A Work Project, presented as part of the requirements for the Award of a Master's degree in Management from the Nova School of Business and Economics.

DATA-DRIVEN INVETORY MANAGEMENT IN FASHION RETAIL: A MACHINE LEARNING APPROACH TO DEMAND FORECASTING

TOMMASO GUASTI

Work project carried out under the supervision of:

Rongjiao Ji

11/06/2024

Abstract

This study develops a machine learning model to enhance inventory management at the “Kilt” clothing store by accurately predicting annual product sales, preventing overstock, and maximizing profitability. Analyzing historical sales data from 2009 to 2023, I evaluated multiple regression models, identifying the Random Forest as the most effective. The model forecasts 2024 profits to reach 2.2 million, significantly higher than 2023’s 1.4 million. By leveraging past sales data and advanced predictive modeling, the study provides strategic insights to optimize inventory decisions and improve overall store profitability.

Keywords

Sales Forecasting, Machine Learning, Inventory Management, Profit Optimization, Random Forest Model, Retail Analytics.

This work used infrastructure and resources funded by Fundação para a Ciência e a Tecnologia (UID/ECO/00124/2013, UID/ECO/00124/2019 and Social Sciences DataLab, Project 22209), POR Lisboa (LISBOA-01-0145-FEDER-007722 and Social Sciences DataLab, Project 22209) and POR Norte (Social Sciences DataLab, Project 22209).

1 Introduction

To ensure sustainability and profitability in the fast-paced world of fashion retail, effective inventory management is crucial. The sector requires sophisticated stock-level management solutions because of its frequently changing trends and variable consumer preferences. Demand forecasting can be effectively addressed by data-driven decision-making, particularly when machine learning is employed. This paper uses machine learning algorithms to enhance demand forecasting. It focuses on transactional data from the Rome-based clothing store “Kilt”, using sales data and specific product attributes to guide the analysis. Conventional inventory management in small to medium sized fashion retail sometimes ignores the multitude of elements impacting fashion purchase trends in favor of oversimplified statistical methods or intuitive projections based on historical sales.

The research holds relevance as it can close this gap by creating a model that can predict the quantity of products sold, for each type of product, with accuracy, over a one-year period, thereby significantly increasing inventory efficiency. This capability addresses a critical point for small stores like “Kilt”, where the owner must place yearly orders at the beginning of the year and always over-orders due to fears of stockouts. Better forecasting models should improve financial performance and decrease overstock. Key questions guide the research:

RQ1. What is the best model for predicting annual product sales at the "Kilt" clothing store, thereby maximizing profit? The demand for different products is predicted using machine learning models like Random Forest and K-nearest neighbors (KNN). These models offer insights that improve inventory management and sales tactics.

RQ2. What types and quantities of products should the Kilt clothing store purchase for 2024 based on predictive sales modeling? This issue seeks to maximize efficiency and

profitability in inventory procurement over the next year by using the insights from machine learning forecasts to real-world decision making.

2 Literature Review

Besides this work, others also engage in examining the use of machine learning models for sales prediction. An overview of the state of the research on machine learning models for sales prediction is given in this section. Ren et al. (2020) demonstrated that the accuracy of demand forecasting is greatly improved when big data analytics are integrated into retail operations, especially for fashionable products. Ashraf (2022) performed a predictive analysis of retail sales forecasting that effectively anticipated real sales. Raizada & Saini (2021) conducted a comparative analysis of supervised machine learning techniques for sales forecasting, including Random Forest Regression, to estimate sales in different geographical locations. Additionally, Ali et al. (2023) employed random forest models for sales prediction in the retail sector. The effectiveness of the Random Forest and XGBoost algorithms in sales recommendation models was noted by Zhao & Keikhosrokiani (2022). All these studies show how well Random Forest works as a product sales predictor in the retail sector.

While the research discussed in this review demonstrates a variety of effective uses of the Random Forest algorithm and other machine learning techniques to forecast product sales in the retail industry, it is important to note that the specific methodology and context used in these studies are different from those used in the current work. This paper presents a methodology that is specifically adapted to the unique product characteristics and operational needs of the Kilt clothes business. It focuses on personalized prediction models that closely align with the owner's requests. With this unique approach, the analysis and results are guaranteed to be directly applicable to Kilt, allowing

for the optimization of inventory decisions based on a thorough comprehension of the particular sales dynamics and product preferences of the store.

Although the studies that are cited offer insightful information and substantiate the usefulness of machine learning in retail environments, the application in this thesis entails unique data preprocessing techniques and feature selection criteria designed to maximize relevance and accuracy in the particular retail setting that is the subject of the study.

3 Problem Description and Data Manipulation

In the context of retail management, precise inventory forecasting represents a critical challenge that directly impacts business efficiency and customer satisfaction. The shop owner in question faces a recurrent issue: the need to determine the precise quantity and type of products to order annually for his clothing store. This decision must be made at the start of each year, necessitating a highly accurate prediction of yearly sales to prevent both overstocking and stock shortages.

The owner's current approach involves ordering significantly more inventory than estimated sales to ensure that the shop never runs out of stock. This approach results in too many products in the inventory, and an increased likelihood that products that have not been sold will become obsolete, although it is efficient in preventing stockouts. The owner said he would like to keep the current product line, and not add any product types or categories. Thus, the primary goal is to create a machine learning model that predicts which and how many products will be sold within a year. With the use of this model, the owner will be able to make data-driven decisions.

3.1 Data Collection and Description of the Dataset

The information technology division of the clothes store “Kilt” provided the dataset directly. The data was supplied in Excel format. Each row in the dataset represents a distinct transaction, these transactions cover a range of goods sales and purchases, each of which is documented with product details and the amount that was exchanged. There are 510876 transactions in total, organized into 36 different features in the dataset. The dataset’s features are outlined in the table in the Appendix (see Table 1).

3.2 Data Cleaning and Data Transformation

According to Liu et al. (2004), effective data cleaning is essential for reliable process monitoring and performance analysis, as it helps in maintaining the integrity of the data by preserving true data structures and eliminating noise caused by outliers. The dataset's relevance and usability for predicting consumer purchasing behaviors were improved by removing unnecessary or improperly used columns, filtering to include only sales transactions, consolidating similar features, strategically addressing missing values, and fine-tuning categorical data to reflect meaningful segments. The final cleaned dataset now contains 182697 entries dispersed over 15 features, each representing a specific attribute of the transaction.

For a detailed overview of the data cleaning processes used in this study, please refer to Appendix section 8.2.

3.3 Univariate Analysis

Univariate analysis, which focuses on a single variable, is the most basic type of data analysis, according to Alexander (2024). The main goal is to characterize and condense this variable while spotting trends without looking into causes or connections. Using this method, one variable's data is taken, summarized, and patterns are found within the data. A summary table is provided (Table 2) to help visualize the frequency and significance of categorical characteristics in the dataset. For every feature, the most common category is displayed in this table along with its frequency and percentage in relation to the total dataset.

