Machine Learning Coursework
Student ID:
Dataset: UCI Online Retail II
2024

Table of Contents

1. Introduction

This paper explores the integration of machine learning (ML) techniques with retail operations, underpinned by the analysis of Point of Sale (POS) data. The central premise posits that retail establishments can capitalize on data intelligence by harnessing transactional data to forge data-driven strategies that elevate operational efficacy, and in turn, amplify sales and profitability. (Di Sia P.,2022)

1.2 Research Objectives

The stakes of navigating the intricacies of contemporary retail are higher than ever, with the digital economy rewriting the rules of engagement between businesses and consumers.

The evolution of marketing from a product-centric to a customer-centric approach underscores the importance of understanding consumer behaviour in the current data rich landscape (Cumby et al., 2004). An emphasis on customer relationship management (CRM) has become a pivotal aspect of competitive differentiation.

Clustering*:* To observe customer segments and enhance CRM strategies. This unsupervised approach will supplement and refine marketing insights by identifying non-predefined clusters.

Classification*:* To construct a predictive model based on clusters. This model will classify new customers into identified segments to guide specific business strategies.

Accurate forecasting is critical for effective business planning and decision-making (Hsieh, Giloni and Hurvich, 2020). Companies aim to balance inventory levels to meet customer demand efficiently, avoiding unnecessary costs in procurement and storage. This research intends to explore the complex, often nonlinear relationships between sales data, where traditional mathematical models fall short (Hwang et al., 2023).

Regression: To this end, regression will be employed to discern and forecast patterns in sales data, both linear and nonlinear, to inform and enhance business strategies.

2. Understanding the Dataset: UCI Online Retail II

The UCI Online Retail II dataset was selected for its comprehensive representation of basic variables that are commonly recorded in retail operations. Its simplicity makes it accessible for a wide range of analytical techniques, while it spans over 2 years providing a substantial temporal framework for understanding customer behavioral trends. Although the data originates from a wholesale focused retailer, the core data characteristics allow for model adaptability across different contexts, showcasing the potential for widespread industry application.

The steps that will be taken to drive business insights consists of 6 pivotal stages:

1. Data Understanding, to comprehend the variables and their interrelations.

2. Data Preprocessing, to refine the dataset for analysis.

3. Modeling Phase, where algorithms will learn from the data.

4. Evaluation Phase, to assess model performance.

5. Go Live, where models are deployed in a real-world environment

6. Business Insight, the empirical application where analytical results are translated into strategic actions.

This framework, inspired by the principles delineated by Nguyen et al., is designed to transform raw data into strategic knowledge, fostering an empirical approach to business enhancement.

2.1 Exploratory Data Analysis

Develop a **df_info** function to succinctly present key data frame attributes. Post-cleanup, the data frame dimensions are **(823,364 rows, 9 columns)**. It's imperative to exclude records with missing values to ensure data integrity. The dataset contains multivariate, sequential, and time series components—each enriching the analysis with categorical, quantitative, identifier, and time-based information, respectively.

#	Column	Non-Null Count	Dtype
0	Invoice	802995 non-null	object
1	StockCode	802995 non-null	object
2	Description	802995 non-null	object
3	Quantity	802995 non-null	int64
4	InvoiceDate	802995 non-null	datetime64[ns]
5	Price	802995 non-null	float64
6	Customer ID	802995 non-null	int64
7	Country	802995 non-null	object
8	Year	802995 non-null	object

Figure 1: Table showing data types in our data set

2.2 Feature Engineering

The Recency, Frequency, and Monetary (RFM) model is an established market analysis tool that segments customer by purchasing habits. The segmentation informs strategies to enhance the value each customer brings to the business (Christy et al., 2018)

	Customer ID	Recency	Frequency	Monetary
0	12346	327	34	77556.46
1	12347	3	253	5633.32
2	12348	76	46	1658.40
3	12349	20	172	3678.69
4	12350	311	16	294.40
...	...	...	...	...

Figure 2: Table showing the calculated RFM values after grouping by customers

Recency: measures the time since a customer's last purchase, with longer periods indicating reduced engagement.

Frequency: measures the total number of transactions a customer has made over a two-year period, with a higher count suggesting greater loyalty.

Monetary: totals the amount spent by a customer over 2 years gauging their revenue contribution.

Aggregating the data by "Customer ID" facilitates the application of the RFM model to individual customers, yielding data-driven insights conducive to business optimization strategies.

2.3 Data pre-processing

This study ensures representativeness by encompassing the entirety of customer transactions and employs scaling to normalize feature ranges, thus preserving model integrity.

Scaling and representativeness: Scaling and attaining clear distribution characteristics are fundamental pre-requisites for application of machine learning and statistical analysis. Logarithmic transformation was preferred for its suitability in handling the dataset's wide value range, preventing data distortion through inappropriate value compression.

This transformation approach prevents the disproportionate compression of data ranges, which could inadvertently assign zero values to non-zero figures, distorting the data's true range.

By calculating the z-scores of the logged values, a set of standardised scores indicate how many standard deviations the logged RFM are away from the mean. Setting a threshold of three standard deviations facilitates the exclusion of data points that significantly diverge from the mean, thus effectively purging extreme outliers from the dataset.

