GROUP PROJECT REPORT

Sales prediction for a big retail firm

OCTOBER 11, 2024

AUCKLAND UNIVERSITY OF TECHNOLOGY (AUT)

Table of Contents

1. Introduction and Business Problem

The retail store aims to boost profitability by leveraging data-driven insights to better understand their consumers. This report will present findings on the successful and poor performance across products, regions, categories, and customer segments, based on insights derived from the analysis of Forest Modelling, Cluster Analysis, and Regression. The presented models will be refined and evaluated for reliability. These will then be analysed to identify areas that the retail store has done well to prioritise as well as areas that may benefit with extra attention. Then, suggestions will be offered on how the retail store can tailor its approach and increase its profitability and consequently its competitive advantage. Finally, data governance principles will be discussed in relation to current operations, and recommendations will be provided to inform for best practice.

2. Data Cleaning and Preparation

Firstly, the “Select” tool was used to determine that the correct data types were selected. These were largely accurate; however, it is of interest to note that though “Postal Code” appears to be numerical, it is categorical and not quantifiable. As there were missing values in three variables, namely “Category”, “Product Name” and “State”, the next step to ensure a complete and accurate dataset was to populate these; Fortunately, the dataset provided by the company included the correct values for the missing values in different fields, so these were filled by using “Multi-Row Formula”, searching for “Product ID” for the former two variables and “Postal Code” for the latter. This process was applied to clean the data, before further manipulation for the following analyses.

2.1 Classification

The “Formula” tool was utilized to create a new column that categorised profitability into a binary "Yes" or "No", based on whether it turned a profit or loss. This prepared data for use in classification analysis and allowed ease in determining profitable and unprofitable transactions. A sample tool was included in the workflow, to split the dataset into training, test, and holdout data, to enable validation and evaluation of the model.

2.2 Clustering

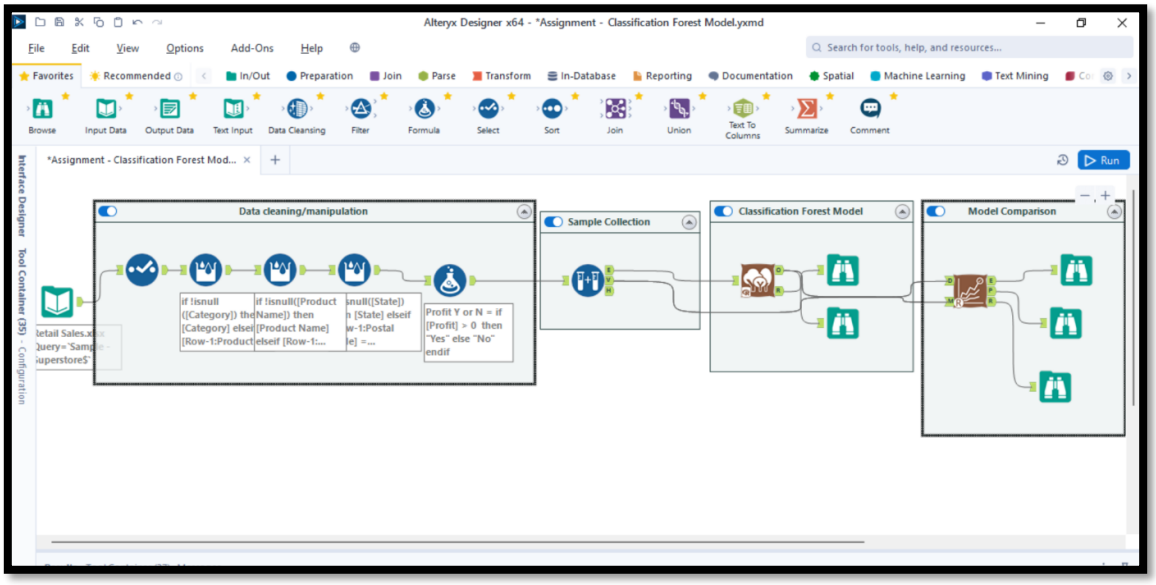
For this analysis, we used the variables "Sales," "Quantity," "Discount," and "Profit." We then ran a “K-Centroid Diagnostics” to determine the optimal number of clusters for clustering analysis. The diagnostics indicated that four clusters were optimal, as the data was evenly distributed across the boxplot, with the median positioned close to the mean. The data was standardised by using the z-score within the “K-Centroids Cluster Analysis” tool to ensure that all variables were on a comparable scale, preventing any singular variable from disproportionately influencing and skewing the clustering results.