Table 2: Univariate Analysis of Categorical Features

Feature	Number of categories	Most common category	Number of categories making up to 80% of entries
Name_supplier	270	"PRL FASHION OF EUROPE", 23% of the total	36
Collection_type	46	"A14" (autumn/winter 14), 7% of the total	26
Model_ID	7946	"A75AW452", 1,8% of the total	2097
Variant_ID	15839	"1", 13% of the total	6667
Brand	285	"Polo Ralph Lauren", 24% of the total	36
Product_group_description	25	"Maglieria" (Knitwear), 12% of the total	11
Color_Name	464	"Blue", 22% of the total	25
Fabric_description	296	"Cot." (Cotton), 21% of the total	28
Pattern_description	202	"Tu", 54% of the total	7

“Brand”, “Color_Name” and “Pattern_description” have a high concentration in their most common categories, with “Polo Ralph Lauren”, “Blue” and “Tu”, dominating 24%, 22%, and 54%, of their respective totals. Despite having many categories, “Variant_ID” and “Model_ID” have a relatively balanced distribution, with the top category constituting only 13% and 1.8% of the total, respectively. "Size" and "Size_scale" were excluded from Table 2 as they did not provide any insights. A comprehensive summary table (Table 3) is provided also for every numerical variable to enhance our comprehension of categorization aspects. Important statistical parameters for each feature are included in this table, including the mean, median, standard deviation, minimum, maximum, 25th, and 75th percentiles. Understanding the distribution, central tendency, and variability of our numerical data depends on these statistics.

Table 3: Univariate Analysis of Numerical Features

Feature	Mean	Median	StdDev	Min	Max	25th Percentile	75th Percentile
Purchase_price	80,6	55	79,2	0	1312	38,5	88
Selling_Price	221,05	222,2	218,2	7	3550	108	230
Quantity	1	1	0,07	1	5	1	1
Year	2015,2	2016	4,7	1980	2023	2011	2019

“Purchase_price” and “Selling_price” exhibit high variability, ranging from 0 to 1312 for the former and from 7 to 3550 for the latter. "Quantity" refers mostly to transactions involving a single item (usually one, up to a maximum of five). Regarding the "Year" feature data spanned from 2005 to 2023. Outliers from 1980 were identified and removed. In 2009, sales increased by about 500% over 2008 levels (as shown in Figure 1). The cause for this expansion was the 2009 merger of "Kilt" with another retailer, which greatly increased sales. The dataset was constrained to include

data just from 2009 to 2023, which concentrated on the time after the merger and provided a more consistent and pertinent set of data for the study.

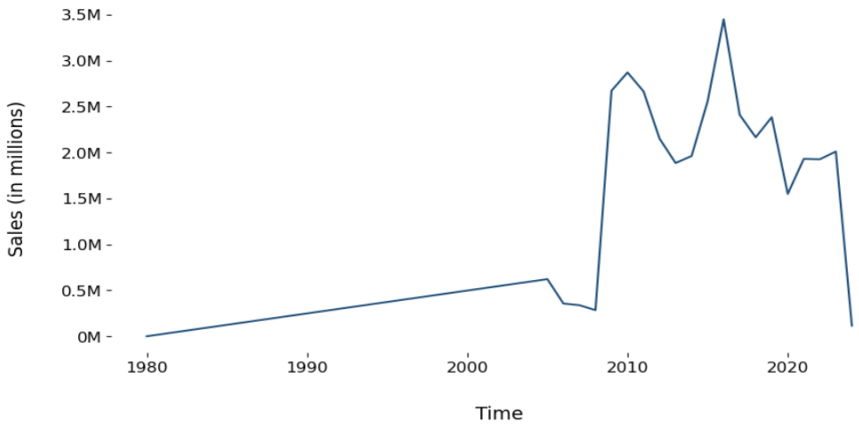

Figure 1: Sales Over Time

3.4 Bivariate Analysis

Bivariate analysis, according to Masud (2023), entails looking at the relationship between two variables. For researchers, this kind of analysis is helpful as it evaluates the degree of correlation and helps establish whether there is one between the variables. The strongest positive correlation (r = 0.98) between "Purchase price" and "Selling price," the high correlation (r = 0,99) between "Selling price" and "Profit," and the strong correlation (r = 0,95) between "Purchase price" and "Profit" are the most significant findings. Given that selling prices are normally determined by purchasing costs and that profit is determined by the difference between buy and selling prices, these correlations are to be expected. Aside from this, no other significant correlations were identified between the variables analyzed, such as quantity sold or the year of transaction.

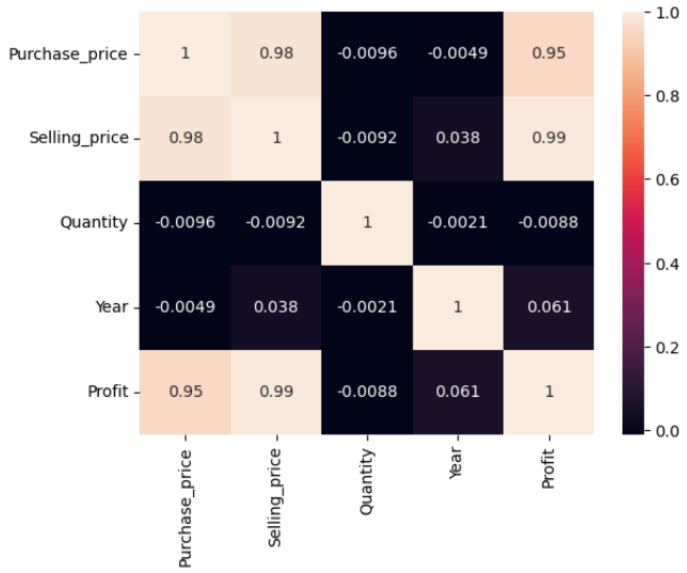

Figure 2: Numerical variables correlation matrix

3.5 Feature Engineering

As Fernández-García et al. (2020) described: Prediction model improvement is significantly impacted by feature engineering. It entails methods like feature deletion, transformation, merging, or splitting which are essential to machine learning. A crucial part of this section is the dimensionality reduction, where the number of categories is reduced within certain features to simplify the model without sacrificing crucial information. Based on a preset threshold, dimensionality reduction was done to some features. With this method, a natural cutoff point in the data distribution is found, beyond which there is a steep fall in the frequency of occurrences. By establishing such a threshold, I can concentrate on the most common categories, which are probably more significant in terms of predicting power and practical significance. For all features, categories falling below their respective thresholds were aggregated into an "Other" category to maintain model simplicity and effectiveness.

“Name_supplier” reduced from 272 categories to the top 43, each with occurrences above 500. “Brand” reduced from 285 categories to the top 130 categories, each with occurrences above 100.

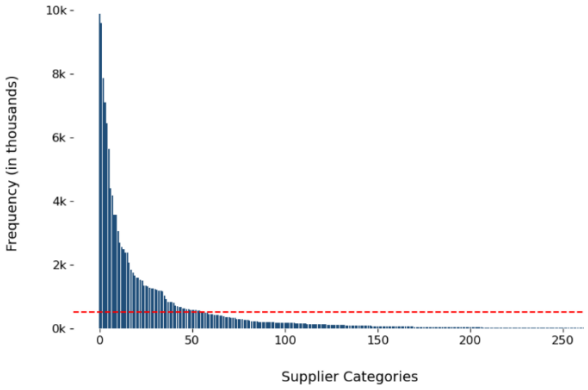

Figure 3: Frequency Distribution of 'Name_supplier' Categories with Threshold Indication

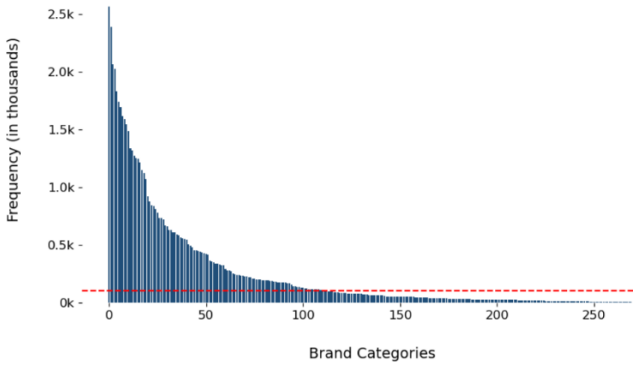

Figure 4: Frequency Distribution of 'Brand' Categories with Threshold Indication

“Product_group_description” reduced from 25 to the top 20, each with occurrences above 1000. “Color_name” reduced from 477 categories to the top 317, each with occurrences above 10.