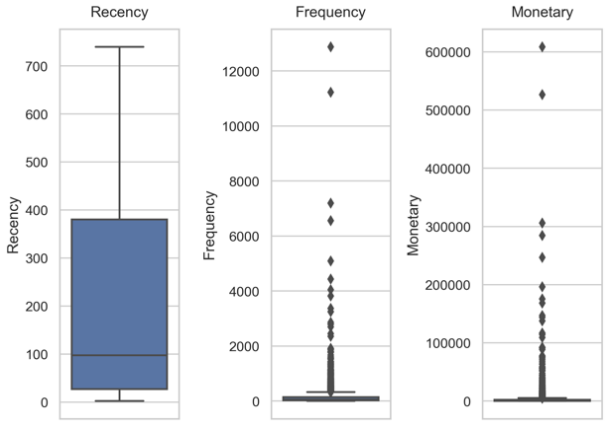


Figure 3: After Robust Scaling the data range is still very large

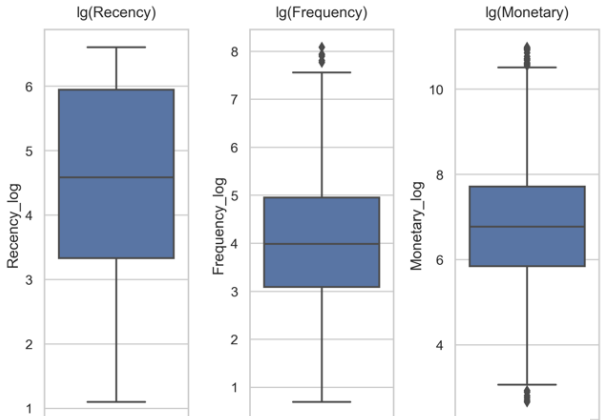


Figure 4: After removing the outliers from the logged scores

3. Clustering – Customer Segmentation

Companies often employ customer analysis to observe the characteristics in data sets to achieve strategic goals (Varad R Thalkar, 2021). Clustering groups data points based on their similar characteristics, ensuring that points in the same group are alike and distinct from those in other groups. Thus, it is is ideal for customer segmentation as it identifies groups within a large dataset using customer attributes, even without pre-existing labels (Christy et al., 2018).

3.1 K-means clustering algorithm

The K-means clustering algorithm calculates the (Euclidean) distance between data points to create K segments. It is preferred for customer segmentation due to its simplicity, efficiency with large datasets, non-overlapping clusters, and scalability. K-means was chosen over other algorithms as its results are easy to interpret. It is beneficial for marketing strategies as it allows for clear delineation of customer segments based on centroid positions within the feature space (Razia Sulthana A et al., 2023)

Additionally, K-Means lends itself to a straightforward analysis and actionability, where marketing teams can readily apply strategies to the defined segments. This ease of application and the actionable nature of the segments formed are often more valuable in a business context than the complex models that require extensive computational resources and yield less interpretable results.

3.2 Enhancing model performance

The Elbow Method observes the decline in the WCSS as a function of K. The optimal number of clusters is then inferred at the point where the rate of decrease attenuates (Ketchen and Shook, 1996). From the plot the elbow appears at K=2 and K=4.

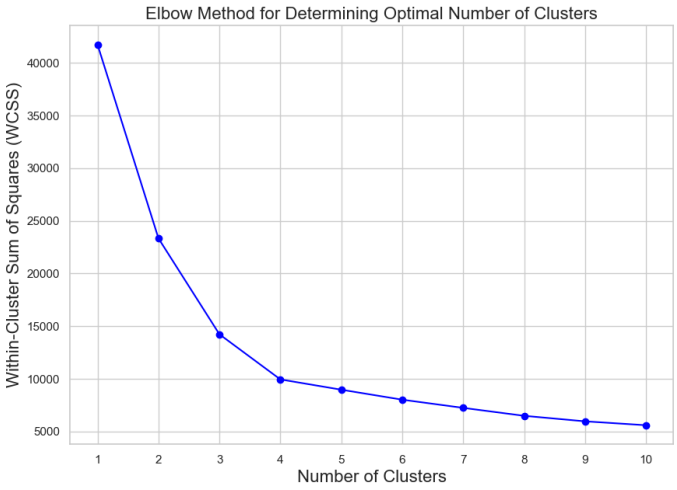


Figure 5: WCSS against K clusters

The silhouette score measures how similar an object is to its own cluster (cohesion) compared to other clusters (separation). A high silhouette score indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.

From the plots, we determined that the optimal number of clusters is 2 or 4, a segmentation of 2 clusters may oversimplify the underlying customer dynamics and miss out on the valuable insights that can be captured with four distinct segments.

Improving Interpretability: The right number of clusters can make the model results more interpretable and meaningful for downstream analysis or decision-making processes.

Enhancing Model Performance: With the optimal set of hyperparameters, the K-means algorithm can perform more efficiently, reducing computational costs and time, especially on large datasets. (Bhade et al., 2018)

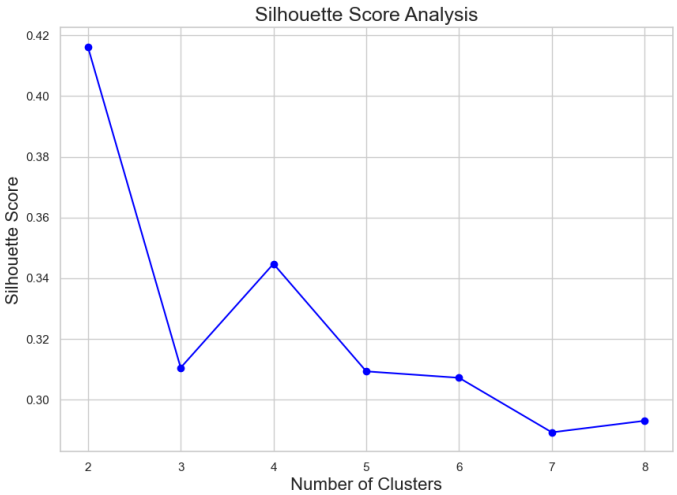


Figure 6: Silhouette score graph

3.3 Analysing the Cluster results

Constructing a pair plot and a 3D plot reveals the delineation of four unique clusters. Analysis of the RFM scores permits the valuation of customer segments, ranking them from most to least valuable as Cluster 2, 1, 0, and 3, respectively. Examination of the pair plot illustrates that Cluster 2 outperforms in all three metrics. Meanwhile, Cluster 1, despite its higher recency indicating less recent interactions, demonstrates comparable frequency and monetary values to the more established customer base in Cluster 0.