2.3 Regression

A “Select” tool was applied to choose the relevant variables, such as “Sales,” “Quantity,” “Discount,” and “Profit,” for regression analysis. Additionally, columns that held valuable insights but were string data, such as “State,” “Region,” “Category,” and “Sub-Category,” were converted into numerical data by one hot encoding with the “Python” tool.

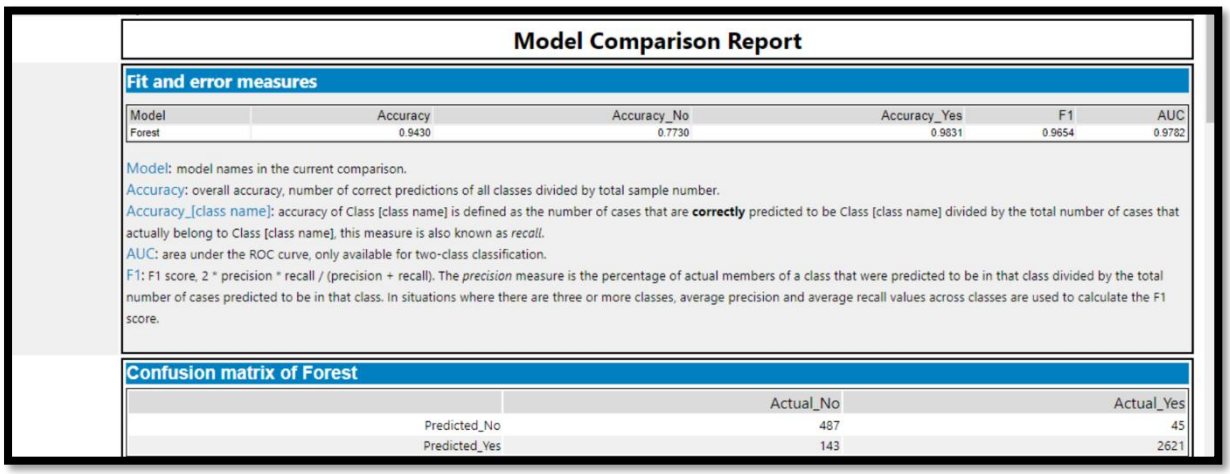
3. Predictive Model Building & Evaluation

3.1 Classification



The new column created in the data preparation stage (“Profit Y or N”) was selected as the target variable for classification to allow forecasting on whether a transaction was likely to be lucrative (“Yes”) or unprofitable (“No”). The predictor variables selected were “Segment”,

"State", "Region", "Category", "Subcategory", "Sales", "Quantity", and "Discount", as provided in the dataset from the retail company. We utilised the "Forest Model" tool to model the data and assessed the model's predictive performance using the tool's inbuilt measures.



Model Comparison Report

Fit and error measures

Model	Accuracy	Accuracy_No	Accuracy_Yes	F1	AUC
Forest	0.9430	0.7730	0.9831	0.9654	0.9782

Model: model names in the current comparison.
Accuracy: overall accuracy, number of correct predictions of all classes divided by total sample number.
Accuracy_[class name]: accuracy of Class [class name] is defined as the number of cases that are **correctly** predicted to be Class [class name] divided by the total number of cases that actually belong to Class [class name], this measure is also known as *recall*.
AUC: area under the ROC curve, only available for two-class classification.
F1: F1 score, 2 * precision * recall / (precision + recall). The *precision* measure is the percentage of actual members of a class that were predicted to be in that class divided by the total number of cases predicted to be in that class. In situations where there are three or more classes, average precision and average recall values across classes are used to calculate the F1 score.