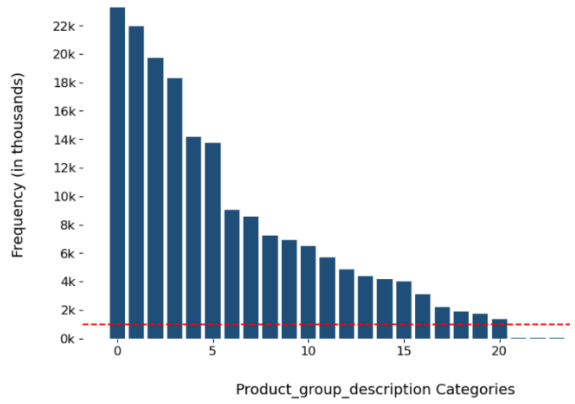

Figure 5: Frequency Distribution of 'Product_group_ description' Categories with Threshold Indication

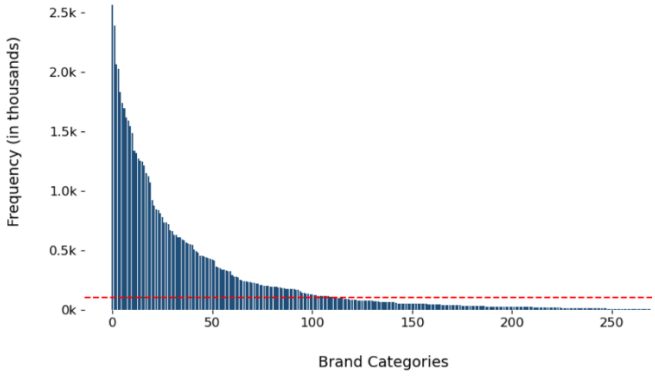

Figure 6: Frequency Distribution of 'Brand' Categories with Threshold Indication

“Fabric_description” reduced from 301 categories to the top 118, each with occurrences above 100.

“Pattern_description” reduced from 209 categories to the top 68, each with occurrences above 100.

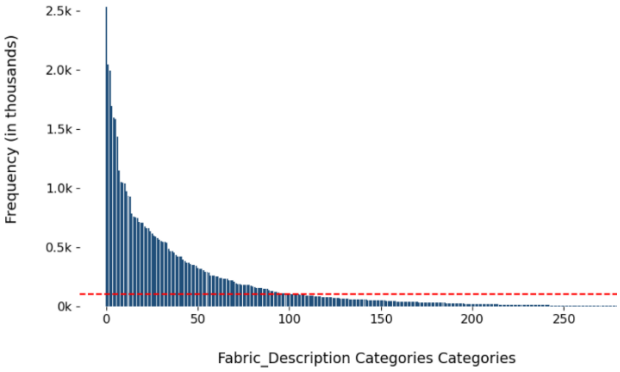

Figure 7: Frequency Distribution of ‘Fabric_ description' Categories with Threshold Indication

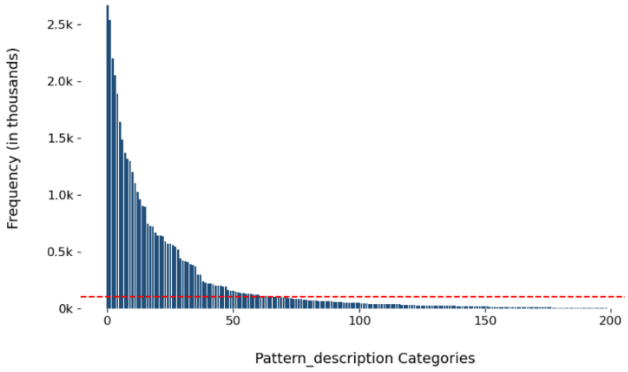

Figure 8: Frequency Distribution of ‘Pattern’ Categories with Threshold Indication

Size_scale reduced from 17 categories to the top 10, each with occurrences above 2500. Size reduced from 79 categories to the top 66, each with occurrences above 10.

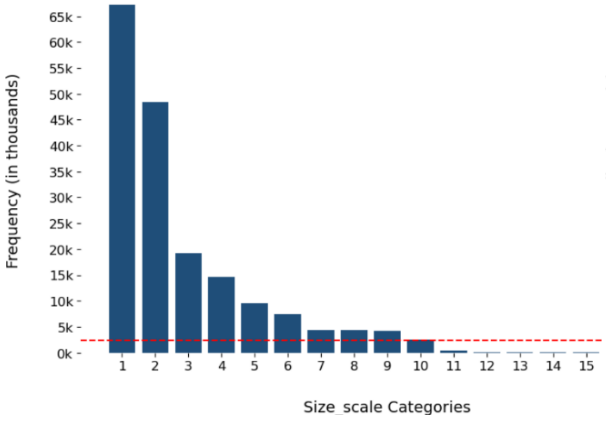

Figure 9: Frequency Distribution of ‘Size_ Scale’ Categories with Threshold Indication

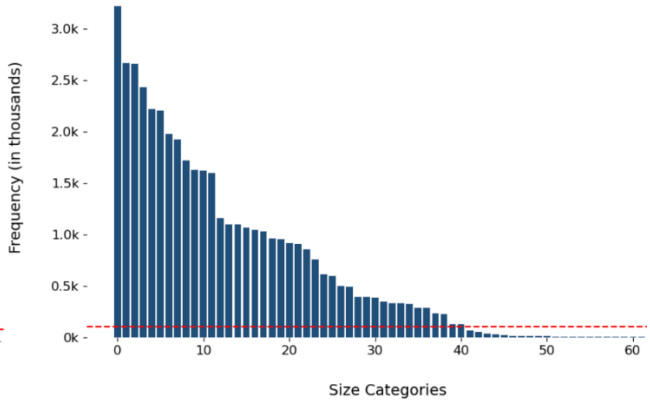

Figure 10: Frequency Distribution of ‘Size’ Categories with Threshold Indication

No dimensionality reduction was applied to “Model_ID” and “Variant_ID”, using these variables in the Machine Learning model, which are unique identifiers, can lead to overfitting because these variables can capture specific details and patterns that are not generalizable. The model might memorize these IDs rather than learn the underlying patterns, thus failing to generalize well to new data. In addition, I used label encoding to convert Year into a number format with values ranging from 1 to 15, which represent the years 2009 through 2023, in that order. Target encoding was used for important categorical features like "Brand,", “Size”, "Product_group_description," "Color_name," "Fabric_description," "Pattern_description," and "Name_supplier." This technique involves calculating the mean of the target variable (“Quantity”) for each category and using this mean as a numerical representation for each category. Target encoding ensures that each category contributes proportionately to its impact on the outcome by recording the average effect of each category on the sales amount. This is especially helpful for features with many categories.