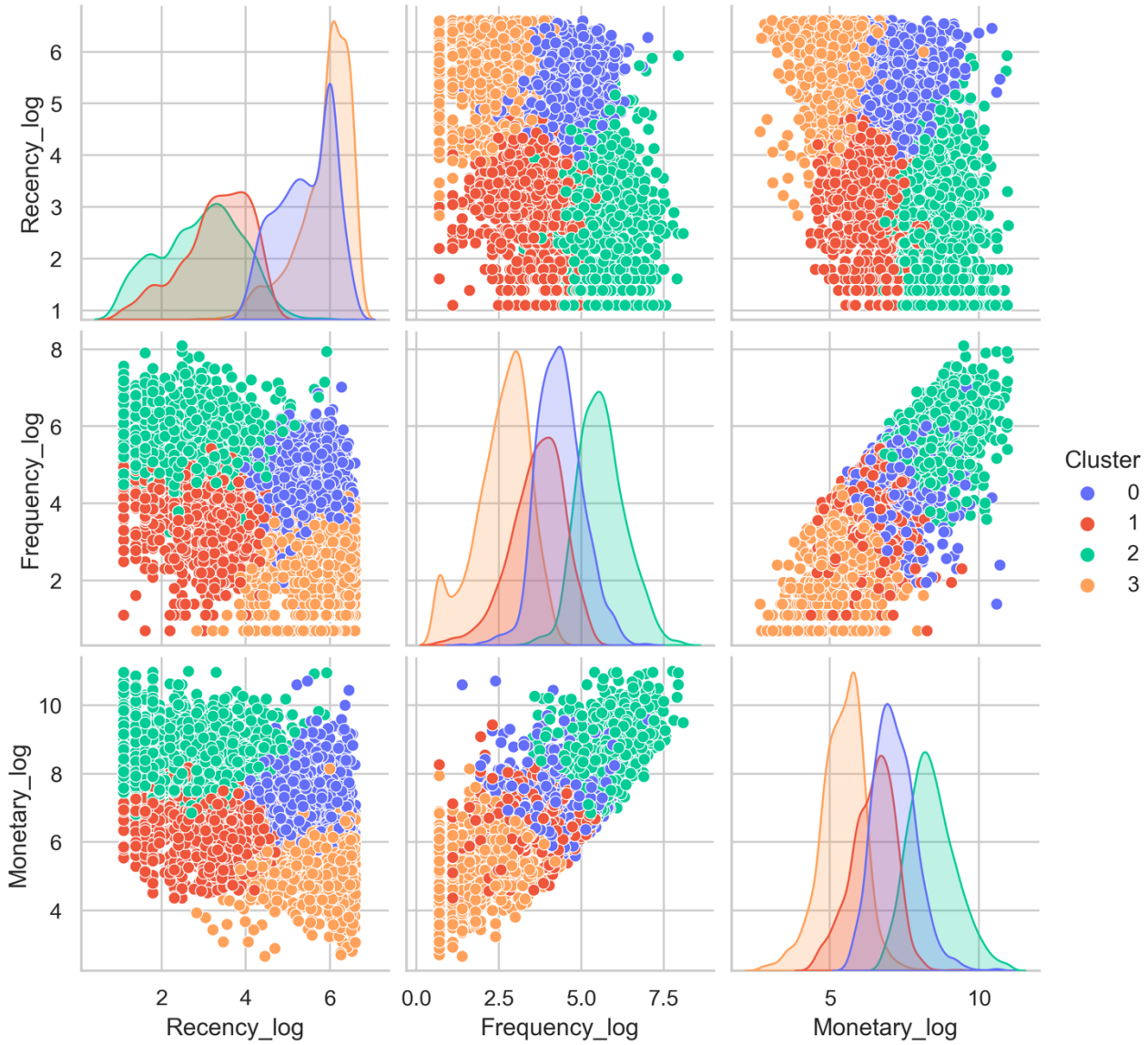


Figure 6: Pairplot of RFM log values

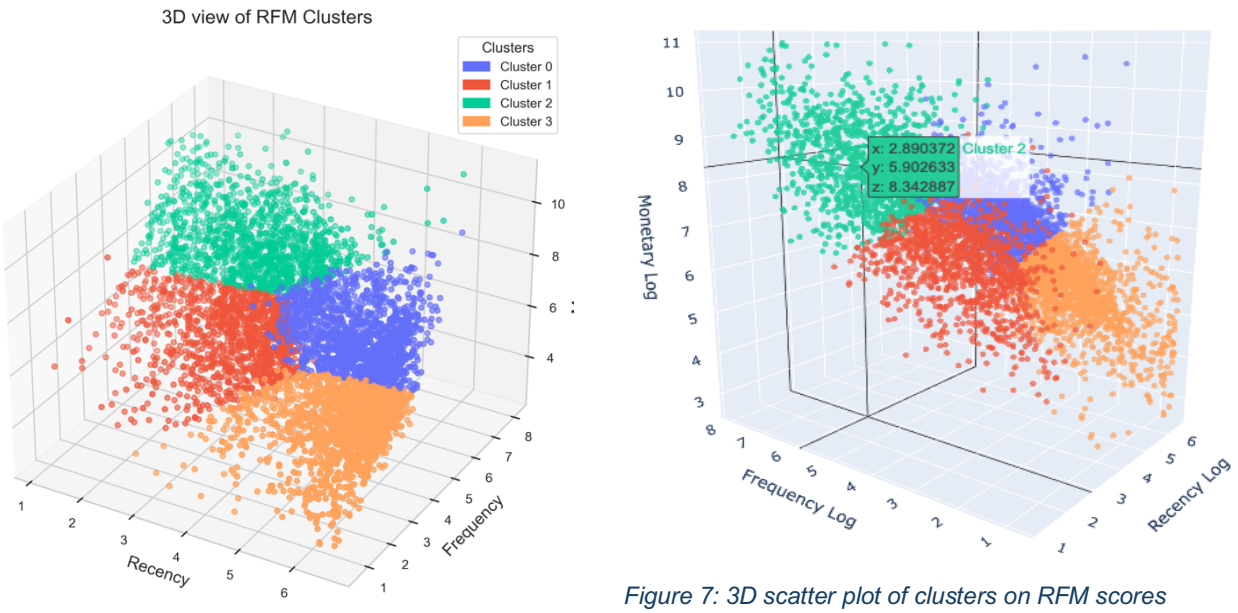


Figure 7: 3D scatter plot of clusters on RFM scores

3.4 Imputing the RFM score onto the cluster

Integrating RFM segments with cluster analysis yields a comprehensive understanding of customer value. This method enhances the granularity with which we appreciate customer behaviors, allowing for more nuanced engagement strategies.

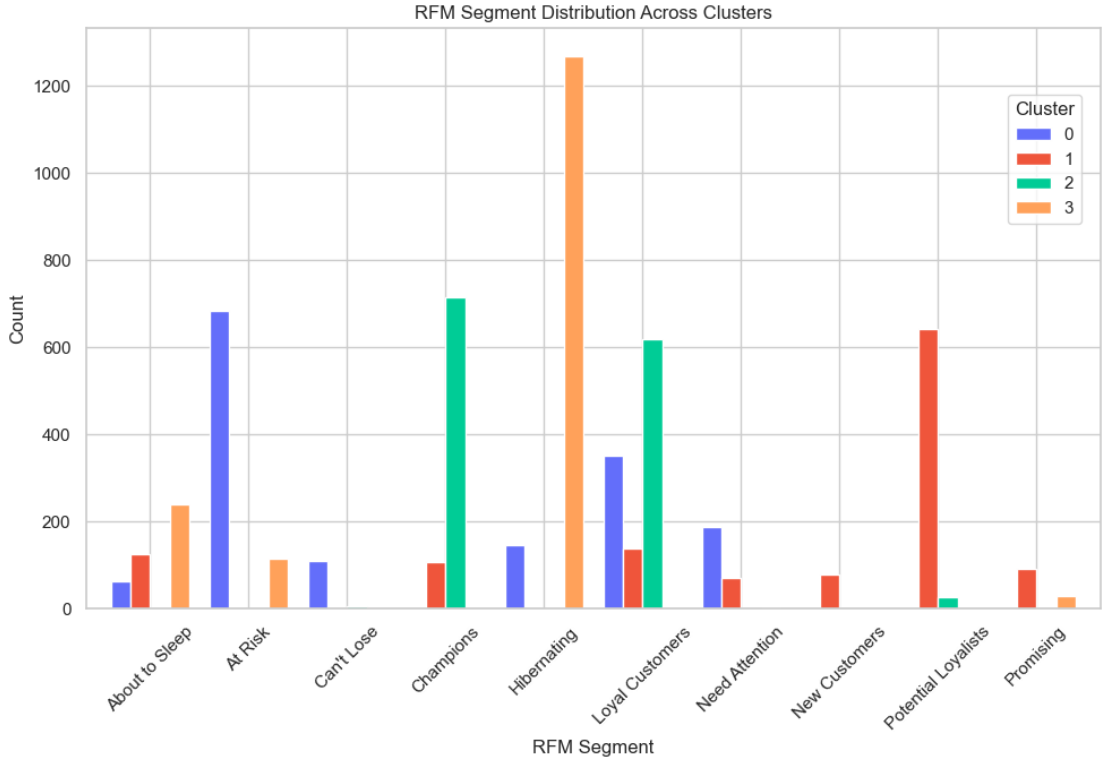


Figure 8: Table to show the distribution of cluster among the segments

3.5 Business implication

The Davies-Bouldin Index, with a value of 0.941, suggests satisfactory cluster definition and separation in a 4-cluster model. This implies moderate delineation between clusters that can now be used to inform business strategies.

Further enhancement of clustering could be pursued through GRFM clustering, contingent upon the availability of categorical data such as product types associated with stock codes. (Chang and Tsai, 2011) This method could refine segmentation by incorporating product categories.

4. Classification

To extend the utility of our insights for new and prospective customers, classification algorithms will be deployed to ascertain the probable cluster for each newcomer. The model will rely on a robust set of features that provide a comprehensive customer profile, allowing for dynamic adjustment of cluster assignments as customer behaviors evolve over time.

4.1 Feature engineering