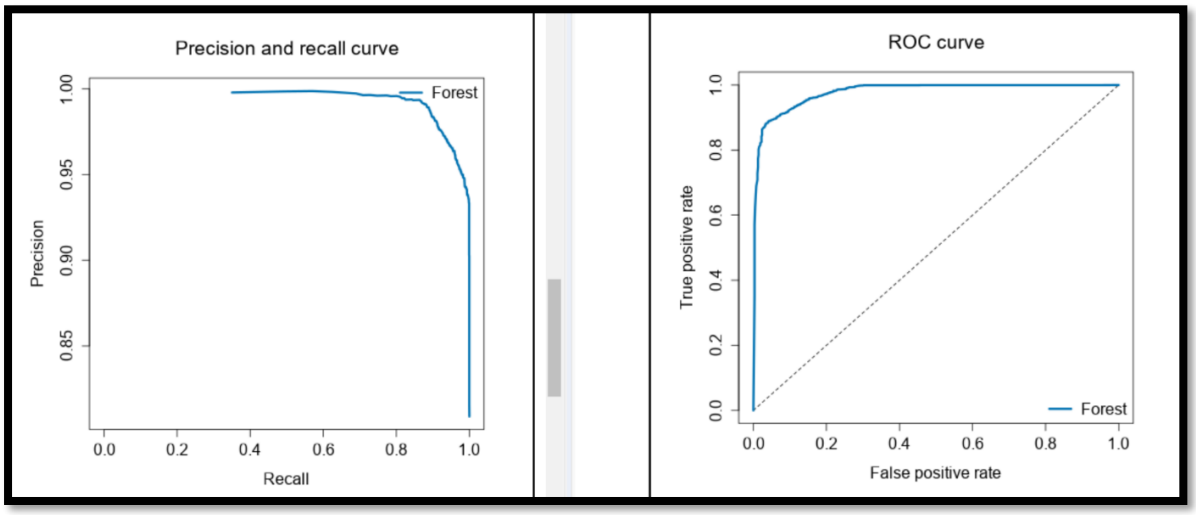
Confusion matrix of Forest

	Actual_No	Actual_Yes
Predicted_No	487	45
Predicted_Yes	143	2621

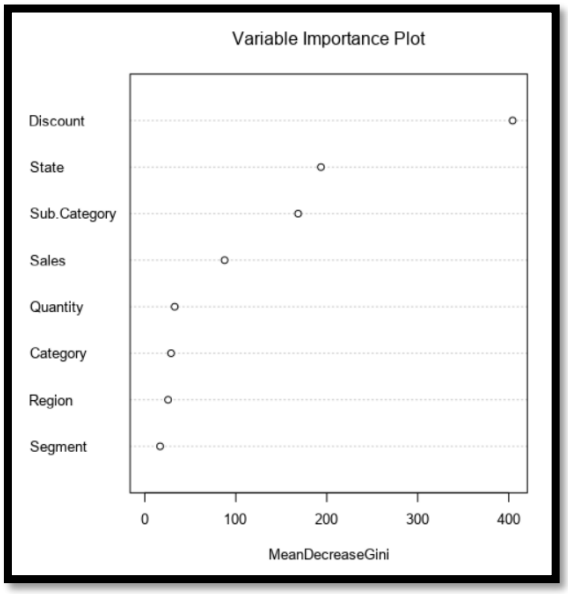
It was noted that the overall accuracy, describing the frequency of data points accurately classified as profitable or non-profitable transactions, of the model was 94.30% indicating that the "Forest Model" was highly accurate in predicting transactional profitability. However, it is of note that the model was significantly better at predicting profitable transactions (98.31%), compared to unprofitable transactions (77.30%), indicating high precision of the model. This occurred due to the significantly smaller proportion of unprofitable data points (651) compared to profitable data points (2747) that the model had to train with, which is unfortunate in terms of model accuracy, but expected and encouraging as it affirms that the retail business is performing at an acceptable level.