4 Methodology

4.1 Feature Selection

As described by Li et al. (2017), Feature selection is a crucial and commonly applied method for reducing dimensions by eliminating unnecessary information from the dataset to achieve an optimal set of features. This phase's goal is to pinpoint the most important characteristics that have a major influence on the target variable—in this example, the number of products sold. Before proceeding with feature selection techniques, a fundamental step involved restructuring the dataset. To provide a comprehensive overview of annual sales performance across various product characteristics, the data was strategically aggregated. This aggregation was executed by grouping the data based on several key attributes: ‘Year’, ‘Product Group Description’, ‘Brand’, ‘Fabric

Description', 'Pattern Description', 'Color Name', 'Name Supplier', 'Size', and 'Collection Type'. The purpose of this aggregation was to transition from examining individual transactions to analyzing the annual sales volume of each product variant, defined by these specific attributes. Additionally, for each group, I calculated the mean values for 'Purchase Price', 'Selling Price', and 'Profit', alongside the sum of quantities sold. Following the data aggregation phase, the next step in our feature selection process involved employing Random Forest to identify the most significant predictors of sales quantity. As described by Menze et al. (2009), a feature selection based on the random forest classifier has been found to provide multivariate feature importance scores which are relatively "cheap" to obtain, and which have been successfully applied to high dimensional data, arising from microarrays, time series, even on spectra. The Random Forest analysis highlighted several key features that significantly impact the sales volume.

Table 4: Features importance

Feature	Importance
Purchase_price	0,154
Year	0,139
Color_name	0,137
Profit	0,128
Selling_price	0,093
Size	0,084
Brand	0,074
Fabric_description	0,056
Product_group_description	0,036
Pattern_description	0,033
Name_supplier	0,031
Collection_type	0,029

The features selected for the final model were: 'Purchase_price', 'Year', 'Color_name', 'Profit', 'Brand', 'Size', and 'Product_group_description'. 'Product_group_description' and 'Brand' were retained despite their lower importance scores. This decision is a strategic one, aimed at aligning

the predictive capabilities of the model with the operational and strategic needs of the business owner. ‘Product_group_description’, which categorizes products into distinct groups like shoes, t-shirts, etc., as well as ‘Brand’, play a crucial role in enhancing the model's practical application. Including these two features ensures that the model provides more than just numerical forecasts about sales volumes; it offers detailed insights into the types of products that are selling. This is particularly valuable for the business owner because it transforms the model from a simple predictive tool into a comprehensive decision-support system. ‘Selling_price’ was also excluded from the model. The inclusion of both ‘Purchase_price’ and ‘Profit’ already provides a complete picture, allowing for the derivation of selling price if needed, thereby simplifying the model without losing critical information.

4.2 Model Selection

In the development of this predictive model, a critical step involved the selection of the most appropriate regression technique to forecast annual sales based on various product characteristics. To this end, a comprehensive evaluation of multiple regression models was conducted, to determine which model best suits our data and forecasting goals.

The dataset was divided into training, validation, and test sets to assess each model's performance across different stages of the sales cycle. The training set, comprising data from years up to and including 2020, was used to train the models. The validation set, which included data from 2021 to 2022, helped in fine-tuning the models and selecting the best parameters. Finally, the test set, consisting of data in 2023, served as an unseen dataset to evaluate the final model's performance, ensuring its effectiveness and generalizability.

Table 5: Performance metrics across all tested models

Model	MSE Train	MSE Validation	MSE Test	MAE Train	MAE Validation	MAE Test	R^2 Train	R^2 Validation	R^2 Test
Linear Regression	14,9	8,13	7,72	1,8	1,56	1,49	0,17	0,08	0,05
Lasso Regression	17,04	8,31	7,72	1,9	1,7	1,63	0,054	0,06	0,05
Ridge Regression	14,96	8,13	7,72	1,8	1,56	1,49	0,17	0,08	0,05
Elastic Net	15,94	7,8	7,23	1,81	1,59	1,5	0,11	0,12	0,11
Kneighbors	**4,98**	**7,06**	**7,3**	**0,97**	**1,3**	**1,3**	**0,72**	**0,2**	**0,1**
SVR	17,7	9,46	8,88	1,5	1,3	1,2	0,01	-0,06	-0,09
Decision Tree	0	9,37	12,06	0	1,4	1,5	0,99	-0,09	-0,48
Random Forest	**0,79**	**6,87**	**7,28**	**0,39**	**1,26**	**1,29**	**0,95**	**0,22**	**0,1**
Gradient Boosting	10,7	7,76	7,75	1,56	1,43	1,36	0,4	0,12	0,04
MLP	11,7	7,79	7,64	1,61	1,47	1,44	0,37	0,12	0,06

The chosen models, Random Forest and KNeighbors, demonstrated superior performance in terms of prediction accuracy, generalizability, and error minimization, making them ideal candidates for further analysis and deployment in predicting sales volumes.

The **Random Forest** model stood out due to its robust performance across all datasets: It consistently showed low Mean Squared Error (MSE) and Mean Absolute Error (MAE) across training, validation, and test sets. This indicates that the model not only predicts with high accuracy but also maintains this across unseen data, which is critical for practical applications. As Hodson (2022) described, the mean absolute error (MAE) is a widely used metric in model evaluation, calculated as the average of the absolute differences between observed values and model predictions. Furthermore, With the highest R-squared (R2) value among all models during the validation phase and respectable scores in training and testing phases, Random Forest explains a significant proportion of the variance in the data. This high R2 value means that the model captures the underlying patterns and dynamics of the dataset effectively. The **KNeighbors** model was selected for its exemplary performance, particularly in the validation and test phases: While it showed a slight decrease in performance from training to testing compared to Random Forest,

KNeighbors maintained lower MSE and MAE values than most other models in the validation and test sets. This performance indicates its ability to generalize well to new data, a desirable attribute in predictive modeling. The R2 values, although lower than those of Random Forest, were still among the highest, demonstrating that KNeighbors could reliably predict sales across different temporal settings without overfitting the training data.

4.3 Random Forest

Random Forest, as mentioned by Lee et al. (2018), is a powerful ensemble learning method that combines the principles of bootstrap aggregating (bagging) and decision trees to enhance the accuracy and robustness of predictions. By growing a set of decision trees on different bootstrap samples and averaging their outputs, Random Forest balances the trade-off between bias and variance effectively, making it one of the most attractive machine learning algorithms.

Having selected the Random Forest model based on its performance across various metrics, the next crucial step was to optimize its parameters to further enhance its predictive accuracy. According to Liao et al. (2022), hyperparameter tuning involves configuring numerous trial models with different hyperparameter settings to find the optimal configuration that maximizes accuracy or minimizes loss. Table 6 lists the best hyperparameters identified through the hyperparameter tuning process. Table 7 demonstrates the model's performance metrics, using the optimized parameters.