Additional features: To categorize customers based on purchasing behaviour, a new DataFrame was created from the cleaned data, and dates were converted into year, month, day, and quarter categories for detailed temporal analysis. RFM metrics were computed. Additional metrics, such as average spend per transaction and seasonal buying patterns, were calculated. After deduplicating to retain unique customer records, the dataset was prepared for classification, enabling prediction of customer segments for targeted marketing and strategic engagement.

This process aligns each customer with a segment reflecting their purchasing patterns, which can be invaluable for targeted marketing efforts and personalized customer engagement strategies.

Encoding: Traditional encoding techniques such as dummy encoding and one-hot encoding were initially considered. However, these methods proved impractical as they significantly increased the dimensionality of the dataset, which could lead model complexity and computational inefficiency, commonly known as the "curse of dimensionality."

Consequently, **binary encoding** was employed as an alternative approach. This method was selected for its efficiency in dealing with categorical variables by converting them into binary columns, but with a lower increase in dimensionality compared to other encoding schemes.

4.2 Data pre-processing

Feature scaling in ML is critical to normalize variables and ensure uniformity across different scales. **MinMaxScaler** was employed This scaling technique adjusts the values within a [0, 1] range, facilitating the comparison of feature magnitudes which is crucial for models sensitive to the absolute size of inputs.

