all weeks from 2016-10-31 to 2018-09-24. In total, the data has 100 weeks (i.e., approximately two years of data).
- SKU: There are 44 SKUs, indexed from 1 to 44. In total, the dataset has 44 * 100 = 4400 rows. An SKU (Stock Keeping Unit) is a unique code or identifier used in the retail industry to distinguish and track individual products within a store or

inventory by encapsulating specific attributes such as size, color, and brand, facilitating efficient inventory management and differentiation.

- Featured on the main page: To boost the sales of specific products, the marketing team may decide to broaden their visibility by featuring these products on the website's homepage. The data records for each week and SKU whether this was the case (i.e., binary indicator).
- Color: Describes the color of products sold.
- Price: The price is fixed for each item during a given week. The pricing team can adjust the price on a weekly basis based on various considerations (e.g., promotional events, excess amounts of inventory).
- Vendor: The Company acts as a retailer for electronics brands. The vendor variable refers to the product brand. The SKUs in our dataset span 10 different vendors.
- Functionality: The functionality is the main function or description of the SKU. Specifically, there are 12 different functionalities, which correspond to the following categories: streaming sticks, portable smartphone chargers, Bluetooth speakers, selfie sticks, Bluetooth tracker, mobile phone accessories, headphones, digital pencils, smartphone-stands, virtual reality headset, fitness trackers, and flash drives.
- weekly_sales: This is the number of items sold during the focal week for the corresponding SKU. This is the variable to be predicted, often called the target or outcome variable.

3.3 Data Pre- Processing

The section outlines the various techniques employed to prepare and clean the raw data for analysis. The data pre-processing steps undertaken include:

3.3.1 Missing Value Imputation

Missing values within the dataset were addressed using the SimpleImputer module from the scikit-learn library. The most frequent strategy was adopted to impute missing values in the 'color' attribute for each SKU. This involved identifying missing values, fitting the imputer to the available data for the respective SKU, and then transforming the missing values based on the most frequent color value within each SKU group.

3.3.2 Handling Outliers

Outliers within the dataset were identified and removed on a per-SKU basis. Statistical measures such as mean and standard deviation were calculated for attributes like 'price' and 'weekly_sales' within each SKU group to identify outlier data points. Data points falling outside a threshold of three standard deviations from the mean were considered outliers and subsequently removed from the dataset.

3.3.3 Feature Engineering

Feature engineering techniques were applied to derive additional features from the existing dataset. This involved the creation of new attributes such as 'avg_price', 'price_difference', and lagged price differences. These engineered features aimed to capture valuable insights and patterns within the data for subsequent analysis.

3.3.4 Feature Scaling

To ensure uniformity and comparability across numerical features, feature scaling was performed using the StandardScaler from the scikit-learn library. Numerical columns relevant for scaling were identified, and the data was transformed to have a mean of 0 and a standard deviation of 1.

3.3.5 Encoding Categorical Variables (One-Hot Encoding)

Categorical variables such as 'functionality' and 'color' were encoded using one-hot encoding to convert them into numerical format suitable for machine learning algorithms. This process expanded each categorical attribute into binary columns, representing the presence or absence of each category. It creates one binary attribute per category: one attribute equal to 1 when an item belongs to a category and 0 otherwise.

3.4 Feature Extraction and Selection

Feature extraction process transforms the original data into a reduced number of features and selection tries to find the element that gives the most information about the problem. Initially, the dataset comprised 33 independent variables encompassing various aspects of the data. The objective was to optimize the feature space to enhance the performance of the predictive model.

Initial Approach: Began the analysis by utilizing all available features in our random forest model. This approach provided a comprehensive understanding of the dataset's predictive capacity and served as a baseline for subsequent optimization techniques.

The **RandomForestRegressor** class in scikit-learn comes with a built-in **feature_importances_** attribute which provides information about the relative importance of each feature in the random forest model.

After fitting the **RandomForestRegressor** model to your data, you can access the **feature_importances_** attribute to retrieve the importance scores assigned to each feature. These scores indicate the contribution of each feature to the predictive performance of the model.