The model achieved an excellent F1 score of 0.9654, which combines precision and recall into a single measure to balance the two. Precision is the ratio of true positive to total positive predictions, i.e., true positives and false positives; while recall (also called sensitivity) is the ratio of true positives to the actual positives, i.e., true positives and false negatives. Both precision (0.9483) and recall (also known as sensitivity, with a value of 0.9831) were high, affirming the reliability of the model. This is reassuring as some of the classes were imbalanced, as aforementioned, so the high F1 score offers confidence that the model performed well in consideration of the true positives and the incorrectly predicted

classes. Moreover, the model's Out-of-Bag (OOB) error rate was 6.4%, which indicates that it is highly accurate in predicting outcomes with data it has not been trained on.

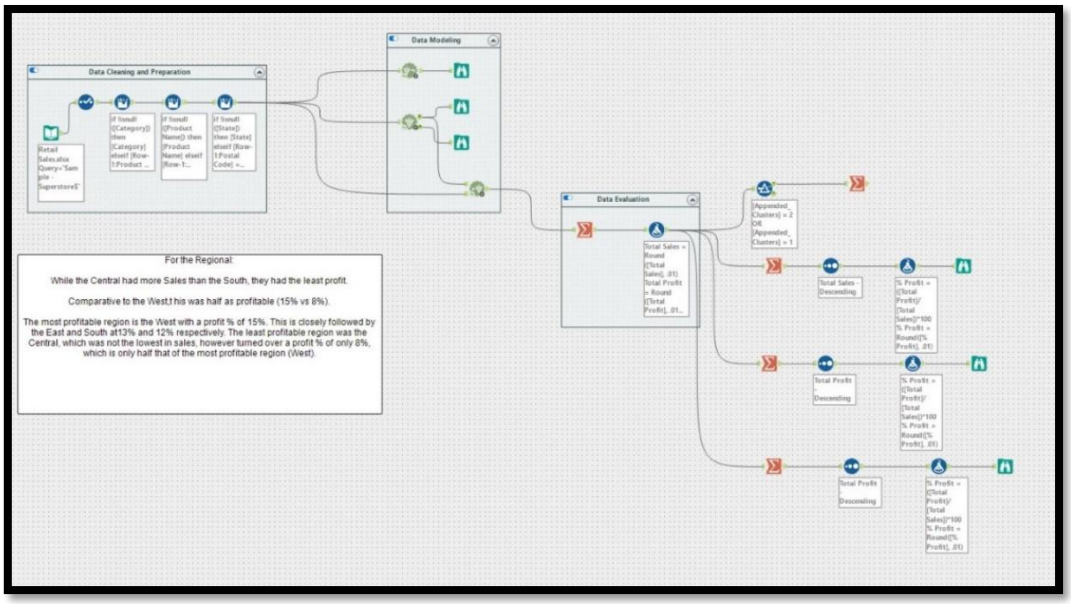


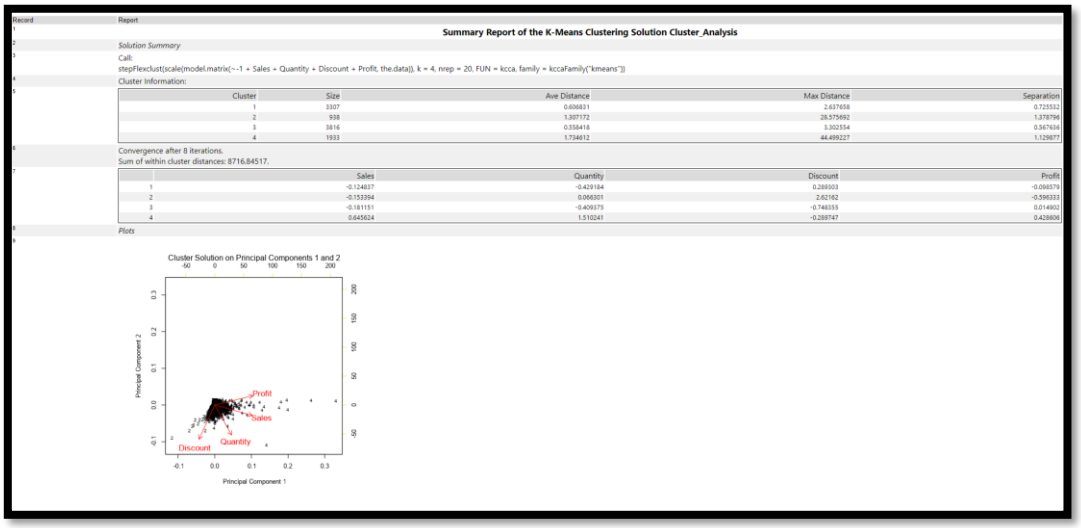
The AUC (area under curve) of the ROC curve, as shown above, is a performance indicator that assesses the positive predictions predictive by the model. The excellent AUC score of 0.9782, in combination of the high top-left-corner leaning curve plotted indicate that the "Forest Model" was very good at predicting profitability.