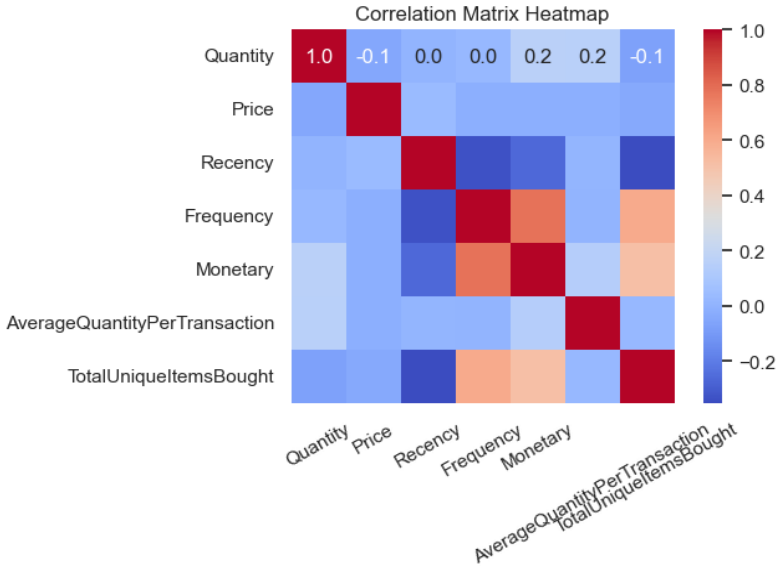


Figure 9: Correlation Matrix of features

Evaluating collinearity among variables is another significant step in the data preparation phase. The **Variance Inflation Factor (VIF)** provides a quantification of the increase in the variance of regression coefficients due to multicollinearity in the model. The dataset presented VIF scores predominantly below the threshold of 5, indicating moderate correlation which typically does not pose a significant threat to the validity of the model.

By applying the MinMax scaling technique, the dataset is thus prepared to meet the demands of the algorithm, laying a robust foundation for the subsequent analytical processes.

Set Seed: To ensure that the result of the model is reproducible and maintain consistency in the experiment. This will ensure that any variation in the performance of the classification is due to the change made in the model rather than randomness in the algorithm's initial condition or data partitioning.

4.3 Model selection

After the data has been treated, the project employs ML classifiers from the scikit-learn library for predictive analysis: Logistic Regression, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Gradient Boosting, and Gaussian Naive Bayes. These classifiers are initialized with their respective default settings, except for the Logistic Regression which is configured for convergence with a higher iteration limit.

A function **train_and_evaluate** is created to orchestrate the training of these models on a designated training dataset and subsequently evaluates their accuracy on a test set. It employs the **.fit()** method for model training, followed by **.predict()** for generating predictions. Performance metrics are then calculated using **accuracy_score** and **classification_report** functions from scikit-learn, providing not only accuracy but also precision, recall, and F1-score for a multifaceted evaluation of each model's predictive capability.

The evaluation of classification models revealed Logistic Regression's moderate accuracy, with underperformance for Cluster 2, and Random Forest's high effectiveness, although overfitting may be a concern. SVM underperformed, favoring Cluster 3, and would require substantial tuning. Neural Networks showed average results with potential for increased precision. Gradient Boosting delivered the highest overall precision and recall, indicating robust performance. Naive Bayes had moderate success with noticeable misclassification for Cluster 2.

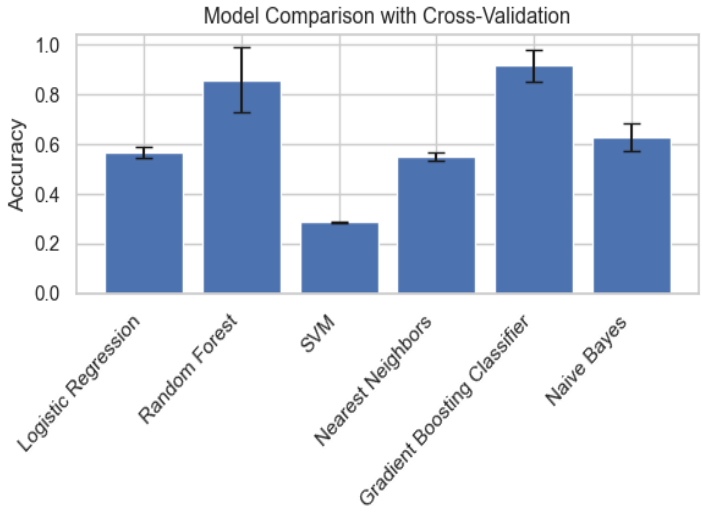


Figure 10: Model's Accuracy Scores and Standard deviations visualised

Gradient Boosting and Random Forest are generally the best for this classification task.

Upon analyzing the results of cross validation, the Gradient Boosting Classifier is the best model for classification as it offers a strong balance of high accuracy and moderate variability.

4.4 Business implication

Notably, Cluster 3 was most readily classified, a possible improvement that can be made is to test with outliers in the model. There is a possibility of excess positive outliers that may have led to Cluster 3 being easier to identify.

The classification model based on gradient boosting can now be integrated into the existing CRM systems to enable better marketing and sales strategies. Furthermore, the effectiveness of the models in live environments can be measures against overall sales growth and retrained as new customer data becomes available.

5. Regression

ML sales forecasting transcends traditional time series analysis by embracing the complexity of sales dynamics through regression techniques. ML techniques are adept at identifying intricate seasonal patterns, including various associated risks, thereby facilitating refined predictions of sales trends. (Kohli, Godwin and Urolagin, 2020) In this section, regression ML models are used to predict future sales.

5.1 Exploratory data analysis

The analysis of time-series data reveals a discernible pattern where, within any given year, sales trends exhibit an upward trajectory with notable peaks typically occurring mid-year.