Table 6: Random Forest Best Hyperparameters

Model	Number of estimators	Max features	Max depth	Min samples split	Min samples leaf
Best Random Forest	300	sqrt	20	2	1

Table 7: Random Forest Best Hyperparameters Performance

Model	MSE Train	MSE Validation	MSE Test	MAE Train	MAE Validation	MAE Test	R^2 Train	R^2 Validation	R^2 Test
Best Random Forest	1,04	6,08	6,36	0,53	1,22	1,23	0,94	0,32	0,22

After obtaining predictions for the test set, I noticed that the model produced decimal values. Upon reviewing these predictions, I found that 2787 (70,7%) instances were predicted to exceed actual values, while 1553 (29,3%) were predicted to be less than actual values. To refine these predictions and make them more practical, I adjusted them to the nearest lower integer. This adjustment further improved the Mean Absolute Error (MAE) of the model, reducing it from 1.23 to **1.08**.

4.4 K-nearest Neighbors (KNN)

As Zhang (2016) described, the k-Nearest Neighbors (kNN) classifier assigns unlabeled observations to the class of the most similar labeled examples based on their characteristics. An in-depth parameter tuning was conducted to enhance its predictive accuracy and ensure robustness. Table 8 details the optimal hyperparameters determined via the hyperparameter optimization procedure. Table 9 presents the performance indicators of the model when employing these refined parameters.

Table 8: K-nearest Neighbors Best Hyperparameters

Model	Algorithm	Number of Neighbors	Weights	Leaf size
Best K-nearest Neighbors	Auto	11	Distance	18

Table 9: K-nearest Neighbors Best Hyperparameters Performance

Model	MSE Train	MSE Validation	MSE Test	MAE Train	MAE Validation	MAE Test	R^2 Train	R^2 Validation	R^2 Test
Best K-nearest Neighbors	0	7,17	7,3	7,3	1,36	1,33	0,99	0,2	0,11

In analyzing the test set predictions, it was observed that the model output decimal values. A detailed examination revealed that 2702 predictions (68,5%) were higher than the actual values, whereas 1238 predictions (31,5%) were lower. To enhance the applicability of these predictions, I rounded them down to the nearest lower whole number. This methodological tweak significantly bettered the model's Mean Absolute Error (MAE), decreasing it from 1.33 to **1.21**.

5 Results and Discussion

In this section, I will talk about the outcomes of the implemented models, focusing firstly on their performance in predicting 2023 sales. This analysis is crucial for selecting the best model to predict 2024 sales. To assess the precision of each model, the predictions were segmented into various bins based on the difference between the predicted and actual sales values: a bin for exact predictions (difference of 0), bins for minor discrepancies (differences of 1-2 and 3-5), a bin for moderate discrepancies (differences of 6-9), and a bin for the largest discrepancies (differences of 10 or more). This categorization facilitates a comparison of the models, allowing us to understand their effectiveness in capturing sales trends (See Figure 11). Regarding the K-nearest Neighbors model, nearly half (48,4%) of all predictions perfectly matched actual sales, exhibiting no errors. A significant 38,1% of predictions deviated by just 1 or 2 units from actual sales. Errors between 3 to 5 units were less common, affecting 9,1% of predictions. Only 2,5% of the predictions had errors ranging from 6 to 9 units. The smallest group, at 1,3% represented predictions with the largest error

exceeding 10 units, indicating the most significant discrepancies between predicted and actual values. Regarding the Random Forest model, a majority, 50,6% of predictions were correct, indicating no error. A substantial 38,6% had a minimal error of 1-2 units. Errors of 3-5 units were seen in 7,6% of the cases. Errors within the range of 6-9 units occurred in 1,9% of predictions. The least frequent, at only 1,3% were predictions off by more than 10 units, showing significant discrepancies.

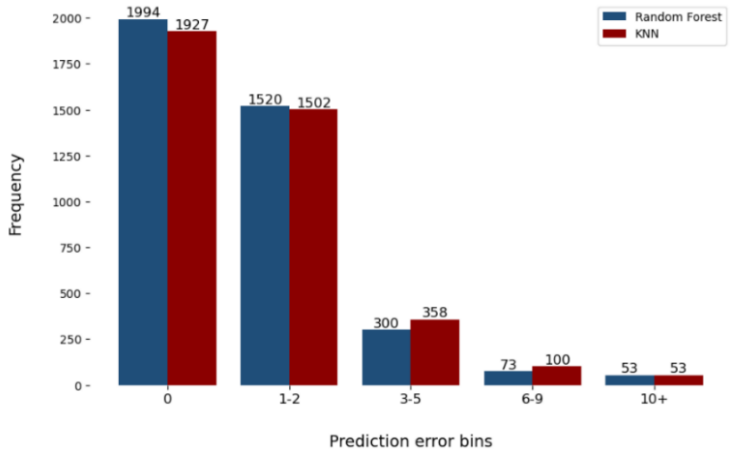

Figure 11: Comparison of Prediction Error Distribution: KNN vs Random Forest

In addition, I conducted a comparison of the predicted profits in 2023 generated by both models (See Figure 12). This comparison was based on the predicted quantity of each product, multiplied by the profit associated with each predicted product. The figure shows the actual profit to be 1.44M, KNN predicted profit as 1.13M, and Random Forest predicted profit as 1.19M. The results are summarized in the bar chart below (See Figure 12). Comparatively, the K-nearest Neighbors model showed less accuracy, resulting in a higher proportion of predictions falling into larger error ranges and less accuracy in predicting profits.

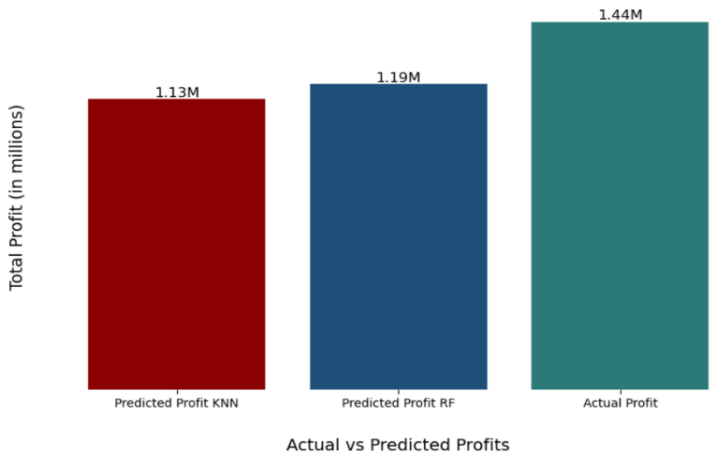

Figure 12: Comparison of Predicted Profits by KNN and Random Forest Models

Therefore, for future predictions, the Random Forest model will be utilized as it has demonstrated slightly superior performance, ensuring reliable and accurate sales forecasting. This conclusion is drawn from addressing **RQ1**, which involved a comparative analysis of various machine learning models. Through testing and evaluation, it was found that the Random Forest model outperformed other models in terms of prediction accuracy. This robustness and higher predictive capability make it the preferred choice for forecasting sales, providing the 'Kilt' clothing store with more reliable sales projections. The Random Forest model was determined to be the most effective for predicting sales at the "Kilt" clothing store, with an MAE of 1.08. In comparison, Raizada and Saini (2021) found that the Extra Tree Regression technique was the best for predicting Walmart sales, followed by the Random Forest Regression technique. Although Extra Tree Regression was not tested in this study, the robust performance of the Random Forest model in both contexts underscores its reliability and accuracy for sales forecasting.

5.1 Random Forest Bias Analysis

The following table categorizes discrepancies between predicted and actual sales data from the Random Forest model using the same bins described earlier. Each group represents a range of

differences between predicted and actual values, providing insights into whether specific features like “Brand” or “Size” might influence prediction errors.