The image shows the rank of each feature in predicting the target variable in descending order:

Feature	Importance
price	4.469968e-01
sku	1.454906e-01
price_lag_1	7.145730e-02
price_lag_2	5.434108e-02
avg_price	5.254022e-02
price_difference	4.836916e-02
price_lag_difference_1	3.905975e-02
color_grey	3.606040e-02
price_lag_difference_2	2.740129e-02
color_green	2.395449e-02
week_number	1.602619e-02
month	7.121306e-03
vendor	6.840727e-03
functionality_portable smartphone chargers	5.791850e-03
functionality_mobile phone accessories	4.518610e-03
year	3.791874e-03
feat_main_page	3.146762e-03
color_red	2.909647e-03
functionality_fitness trackers	1.357397e-03
functionality_selfie sticks	1.234439e-03
functionality_smartphone stands	7.389893e-04
color_blue	5.466292e-04
color_none	1.082731e-04
functionality_streaming sticks	7.922647e-05
functionality_flash drives	4.982846e-05
functionality_bluetooth tracker	4.633206e-05
functionality_digital pencils	9.328171e-06
color_pink	6.747377e-06
color_white	2.320046e-06
color_gold	2.196788e-06
color_purple	1.055131e-07
functionality_headphones	6.419708e-08
functionality_vr headset	5.349083e-08

Figure 2 Feature importance

Feature Space Optimization: To improve model performance and mitigate the potential for overfitting, we conducted feature selection utilizing grid search. Specifically, we focused on tuning the **max_features** parameter of the random forest algorithm. By systematically varying **max_features** and evaluating model performance, we aimed to identify the optimal subset of features that maximized predictive accuracy while minimizing model complexity.

Results: Through grid search optimization, it was determined that a max_features value of 17 yielded the most favorable balance between model performance and complexity. This selection effectively reduced the feature space from the original 33 independent variables to a more refined subset, enhancing the efficiency and interpretability of the model without sacrificing predictive accuracy.

3.5 Model Development and Training

3.5.1 Model Development Requirements

The development of the predictive model necessitated a combination of hardware and software resources to facilitate the implementation, training, and evaluation processes.

Hardware Requirements:

1. Computational Resources: A computer with sufficient computational capabilities to handle data preprocessing, model training, and evaluation tasks efficiently.
2. Memory (RAM): Adequate RAM to accommodate the size of the dataset and the computational demands of the machine learning algorithms utilized.

3. Storage Space: Sufficient storage space to store datasets, code files, and model artifacts generated during the development process.

Software Requirements:

1. Programming Language: Python programming language was conducive to perform data manipulation, analysis, and machine learning implementation.
2. Development Environment: Utilization of integrated development environments (IDEs) conducive to software development and data analysis. Popular choices include Jupyter Notebooks and Visual Studio Code.
3. Machine Learning Libraries: Access to machine learning libraries and frameworks for model development and training. The project primarily leveraged scikit-learn, a powerful machine learning library in Python, for its extensive range of algorithms and utilities.

3.5.2 Model Development Algorithm

In our endeavor to construct an effective unit sales predictive model, various machine learning algorithms were evaluated to identify the most suitable approach for our dataset and objectives. The selection of the model development algorithm was driven by the characteristics of our data, the nature of the problem, and the desired predictive performance.

Initial Exploration:

The initial exploration involved the application of a Seasonal Autoregressive Integrated Moving Average (SARIMA) model, a powerful technique for time series forecasting.

However, upon conducting autocorrelation function (ACF) analysis on our data, it became evident that many SKUs exhibited characteristics of white noise, rendering traditional time series modeling less effective for our purposes.

The plot below shows the ACF for SKU 18, the autocorrelation at all lags are predominantly clustered around zero, indicating no correlation between successive observations.