Further examination uncovered a discrepancy between sales revenue and the quantity of products sold over the years. This inconsistency suggests a potential shift in pricing strategies or changes in the types of items purchased, such as an increase in the sale of higher-quantity but lower-priced items.

Across different clusters, it was observed that these patterns closely mirrored the general sales trends, presenting no significant deviations.

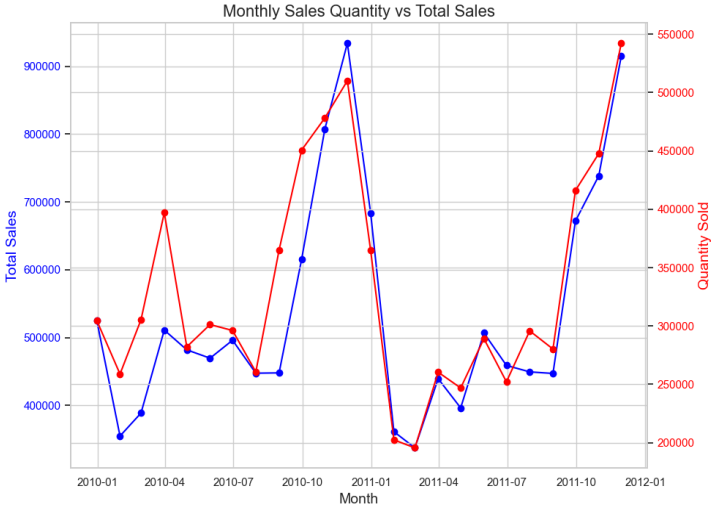


Figure 11: Sales vs Quantity sold trend

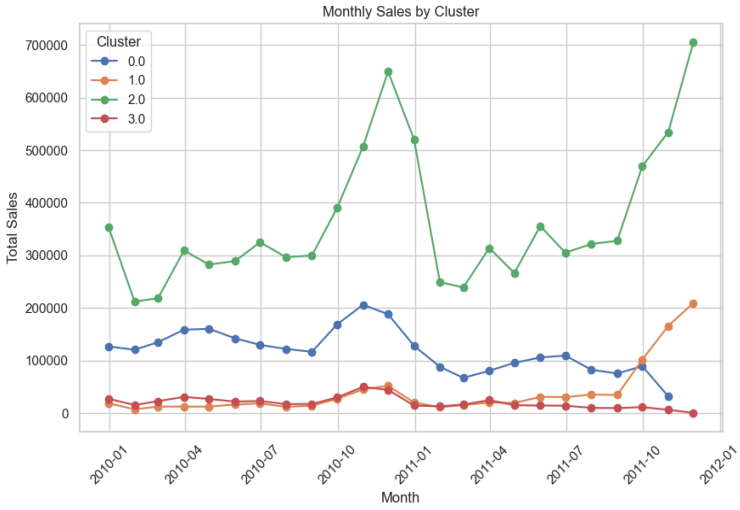
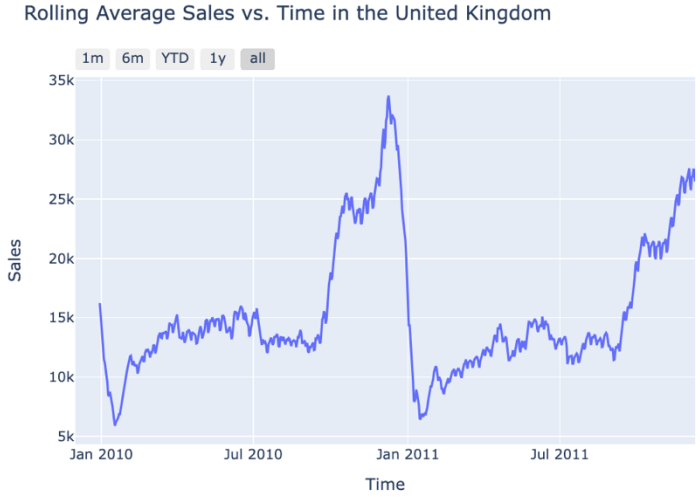
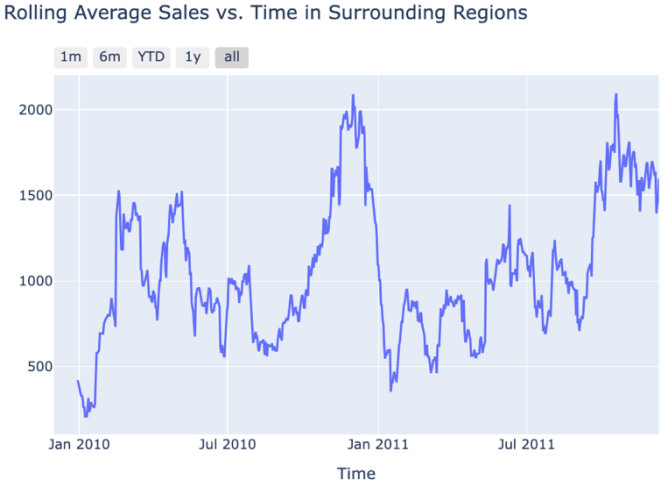


Figure 12: Cluster over Total sales trend

UK is the country with the most sales; hence the UK and its surrounding regions are considered. The comparative analysis between UK sales and those in proximate regions highlights minor disparities with an insignificant impact on overall sales. The analysis also uncovered a clear seasonal pattern: sales peaked during the winter months, remained relatively flat throughout the rest of the year, with a noticeable decline in February. These findings underscore the importance of the seasonal nature of products in sales performance.





Time-series decomposition is an analytical method that dissects a time series into three primary components: trend, which shows the overall direction of the data over time; seasonality, which captures regular patterns or cycles; and noise, which consists of random fluctuations that cannot be attributed to the trend or seasonal factors.