Table 10: Random Forest Bias Analysis

				Most common categories across these features			
Groups	Number of products	Mean purchase price	Mean Profit	Brand	PGD	Color_name	Size
10+ group	53	**58,3**	**111,4**	RAINS	**Access**	Blu	**UNI**
6-9 group	73	**83,9**	**163,8**	Polo	**Access**	Blu	**UNI**
3-5 group	300	**81,2**	**154,0**	Polo	Pants	Blu	50
1-2 group	1520	**86,9**	**167,9**	Polo	Maglieria	Blu	M
0 group	1994	**100,1**	**193,8**	Other	Maglieria	Blu	48

The analysis highlights a potential model bias where predictions tend to be less accurate for “Access” (Accessories) and products with Size “UNI” (Universal). The analysis of the Mean Profit and Mean Purchase Price across these different groups, suggests that the largest discrepancies in prediction accuracy tend to occur with less expensive items. This observation can be gleaned from the trend showing decreasing Mean Profit and Mean Purchase Price as the error in prediction increases.

5.2 Forecasting Product Demand for 2024

To determine a list of products for my model to predict the quantities for 2024, an analytical method was employed to maximize profitability. Historical sales data from 2009 to 2023 revealed that, on average, 3,900 distinct products are sold each year. To ensure the product mix remains consistent with the shop's typical inventory, the percentage of each category within the "Product_group_description" was calculated based on this comprehensive dataset. Using data from

the past three years (2021-2023) for relevance, products within each category were sorted by absolute profit. The most profitable products were then selected until the required count for each category was met, according to the pre-determined percentages. For example, if T-shirts, in the past sales, constituted 4.57% of the product line, approximately 175 T-shirts (4.5% of 3900) with the highest profitability were chosen. This selection method aimed to maximize profitability and improve model accuracy, as previously observed in Table 10, where products with the highest profitability had better prediction accuracy. By maintaining the historical category distribution and selecting the most profitable products within each category, the chosen product line ensures a balance between representativeness and profitability. This curated list of 3900 unique products was then provided to the model to predict the quantities for the upcoming year.

Table 11 provides an example of eight products forecasted to be sold in 2024. These products were selected based on prior analysis, and their predicted sales quantities were determined by the Random Forest model. This conclusion effectively addresses **RQ2**, which examines what types and quantities of products the Kilt clothing store should purchase for 2024 based on predictive sales modeling. The Random Forest model's accurate predictions offer strategic insights into inventory planning, ensuring that the store stocks optimal product types and quantities to meet anticipated demand, thereby maximizing sales and profit and minimizing overstock effectively.

Table 11: Example of eight products forecasted for 2024.

Brand	PGD	Color	Size	Purchase Price	Profit	Predicted Quantity
Suite191	KnitWear	Blue	50	102	202	8
SANTANIELLO	Pants	Blue	50	61,5	138,7	11
XACUS	Shirt	White	40	44	94	10
Polo Ralph Lauren	Hoodie	Black	L	69	99	9
Fabio Toma	Accessory	White	UNI	7,3	31,4	25
Caruso	Suit	Blue	52	457,6	900,7	3
Doucal's	Shoes	Beige	42	133	236	7
XACUS	Shirt	Blue	39	47,3	95,3	19

Based on the previously conducted model bias analysis (see Table 10), there is a noted probability that the forecast in Table 11 for the ‘Fabio Toma’ accessory might be less accurate. As highlighted earlier, products characterized by lower profits, classified under 'Accessory' in product group descriptions, and sized as 'Uni' were more frequently associated with the largest discrepancies between predicted and actual sales figures. This insight is critical for managing expectations and planning strategies to mitigate potential inaccuracies in future inventory decisions.

The model's utility extends beyond general forecasting; it can be tailored to specific requests from the business owner, ensuring its adaptability to varying business needs. For instance, if the owner wishes to understand the projected sales for a product with specific attributes in 2024—such as a "Patagonia blue hoodie, size M, costing $51 with a profit of $98"—the model is capable of providing this precise forecast. In this example, the model predicted that 12 units of this specific product configuration would be sold in 2024, demonstrating the model's capability to deliver detailed, actionable insights based on customized parameters.

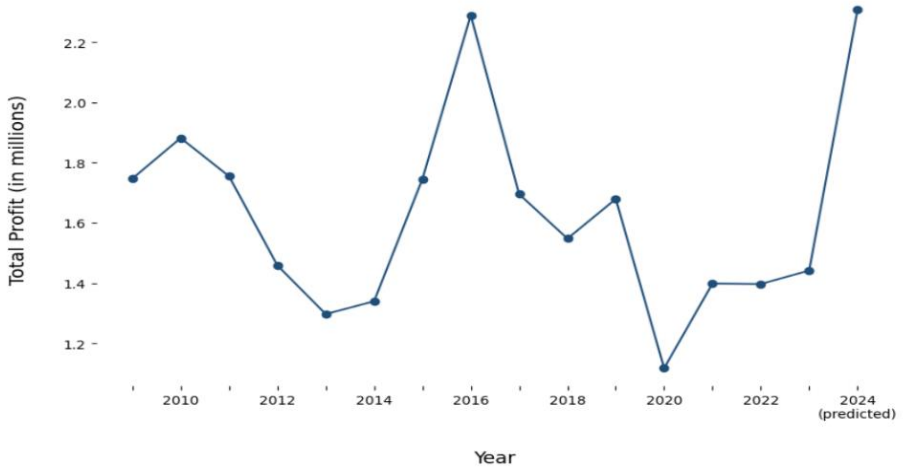

Figure 13: Actual Profits (2009-2023) and Predicted Profit for 2024

After predicting the products for 2024 and calculating the profit based on those predictions, the profit for 2024 is projected to be around 2.2 million, reaching the highest peak alongside 2016. This represents a significant improvement from the 1.4 million profit in 2023. (See Figure 13)

6 Limitations

The model developed for forecasting sales operates within a specific limitation tied to its training data. It is designed to predict sales volumes but not the likelihood of zero sales for any given product. This is because the model was exclusively trained on transactional data where products were sold at least once. This limitation underscores the importance of understanding that while the model provides valuable insights into likely sales volumes, it does not encompass predictions of unsold inventory. This constraint is significant for planning and risk management, suggesting areas for potential model refinement and future research to integrate data reflecting zero-sales instances, thereby enhancing the model's applicability in real-world scenarios. Another limitation of the model emerges from the findings in the model bias analysis (Section 5.1). The model tends to make larger prediction errors when dealing with 'Accessories', products listed as 'UNI' for size, and when products have low purchase prices and profits. This pattern suggests a systematic bias in which the model struggles with accurate predictions for lower-cost items and non-standard product sizes, which often have distinct sales patterns and are less frequently represented in the training data. Such discrepancies underline the need for the model's refinement to enhance its reliability, ensuring uniform prediction quality across the board. Moreover, the model does not account for external factors like market trends and assumes future consumer behavior will follow past patterns. Consequently, significant changes in preferences or behavior could reduce its accuracy.