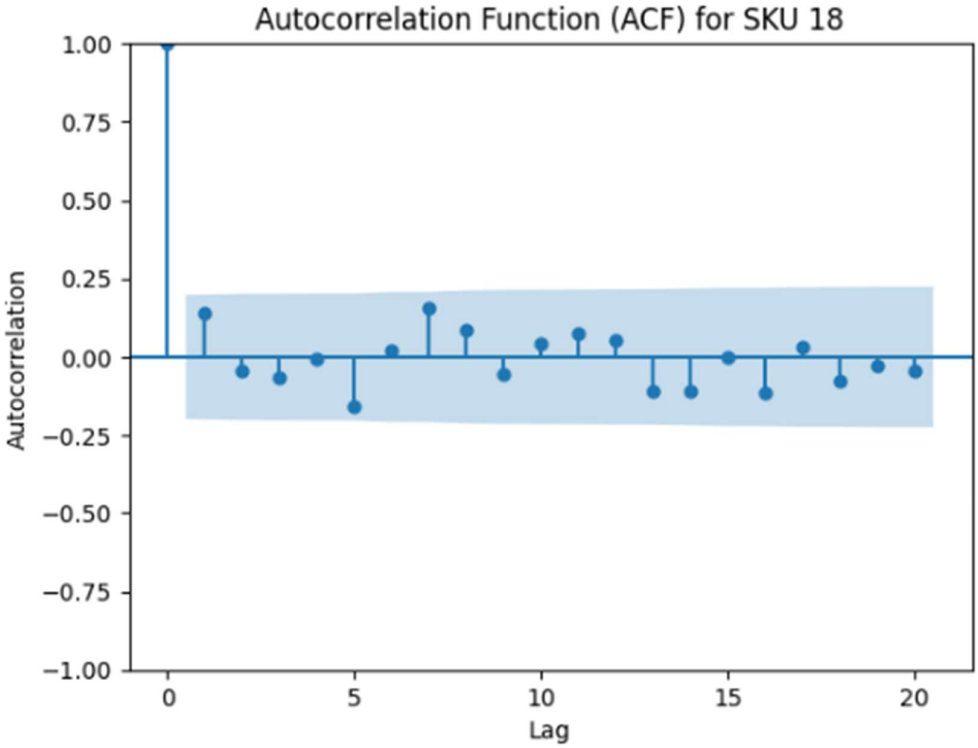

Figure 3 ACF Plot

White noise is a special type of time series where each data point is independent and identically distributed with a mean of zero and constant variance. In white noise, the autocorrelation at all lags are close to zero. It is characterized by randomness and lack of structure. Therefore, if the ACF plot appears as a flat line or a series of random spikes that hover around zero without any discernible pattern, it's indicative of white noise.

Machine Learning Approach:

Subsequently, we transitioned to a machine learning paradigm, recognizing its versatility and ability to capture complex patterns present in the dataset. After careful consideration and experimentation, it was opted for the following algorithms:

1. **Random Forest Algorithm:**

The random forest algorithm was chosen for its robustness, scalability, and ability to handle high-dimensional data with complex relationships. It excels in both regression and classification tasks, making it suitable for our predictive modeling needs. Additionally, random forests are less prone to overfitting compared to individual decision trees, providing a balance between accuracy and generalization.

2. **AdaBoost Regressor:**

AdaBoost (Adaptive Boosting) is an ensemble learning technique that combines multiple weak learners to create a strong predictive model. We employed AdaBoost Regressor for its capability to improve upon the performance of weak learners by focusing on instances that are difficult to predict. This algorithm adapts iteratively to emphasize the data points that were misclassified in previous iterations, enhancing the overall predictive accuracy.

3.5.3 Training and Optimization

To ensure temporal consistency and account for inherent sequential dependencies within the data, a time-based splitting strategy was employed to partition the dataset into distinct training and testing subsets. This approach involved delineating the dataset based on

chronological order, with the training set encompassing historical data preceding a specified cutoff date and the test set comprising more recent observations.