Considering the high accuracy, precision, recall, F1 score and AUC measures, it can be safe to say that the model was satisfactory. Thus, it can be inferred from the Variable Importance Plot generated by the "Forest Model" that Discount was substantially found to be the strongest predictor of profitability, followed by State and Subcategory.

3.2 Clustering





Based on initial analysis, clusters 1 and 3 were the largest with over 3000 data points, followed by cluster 4, with 1933 data points. Cluster 4 had the highest average and maximum distances, indicating diversity of its data points. Further analysis revealed that it is the most successful cluster, showing high sales, quantity, and profit alongside low discounts. On the other hand, cluster 2 had the highest discounts and negative profits, suggesting that excessive discounting adversely affected the profitability of the products in this cluster. This is in corroboration with findings in the “Forest Model” that discounting has strong influence on profit. These are in alignment with the data presented in the summary report below, which supports that cluster 4 has the highest sales and profit.

Results - Summarize (56) - Output

4 of 4 Fields | Cell Viewer | 4 records displayed

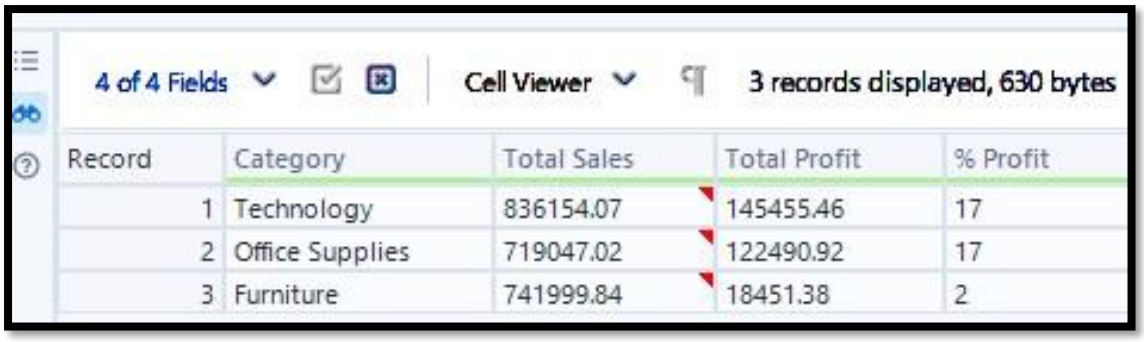
Record	Appended_Clusters	Sum_Total Sales	Sum_Total Profit	Sum_Total Quantity
1	4	1222119.99	249477.46	13821
2	1	502841.83	18400	9374
3	3	446307.1	122675.97	10985
4	2	125932.01	-104155.67	3693