7 Conclusion

The "Kilt" clothing store faced challenges in accurately forecasting annual product sales, leading to overstocking and stock shortages due to an approach of over-ordering inventory. To solve this, historical sales data from 2009 to 2023 was analyzed. The data was cleaned, transformed, and dimensionality was reduced to focus on significant categories. Multiple regression models, including Random Forest and K-nearest Neighbors (KNN), were evaluated to determine the best predictive model. **RQ1: What is the best model for predicting annual product sales at the "Kilt" clothing store, thereby maximizing profit?** The Random Forest model outperformed others, demonstrating lower Mean Absolute Error (MAE), making it the most effective for predicting sales and providing reliable forecasts. **RQ2: What types and quantities of products should the "Kilt" clothing store purchase for 2024 based on predictive sales modeling?** An analytical method was used to select 3900 distinct products, reflecting the average annual sales. Categories of "Product_group_description" were weighted by their historical percentages to maintain a consistent product mix, and the most profitable products within each category were chosen. The quantities were then forecasted for these selected products for 2024. The 2024 profit is projected to be around 2.2 million, significantly higher than the 1.4 million in 2023, aligning with the peak profit of 2016. This demonstrates the model's effectiveness in enhancing sales forecasting and profitability. In conclusion, the Random Forest model provides accurate sales predictions, optimizing inventory decisions, and maximizing profitability for the "Kilt" clothing store. However, further research could address the identified bias, where the model tends to make larger errors with accessories, non-standard size, and lower profit/purchase price items. Additionally, refining the model to predict zero-sales instances would enhance its applicability in real-world scenarios.

8 References

Ren, Shuyun, Hau-Ling Chan, and Tana Siqin. 2019. “Demand forecasting in retail operations for fashionable products: methods, practices, and real case study.” *Annals of Operation Research/Annals of Operations Research* 291 (1–2) : 761–77. https://doi.org/10.1007/s10479-019-03148-8.

Ashraf, M. Usman. 2022. “A Predictive Analysis of Retail Sales Forecasting using Machine Learning Techniques.” *Lahore Garrison University Research Journal of Computer Science and Information Technology* 6 (04) : 23–33. https://doi.org/10.54692/lgurjcsit.2022.0604399.

Raizada, Stuti, and Jatinderkumar R. Saini. 2021a. “Comparative analysis of supervised machine learning techniques for sales forecasting.” *International Journal of Advanced Computer Science and Applications/International Journal of Advanced Computer Science & Applications* 12 (11). https://doi.org/10.14569/ijacsa.2021.0121112.

Ali, Rao Faizan, Amgad Muneer, Ahmed Almaghthawi, Amal Alghamdi, Suliman Mohamed Fati, and Ebrahim Abdulwasea Abdullah Ghaleb. 2023. “BMSP-ML: big mart sales prediction using different machine learning techniques.” *IAES International Journal of Artificial Intelligence* 12 (2) : 874. https://doi.org/10.11591/ijai.v12.i2.pp874-883.

Zhao, Xian, and Pantea Keikhosrokiani. 2022. “Sales prediction and product recommendation model through user behavior analytics.” *Computers, Materials & Continua/Computers, Materials & Continua (Print)* 70 (2) : 3855–74. https://doi.org/10.32604/cmc.2022.019750.

Liu, Hancong, Sirish Shah, and Wei Jiang. 2004. “On-line outlier detection and data cleaning.” *Computers & Chemical Engineering* 28 (9) : 1635–47. https://doi.org/10.1016/j.compchemeng.2004.01.009.

Alexander. 2024. “Univariate Analysis: Definition, Examples - Statistics How to.” Statistics How To. January 20, 2024. https://www.statisticshowto.com/univariate/.

Masud, Anas Al. 2023. “Bivariate Analysis: What is it, Types + Examples.” QuestionPro. October 17, 2023. https://www.questionpro.com/blog/bivariate-analysis/.

Fernandez-Garcia, Antonio Jesus, Juan Carlos Preciado, Fran Melchor, Roberto Rodriguez-Echeverria, and Jose Maria Conejero. 2021. “A Real-Life machine learning experience for predicting university dropout at different stages using academic data.” *IEEE Access* 9 (January) : 133076–90. https://doi.org/10.1109/access.2021.3115851.

Li, Yun, Tao Li, and Huan Liu. 2017. “Recent advances in feature selection and its applications.” *Knowledge and Information Systems* 53 (3) : 551–77. https://doi.org/10.1007/s10115-017-1059-8.

Menze, Bjoern H, B Michael Kelm, Ralf Masuch, Uwe Himmelreich, Peter Bachert, Wolfgang Petrich, and Fred A Hamprecht. 2009. “A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data.” *BMC Bioinformatics* 10 (1). https://doi.org/10.1186/1471-2105-10-213.

Hodson, Timothy O. 2022. “Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not.” *Geoscientific Model Development* 15 (14) : 5481–87. https://doi.org/10.5194/gmd-15-5481-2022.

Lee, Tae-Hwy, Aman Ullah, and Ran Wang. 2019. “Bootstrap Aggregating and Random Forest.” Dans *Advanced studies in theoretical and applied econometrics*. , 389–429. https://doi.org/10.1007/978-3-030-31150-6_13.

Zhang, Zhongheng. 2016. “Introduction to machine learning: k-nearest neighbors.” *Annals of Translational Medicine* 4 (11) : 218. https://doi.org/10.21037/atm.2016.03.37.

Liao, Lizhi, Heng Li, Weiyi Shang, and Lei Ma. 2022. “An empirical study of the impact of hyperparameter tuning and model optimization on the performance properties of deep neural networks.” *ACM Transactions on Software Engineering and Methodology* 31 (3) : 1–40. https://doi.org/10.1145/3506695.

8 Appendix

Table 1: Overview of Dataset Features

Feature	Description
Transaction_type_ID	Numeric code representing transaction type (e.g., 1 for «Sell», 2 for «Purchase»)
Transaction_type_description	Textual label indicating transaction type (e.g., «Sell» or «Purchase»)
Brand	Name of the clothing brand associated with the item
Date	The specific date on which the product was sold or purchased
Supplier_ID	A unique numerical identifier assigned to each distinct supplier in the dataset.
Name_Supplier	The name of the supplier who sold the item to the clothing shop being analyzed
Collection_type	Indicates the seasonal collection for which the clothing product was released by the brand, distinguishing between «Spring/Summer» and «Autumn/Winter»

Feature	Description
Collection_type2	Indicates the seasonal collection for which the clothing product was released by the brand, distinguishing between «Spring/Summer» and «Autumn/Winter»
Model_ID	A unique identifier assigned by the clothing brand to each distinct clothing model, facilitating differentiation between models
Model_description	Provides a plain language description of the specific clothing model, in understandable terms.
Variant_ID	A unique identifier assigned by the clothing brand to distinguish different variants or versions of the same clothing model.
Not_used	The column is unused or incorrectly utilized by the clothing shop, rendering it irrelevant for analysis.
Brand2	This column represents the brand name of the clothing product, serving the same purpose as the «Brand» column.
Product_group_ID	A numerical unique identifier assigned to categorize the type or group of products being sold such as «shirt», «suit», «shoes», «accessories».
Product_group_description	Provides a descriptive label or name for the product group identified by the «Product_group_ID», indicating the type of category of products being sold (e.g. «shirt», «suit», «shoes», «accessories»).