Time-Based Splitting Details:

1. **Training Data:**
    - **Duration:** From 2016-11-14 to 2018-05-14
    - **Number of Rows:** 3356
    - **Filename:** training_data.csv
    - The training set contains historical data spanning from November 14, 2016, to May 14, 2018. This subset, consisting of 3356 rows, served as the foundation for model learning and parameter estimation, capturing patterns and trends in the data up to the designated cutoff date.
2. **Testing Data:**
    - **Duration:** From 2018-05-14 to 2018-09-24
    - **Number of Rows:** 839
    - **Filename:** testing_data.csv
    - The testing set comprises more recent observations ranging from May 14, 2018, to September 24, 2018, totaling 839 rows. Withheld from both the training and validation phases, this subset served as an independent benchmark to evaluate the model's predictive accuracy on unseen data.

Model Training and Optimization:

The training of our predictive models encompassed the following steps:

1. **Parameter Optimization**:

Randomized Search/Grid Search was employed to explore the hyper parameter space of our models systematically. By specifying a range of hyper parameters and evaluating the performance of the model across different combinations, the optimal configuration that maximized predictive accuracy and minimized overfitting was identified.

2. **Training Process:**

Algorithm Implementation: The selected machine learning algorithms (random forest and AdaBoost regressor) were implemented. The model was trained using the training dataset obtained through the data splitting process.

3. **Model Evaluation:**

After training, the models were evaluated using the testing dataset to assess their performance and generalization capabilities. The evaluation metrics considered included mean absolute error (MAE), and R-squared (R^2) to gauge predictive accuracy and model fit.

CHAPTER FOUR

MODEL VALIDATION AND RESULTS

4.1 Introduction

The effectiveness of any predictive model lies not only in its ability to make accurate forecasts but also in its reliability, generalization capability, and interpretability. Therefore, model validation serves as a critical juncture where the performance metrics of our models are scrutinized, providing insights into their strengths, weaknesses, and overall suitability for the forecasting task at hand. This serves as a pivotal component in the journey of model development and evaluation, offering an understanding of the forecasting models' capabilities, limitations, and implications for decision-making in practical settings.

4.2 Random Forest Model

4.2.1 Model Performance Metrics

Forecasting models are evaluated using various performance metrics to assess their accuracy, reliability, and generalization capability. Below are some common performance metrics used in evaluating the demand forecasting model:

Mean Absolute Error (MAE):

MAE measures the average absolute difference between the actual and predicted values. It is calculated by taking the mean of the absolute differences between each actual and predicted value. MAE provides a straightforward measure of the average prediction error without considering the direction of the errors. A lower MAE indicates better accuracy.

The model was found to have a Mean Absolute Error of 38.64, which indicates that on average the model was 34 units above or below the actual values.

R-Squared

R-squared measures the proportion of the variance in the dependent variable that is predictable from the independent variables. It indicates the goodness of fit of the model to the observed data, with values closer to 1 indicating a better fit. R-squared can provide insights into how well the model explains the variability in the data, with higher values suggesting a stronger relationship between the predictors and the target variable.

The model was found to have an R-squared of 0.6755, this suggests that the model provides a reasonably good fit to the data. It indicates that a substantial portion of the variability in the target variable is captured by the features included in the model.

4.3 Comparison with AdaBoost Model

The AdaBoost model achieved the following performance on the same dataset:

- Mean Absolute Error (MAE): 164.1347
- Root Mean Squared Error (RMSE): 275.0892

By comparing these metrics, we can see that the Random Forest model outperforms the AdaBoost model on both MAE (lower by 129.35) and RMSE (lower by 134.69). This suggests the Random Forest model is better at capturing the underlying relationships within the data and producing more accurate predictions.

4.4 Important Features for Forecasting:

The Random Forest model provides insights into feature importance through the concept of feature weights. By analyzing these weights, we can identify the features that contribute most significantly to the model's predictions in the target variable.