Results - Browse (38) - Input

5 of 5 Fields | Cell Viewer | 4 records displayed, 762 bytes

Record	Region	Total Sales	Total Profit	Sum_Total Quantity	% Profit
1	West	725457.81	108418.77	12266	15
2	East	678781.33	91522.88	10618	13
3	Central	501239.9	39706.41	8780	8
4	South	391721.89	46749.7	6209	12

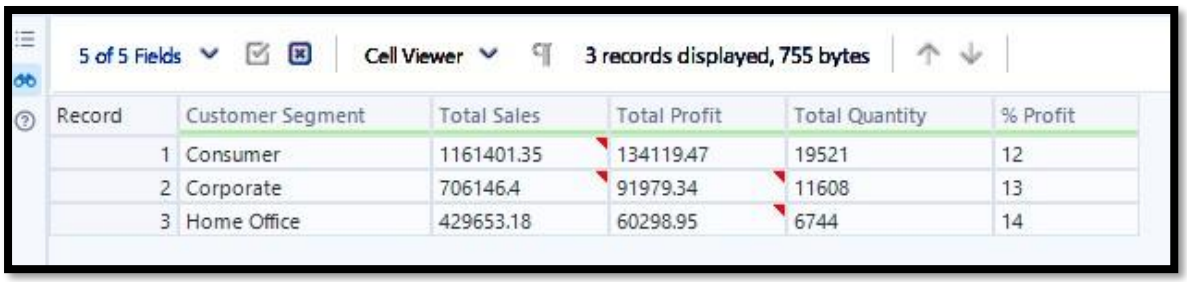
By appending cluster information with a focus on “Region” information, it was found that the most profitable region is the West with a profit percentage of 15%, calculated by applying a formula to find total profit over total sales. This is closely followed by the East and South at 13% and 12%, respectively. The least profitable region was Central, which was not the lowest in sales but turned over a profit percentage of only 8%, which is only half that of the most profitable region (West). Conversely, while the South had the least sales, the stores managed to nearly match the top two most profitable stores in profit percentage and outperformed the Central region in terms of total profit, indicating a possibility of efficient cost management or better pricing strategies with less requirement of discounts to move stock. Marketing efforts have been successful in developing the high-performing West and East regions and should continue, but the company could also consider putting resources towards increasing its market share in the equally profitable South to drive higher sales and increase profit.



4 of 4 Fields | Cell Viewer | 3 records displayed, 630 bytes

Record	Category	Total Sales	Total Profit	% Profit
1	Technology	836154.07	145455.46	17
2	Office Supplies	719047.02	122490.92	17
3	Furniture	741999.84	18451.38	2

When analysing “Category”, the Technology and Office Supplies categories performed strongly, each achieving a profit margin of 17%. The retail company could capitalize on the strong performance of Technology and Office Supplies by implementing cross-selling and expanding product offerings. In contrast, despite high total sales, the Furniture category was distinctly the least profitable, with a profit margin of just 2% which put it only barely in the black. This finding appears to be due to large discount margins which, while effective in driving sales, substantially reduced overall profitability. The poor performance in the Furniture category aligns with the characteristics of clusters 1 and 2, which were characterised by high discounts and low profits. This echoed the profitability challenges faced by the Central region. Tailoring strategies to this large but less profitable customer segment will be critical in enhancing engagement and overall profitability. The retail store should focus on optimizing strategies, particularly in the Furniture category, to increase profitability. An idea could be offering bundling or financing options, such as Zip or Afterpay, in place of excessive discounting to enable the retail company to keep consumers who may struggle with the cost of big-ticket items without losing on potential profit.