Seasonal decomposition of the sales data was conducted on the monthly aggregated totals. The **seasonal_decompose** function from the **statsmodels.tsa.seasonal** library was utilized, applying an additive model.

The decomposition indicates a peak in sales towards the middle of the third quarter. This could reflect cyclic customer behaviour or strategic marketing efforts impacting sales volume. Furthermore, the analysis uncovers that seasonal peaks align with the fourth quarter, which could correspond with key commercial events or holiday seasons known to drive consumer spending. Seasonal decomposition not only enhances the accuracy of sales forecasts but also informs a proactive business strategy tailored to temporal sales patterns.

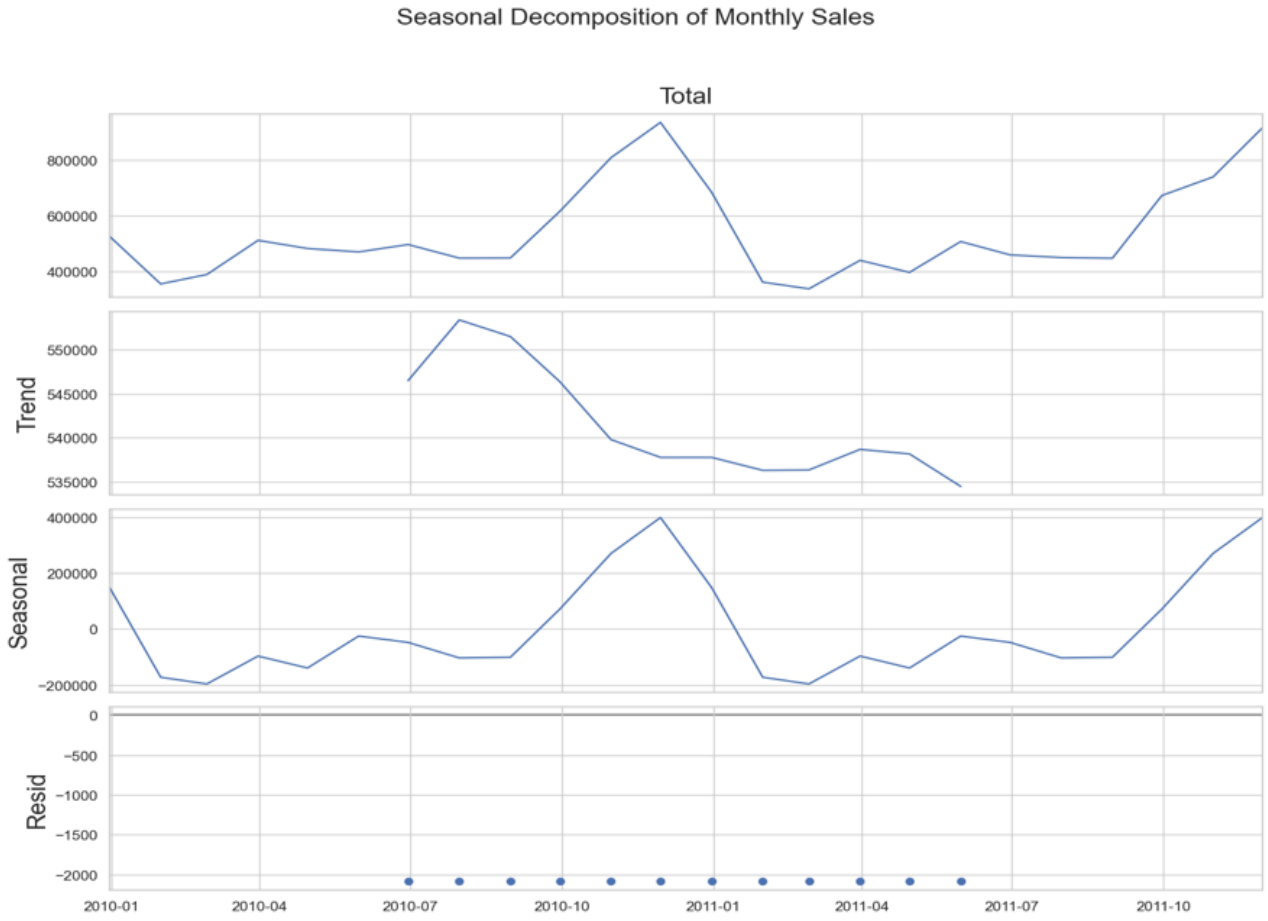


Figure 12: Seasonal decomposition graph of total sales

5.2 Model selection

Using a 3-month lag regression feature, we achieved a mean squared error (MSE) of 2347.72, indicating a modest level of accuracy in sales forecasting. When comparing more complex models, the Random Forest and Gradient Boosting methods demonstrated lower MSEs of 2331.65 and 2222.70, respectively, suggesting better predictive performance than the lag-based regression.

After implementing a time-sequential split, the Linear Regression model's MSE dropped to 204.89, and Gradient Boosting achieved an impressive 134.73, whereas the MSE for the Random Forest increased significantly to 5493.83. This stark improvement in MSE for the Linear Regression and Gradient Boosting models, and the deterioration for Random Forest, highlights the critical impact of temporal alignment in model training. The exceptional performance of Gradient Boosting suggests

its superior suitability for this dataset, likely due to its ability to effectively leverage past data trends and correct from previous errors incrementally.

Cross validation of Random Forest and Gradient Boosting further indicates that Gradient Boosting has better potential to generalize the data. Notably, a pronounced spike in MSE during fold 4 for Gradient Boosting, as opposed to Random Forest, signals an area of the dataset that may benefit from further investigation. It implies that certain characteristics of the data, perhaps peculiar to specific periods such as holiday seasons, or under certain trends, are more effectively captured by the Random Forest model. This discrepancy suggests that while Gradient Boosting may excel in general conditions, Random Forest could offer superior performance in analyzing data with seasonal trends or during atypical time frames.