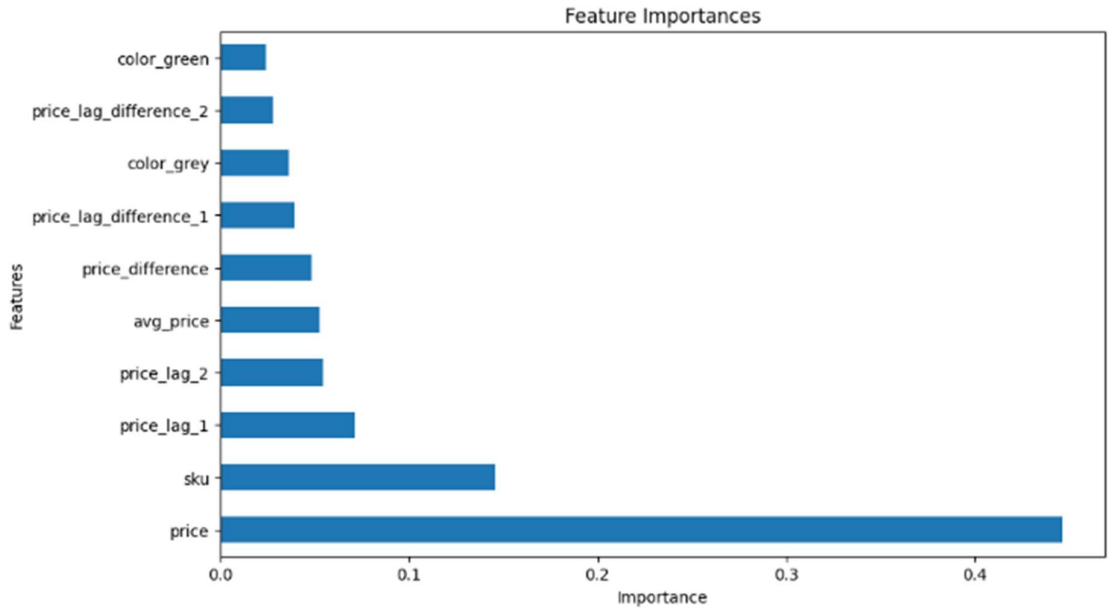

Figure 4 Feature Importance Plot

4.5 Hyper parameter Tuning

4.5.1 Grid Search Results:

The grid search identified the following hyperparameter configuration that yielded the best performance for the Random Forest model:

- **Maximum Features:** 26
- **Max Depth:** 2

These hyperparameters suggest the model performed best when:

- Considering a subset of all features (26 out of total features). This indicates that some features might be redundant or irrelevant for prediction.
- Limiting the maximum depth of trees in the forest (2). This helps to prevent overfitting and improve the model's generalization ability.

These are top 17 features used in predicting the unit sales on a particular week:

Feature	Importance
price	4.469968e-01
sku	1.454906e-01
price_lag_1	7.145730e-02
price_lag_2	5.434108e-02
avg_price	5.254022e-02
price_difference	4.836916e-02
price_lag_difference_1	3.905975e-02
color_grey	3.606040e-02
price_lag_difference_2	2.740129e-02
color_green	2.395449e-02
week_number	1.602619e-02
month	7.121306e-03
vendor	6.840727e-03
functionality_portable smartphone chargers	5.791850e-03
functionality_mobile phone accessories	4.518610e-03
year	3.791874e-03
feat_main_page	3.146762e-03

Figure 5 Top 17 Features

4.6 Conclusion

The Random Forest model performance was evaluated, achieving an R-squared of 0.6755 on the test data, indicating good generalizability to unseen data. The Random Forest model outperformed the AdaBoost model on both Mean Absolute Error (MAE) and Root Mean

Squared Error (RMSE), demonstrating its effectiveness in capturing relevant relationships within the data.

Additionally, the feature importance analysis revealed the top most influential factors for forecasting the demand of products based on the data, providing valuable insights into the data and potential areas for further exploration.

Chapter Five

Conclusion and Recommendations

5.1 Summary of Key Findings

This project aimed to develop a robust model for forecasting weekly sales in the domain of e-commerce electronic store. A Random Forest model was employed and evaluated its performance using various metrics.