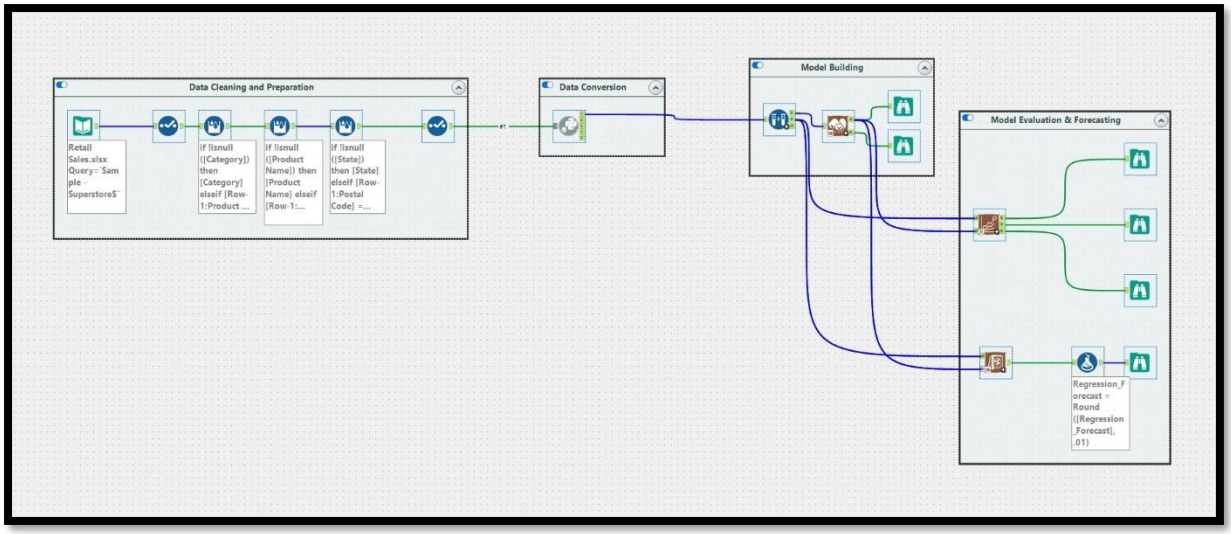
5 of 5 Fields | Cell Viewer | 3 records displayed, 755 bytes

Record	Customer Segment	Total Sales	Total Profit	Total Quantity	% Profit
1	Consumer	1161401.35	134119.47	19521	12
2	Corporate	706146.4	91979.34	11608	13
3	Home Office	429653.18	60298.95	6744	14

As an extra note, there was low variance in profit percentage, but consumers make up the largest proportion of our sales and total profits. This could be a customer segment that the company could focus on, such as through increasing stock of popular items for the Consumers segment.

3.3 Regression Forecasting

Regression forecasting was conducted to predict profit with the aim to gain insights into which factors most significantly influence profitability, to allow the store to make data-driven decisions on based on profit. Before conducting the regression analysis, we set up the Forest Model tool to utilize "Profit" as the target variable along with the chosen predictors, as aforementioned in question 2.3. Additionally, we specified the number of trees to be 1,000.



Model Comparison Report

Fit and error measures

Model	Correlation	RMSE	MAE	MPE	MAPE
Regression_Forecast	0.7679	209.9945	34.8588	3.4616	112.2731

Model: model names in the current comparison.
Correlation: correlation between the predicted values and the actual values.
RMSE: root mean square error.
MAE: mean absolute error.
MPE: mean percentage error. Note: based on its definition, MPE may be positive or negative infinity if the target variable has 0 values. In this case, we return a weighted percentage error (WPE).
MAPE: mean absolute percentage error. Note: based on its definition, MAPE may be positive or negative infinity if the target variable has 0 values. In this case, we return a weighted absolute percentage error (WAPE).