Feature	Description
Target_demographic_ID	A unique identifier indicating the target demographic for which the product is intended, such as «Men», «Women», or «Kids»
Target_demographic_description	Provides a descriptive label or name for the target demographic identified by the «Target_demographic_ID», indicating whether the product is designed for «Men», «Women», or «Kids».
Color_ID	A numerical code representing the color of the clothing item, facilitating categorization and identification
Color_name	Descriptive name or label for the color of the clothing item, providing clear identification of the color.
Fabric_ID	A unique identifier assigned to categorize the type of fabric used in the clothing item's construction
Fabric_description	Provides a descriptive label or name for the type of fabric identified by the «Fabric_ID», indicating the material used in the clothing item's fabrication.
Pattern_ID	A unique identifier assigned to categorize the type of pattern used in the clothing item, such as «checked», «striped», or «floral»
Pattern_description	Provides a descriptive label or name for the type of pattern identified by the «Pattern_ID», indicating the design or pattern featured on the clothing item.

Feature	Description
Collection_type3	Indicates the seasonal collection for which the clothing product was released by the brand, distinguishing between «Spring/Summer» and «Autumn/Winter»
Collection_type4	Indicates the seasonal collection for which the clothing product was released by the brand, distinguishing between «Spring/Summer» and «Autumn/Winter»
Purchase_price	The price at which the clothing shop purchased the item from the supplier
Not_used2	The column is unused or incorrectly utilized by the clothing shop, rendering it irrelevant for analysis
Selling_price	The price at which the clothing shop sold the item to the client.
Not_used3	The column is unused or incorrectly utilized by the clothing shop, rendering it irrelevant for analysis
Size_ID	A unique identifier assigned to categorize the scale or category for the clothing item. It links to the «Size_scale» column, providing a numerical code to represent different size categories.
Size_scale	Describe the scale of the category of sizes for the clothing item, indicating the range or type of sizes available, such as «S to XL» for clothing or specific size ranges for shoes.

Feature	Description
Size	Represents the actual size of the clothing item within the scale or category described in «Size_scale» column.
Quantity	Represents the number of units of the specific product involved in the transaction, whether it’s a purchase or a sale.
Not_used4	The column is unused or incorrectly utilized by the clothing shop, rendering it irrelevant for analysis
Not_used5	The column is unused or incorrectly utilized by the clothing shop, rendering it irrelevant for analysis

8.1 Table of Contents

1 Introduction …… 2

2 Literature Review…… 3

3 Problem Description and Data Manipulation…… 4

3.1 Data Collection and Description of the Dataset…….. 5

3.2 Data Cleaning and Data Transformation……. 5

3.3 Univariate Analysis……. 6

3.4 Bivariate Analysis…… 8

3.5 Feature Engineering……. 9

4 Methodology……..... 12

4.1 Feature Selection…… 12

4.2 Model Selection…….. 13

4.3 Random Forest…… 16

4.4 K-nearest Neighbors (KNN)……... 17

5 Results and Discussion……..... 18

5.1 Random Forest Bias Analysis…… 20

5.2 Forecasting Product Demand for 2024…….. 21

6 Limitations…… 24

7 Conclusions……... 25

8 References…… 26

9 Appendix……... 29

8.2 Data Cleaning and Data Transformation Detailed

The cleaning steps are organized by column, detailing the specific actions taken to address issues such as missing values, outliers, and irrelevant data.

- **Not_used", "Not_used2", "Not_used3", "Not_used4" and "Not_used5"**: These columns were eliminated after it was determined that they included no relevant data for the analysis and had been improperly used.

- **"Transaction_type_ID" and "Transaction_type_description"**: It was imperative to exclude non-sales transactions from the dataset in order to bring it into close alignment with the study's goal of forecasting client purchasing behaviors. Originally, the dataset included a variety of transaction kinds, such as sales to customers and purchases from suppliers. I filtered the dataset to include only transactions marked as "Vendita" (Sales). Considering the store's inventory policies, this emphasis on sales transactions is justified. As I mentioned, the store usually stocks quantities well beyond anticipated demand, so that items are rarely, if ever, out of stock. This approach means that sales data reflect real consumer demand, rather than being limited by inventory shortages. Furthermore, the owner indicated that the purchase data was not reliable, as it failed to account for numerous transactions. Post-refinement, it was determined that the columns "Transaction_type_ID" and "Transaction_type_description" were unnecessary because every data entry related solely to sales. The dataset was made simpler by removing these columns.

- **"Brand" and "Brand2":** The "Brand" column was removed, "Brand2" was renamed to "Brand" because there were more missing values in "Brand2" and fewer missing values in "Brand".

- **“Collection_type”, “Collection_type2”, “Collection_type3” and “Collection_type4”**: “Collection_type2”, “Collection_type3” and “Collection_type4” contained incomplete or irrelevant data. They were removed.

- **“Model_ID” and “Model_description”**: “Model_ID” is a code that the Brand assigns to every single clothing model, representing the official name given to a model. On the other hand, “Model_description” is a column filled in by the shop owner, containing a subjective interpretation or explanation of the model. Although helpful, these descriptors are not uniform and change depending on the owner’s viewpoint or communication style, as a consequence the decision was made to eliminate “Model_description” due also to the substantial number of missing values in it.

- **Color_ID and Fabric_ID**: they were originally denoted in the dataset by randomly issued numerical values that fell between -1000 and +1000 respectively, to indicate distinct colors and textiles. Despite being unique, these identifiers provided no intuitive understanding of the color or fabric’s true properties because the numbers did not convey any descriptive information. In order to improve the dataset’s readability and use, it was decided to delete “Color_ID” and “Fabric_ID” and retain “Color_name” and “Fabric_description”.

- **“Target_demographic_ID” and “Target_demographic description”**: Over 99% of the data in the dataset related to men’s products, which was shown to be the overwhelming majority of entries. Kilt, the shop in question, runs distinct retail outlets for women’s and children’s clothes, which are not included in this dataset. In order to preserve a targeted and pertinent dataset, any leftover rows related to women or kid clothes were removed from the analysis, as a result, these 2 columns have been deleted as they contain only one category.

- **“Brand”, “Purchase_price” and “Size_scale”**: Due to the minimal presence of missing data within these columns, a decision was made to enhance dataset integrity by excluding rows that contained any missing values. This selective removal was predicated on the assumption that the minimal loss of data would not significantly impact the overall analytical outcomes, while simultaneously ensuring a higher quality and consistency across these critical variables.
- **“Color_name”, “Fabric_description” and “Pattern_description”**: These columns exhibited a significant proportion of missing data. To address this without discarding substantial portions of the dataset, missing entries were uniformly assigned the label "Not specified". This approach allowed for the retention of valuable data points, facilitating broader analytical coverage while acknowledging the incompleteness of some records in these attributes.
- **“Date”**: Converted and renamed the "Date" feature into a "Year" feature, due to the strategic focus on yearly sales forecasting. Given that the shop owner's primary concern is understanding and predicting annual product demand to inform yearly ordering processes, detailed granularity such as day or month is superfluous and could introduce unnecessary complexity into the model.

The final cleaned dataset now contains 182,697 entries spread across 15 features, each representing a specific attribute relevant to the study. This structure is a result of meticulous efforts to enhance data quality and integrity, ensuring that each entry contributes meaningfully to the insights and predictions generated by subsequent models.

After completing the data cleaning process, data type transformations were implemented to optimize the dataset for analysis. Here's a concise overview of the changes:

- Category: Name_supplier, Collection_type, Model_ID, Variant_ID, Brand, Product_group_description, Color_name, Fabric_description, Pattern_description, Size_scale, Quantity
- Integer: Year, Quantity
- Float: Purchase_price, Selling_price