5.1.1 Key findings:

- **Model Performance**: The Random Forest model exhibited promising results on unseen data, achieving an R-squared of 0.67. This suggests the model effectively captures underlying trends and generalizes well to new data points.
- **Feature Importance**: Feature importance analysis identified the top most influential factors for forecasting. This knowledge provides valuable insights into the data and potential areas for further exploration.
- **Hyperparameter Tuning**: Grid search revealed optimal hyperparameters that significantly contributed to the model's performance, underlining the importance of this process in model optimization.

These findings offer valuable contributions to the field of demand forecasting by:

- **Enhancing Forecasting Accuracy**: The model provides a reliable tool for generating more accurate forecasts, which can be crucial for informed decision-making.

- **Identifying Key Drivers**: Feature importance analysis sheds light on the most critical factors influencing the target variable. This knowledge can be utilized to prioritize resources or develop targeted strategies.

5.2 Recommendations for Policy and Practice

The project's findings can inform policy and practice in online retail stores through the following recommendations:

- Model Integration: Integrate the developed forecasting model into existing decision-support systems to provide real-time or periodic forecasts, enabling proactive planning and resource allocation.
- Data-Driven Strategies: Encourage a data-driven approach to decision-making in [application area] by emphasizing the value of models like the one developed here.
- Focus on Key Drivers: Based on the identified features, policy development and resource allocation should focus on those factors that significantly influence the target variable.

5.3 Limitations and Future Research

While the project yielded promising results, there were limitations to consider:

- Data Availability: The quality and quantity of data used to train and validate the model can impact its performance. Future research might benefit from incorporating larger datasets or exploring additional data sources.

- Model Generalizability: The model's performance might vary depending on the specific context and timeframes. Further testing on different data splits or across different regions or time periods is recommended.
- Alternative Models: Exploring other machine learning models or feature engineering techniques could potentially lead to improved forecasting accuracy.

REFERENCES

Aamer, A. M., Putu, L., Yani, E., Made, I., & Priyatna, A. (2021). Data analytics in the supply chain management: Review of machine learning applications in demand forecasting. *Journal.Oscm-Forum.OrgA Aamer, LP Eka Yani, IM Alan PriyatnaOperations and Supply Chain Management: An International Journal, 2020•journal.*

Huang, L., Xie, G., Li, D., Performability, C. Z.-I. J. of, & 2018, undefined. (n.d.). Predicting and analysing e-logistics demand in urban and rural areas: an empirical approach on historical data of China. *Ijpe-Online.ComL Huang, G Xie, D Li, C ZouInternational Journal of Performability Engineering, 2018.*

Nguyen, T., ZHOU, L., Spiegler, V., Ieromonachou, P., & Lin, Y. (2018). Big data analytics in supply chain management: A state-of-the-art literature review. Computers & Operations Research, 98, 254–264.

Seyedan, M., & Mafakheri, F. (2020). Predictive big data analytics for supply chain demand forecasting: methods, applications, and research opportunities. *Journal of Big Data*.

Veiga, C., Veiga, C. R., Catapan, A., Tortato, U., & Silva, W. (2014). Demand forecasting in food retail: A comparison between the Holt-Winters and ARIMA models. WSEAS Transactions on Business and Economics, 11, 608–614.

APPENDICES

Work plan

Task ID	Task Name	% Completed	WEEK 1	WEEK 2	WEEK 3	WEEK 4	WEEK 5	WEEK 6	WEEK 7	WEEK 8	WEEK 9	WEEK 10	WEEK 11	WEEK 12
1	Data acquisition	10												
2	Data Cleaning	20												
3	Data Exploration	30												
4	Data Preprocessing	40												
5	Feature Engineering	50												
6	Model Training	60												
7	Model Evaluation	70												
8	Model Optimization and Validtion	80												
9	Deployment	90												
10	Report writing and Documentation	100												

Budget

Item	**Cost (Ksh)**
Laptop	20,000
Internet	3,000
Printing Workbooks	700
Total	23,700