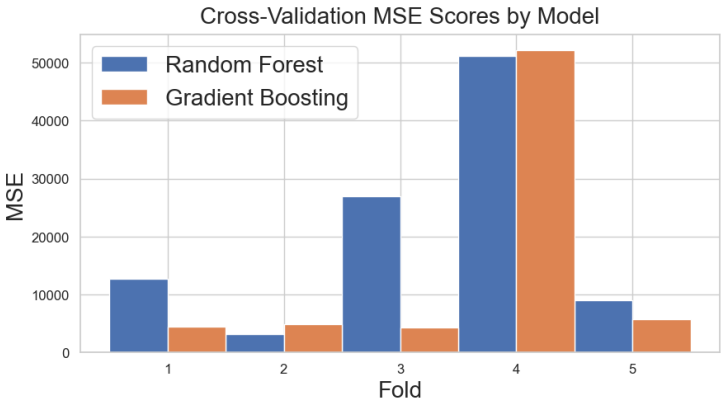


Figure 9: K-folds cross validation results of MSE

5.3 Business implication

When the model performance has been optimized, project managers may prepare it for deployment by wrapping the model in an API for integration with business processes or creating a dashboard for easy forecasts. Continuous analysis and tuning of the model should be done routinely with the collection of new sales data to ensure model and result consistency.

With an ever-increasing number of competing services available, businesses need to focus efforts on maintaining continuous consumer satisfaction, optimizing inventory, and capitalizing on data intelligence (Tijan and Sameer, 2016).

6. Conclusion

The study demonstrates how ML transforms POS data into actionable insights for retail strategy. From data refinement to the deployment of clustering, classification, and regression models, ML proves to be a robust tool for dissecting complex patterns and guiding business decisions.

Clustering was used to segment customers and refine marketing efforts, while classification models predict new customer segments to improve engagement. Regression was used to forecast sales with higher accuracy, identifying patterns crucial for inventory and pricing strategies.

The initial premise stands validated: ML significantly enhances retail operational efficacy. The challenges posed by data integrity and model optimization have been successfully navigated, confirming the value of linear and nonlinear ML techniques in retail strategy enhancement.

In summary, this research underscores the adaptability and potency of ML in retail analytics, serving as a beacon for ongoing strategic decision-making and business intelligence.

Citation:

(Di Sia P.,2022). Di Sia, P. (2022). Industry 4.0 Revolution: Introduction. In: Hussain, C.M., Di Sia, P. (eds)
Handbook of Smart Materials, Technologies, and Devices. Springer, Cham. https://doi.org/10.1007/978-3-030-84205-5_88

(Cumby *et al.*, 2004) Cumby, C. et al. (2004) 'Predicting customer shopping lists from point-of-sale purchase data', Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '04 [Preprint]. Available at: https://doi.org/10.1145/1014052.1014098.

(Hsieh, Giloni and Hurvich, 2020) Hsieh, M.-C., Giloni, A. and Hurvich, C. (2020) 'The propagation and identification of ARMA demand under simple exponential smoothing: forecasting expertise and information sharing', IMA Journal of Management Mathematics, 31(3), pp. 307–344. Available at: https://doi.org/10.1093/imaman/dpaa006.

(Hwang *et al.*, 2023) Hwang, S. et al. (2023) 'A Sales Forecasting Model for New-Released and Short-Term Product: A Case Study of Mobile Phones', Electronics, 12(15), pp. 3256–3256. Available at: https://doi.org/10.3390/electronics12153256.

(Christy et al., 2018) Christy, A.J., Umamakeswari, A., Priyatharsini, L. and Neyaa, A. (2018). RFM Ranking – an Effective Approach to Customer Segmentation. Journal of King Saud University - Computer and Information Sciences, 33(10). doi:https://doi.org/10.1016/j.jksuci.2018.09.004.

(Varad R Thalkar, 2021) Varad R Thalkar (2021) 'Customer Segmentation Using Machine Learning', International Journal of Scientific Research in Computer Science, Engineering and Information Technology, pp. 207–211. Available at: https://doi.org/10.32628/cseit217654.

(Razia Sulthana A et al., 2023) Razia Sulthana A, Jaiswal, A., Supraja P and Sairamesh L (2023). Customer Segmentation using Machine Learning. Uppsala University Publications (Uppsala University). doi:https://doi.org/10.1109/icaect57570.2023.10117924.

(Ketchen and Shook, 1996) Ketchen, D.J. and Shook, C.L. (1996). The Application of Cluster Analysis in Strategic Management Research: An Analysis and Critique. Strategic Management Journal, [online] 17(6), pp.441–458. Available at: https://www.jstor.org/stable/2486927.

(Bhade et al., 2018) K. Bhade, V. Gulalkari, N. Harwani and S. N. Dhage, "A Systematic Approach to Customer Segmentation and Buyer Targeting for Profit Maximization," 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Bengaluru, India, 2018, pp. 1-6, doi: 10.1109/ICCCNT.2018.8494019.

(Chang and Tsai, 2011) Chang, H.-C. and Tsai, H.-P. (2011) 'Group RFM analysis as a novel framework to discover better customer consumption behavior', Expert Systems with Applications, 38(12), pp. 14499–14513. Available at: https://doi.org/10.1016/j.eswa.2011.05.034.

(Kohli, Godwin and Urolagin, 2020) Kohli, S., Godwin, G.T. and Urolagin, S. (2020). Sales Prediction Using Linear and KNN Regression. Algorithms for Intelligent Systems, pp.321–329. doi:https://doi.org/10.1007/978-981-15-5243-4_29.

(Tijan and Sameer, 2016) Icrie2016, U.G. of I. - (2016) 'Predictive Modelling for Assessing the Sales Potential of the Customer', www.academia.edu [Preprint]. Available at: https://www.academia.edu/28362014/Predictive_Modelling_for_Assessing_the_Sales_Potential_of_the_Customer