Our regression model for forecasting profit demonstrates several key performance metrics. Firstly, with a correlation value of 0.7679, this indicated a moderately positive relationship between the predictors and profit. This suggests that as the predictors increase, profit tends to increase as well. The model's Root Mean Square Error (RMSE) of $209.99, Mean Absolute Error (MAE) of $34.86, and Mean Percentage Error (MPE) of 3.46% indicate the profit predictions deviation from the actual values which are acceptable but with room for improvement. Meanwhile, the high Mean Absolute Percentage Error (MAPE) of 112.27%

points to significant variation in the forecast, likely due to outliers or complex patterns in the data that the model struggles to capture effectively.

Overall, the model demonstrates a moderate correlation of the predictors with profit, but the error measures suggest a need for further refinement, such as using more relevant predictors or enhancing data preprocessing techniques.

10 of 10 Fields | Cell Viewer | 3,298 records displayed, 87 KB | Search

Record	Segment	State	Region	Category	Sub.Category	Sales	Quantity	Discount	Profit	Regression_Forecast
1	0	3	3	0	4	239.666	2	0.15	14.098	4.05
2	1	3	3	2	13	1626.192	9	0.2	121.9644	180.64
3	2	3	3	2	0	399.96	4	0	139.986	101.32
4	0	3	3	1	3	61.12	5	0.2	22.156	22.8
5	1	25	0	2	13	79.96	4	0	22.3888	73.91
6	2	25	0	0	9	15.92	2	0	7.0048	12.4
7	1	25	0	1	3	53.9	5	0	25.872	24.19
8	1	41	0	1	3	1.964	2	0.8	-3.2406	-15.77
9	0	41	0	1	3	10.024	4	0.8	-16.5396	-58.63
10	0	41	0	1	14	118.16	2	0.2	-25.109	-4.74
11	2	41	0	1	3	5.792	2	0.8	-9.5568	-15.57
12	1	41	0	2	13	492.768	4	0.2	55.4364	55.18

4. Data Governance and Ethical Considerations

Data governance is of great importance and is unsurprisingly so in the case of our retail company, which unavoidably collects and stores consumer data due to the nature of its business. Each country and state have their own regulations around data; however, the Institute of Internal Auditors (2020) summarizes a few key principles that are commonly cited in legislation internationally:

- Ownership of personal data belongs to the individual.
- Transparency on usage of personal data and access to aggregate datasets.
- Consent requirements for other individuals or legal entities to use personal data.
- The highest reasonable effort must be expended to protect privacy.
- Right to know if financial transactions involving personal data occur, and on what scale.

Domestically, any company conducting business within New Zealand must abide by the Privacy Act (2020), which emphasizes 13 Information Privacy Principles (IPPs). In addition to those principles stated by the Institute of Internal Auditors (2020), the Privacy Act (2020) references the right to access and correct personal information. Specifically, the IPPs require companies to consider the purpose, source, users, collection methods, storage, security, access, retention, usage, privacy, and legal and cross-border disclosure of their consumers'

personal data, as well as give provisions to their consumers to edit personal data for accuracy. Such legislation provides the retail company with a framework for dealing with consumers and guides them on best practice when working with personal data.

The retail company can develop and establish an ethical and compliant data environment by outlining data governance policies that employees must abide by in their organisational code of conduct. Further, they can selectively grant access to personal data to relevant personnel and provide training on ethics and procedures around handling data, so that they can be held accountable and become more vigilant around securing company devices to prevent data leakage. The retail company could also consider applying principles of the CAN-SPAM Act (U.S.A) in their communications and include a privacy policy, avoid misleading information, create clear ads, provide company details and location, give consumers an opt-out option and process these promptly upon request, and monitor any external parties engaged in marketing to ensure they treat sensitive data in an acceptable manner. Retention of consumers' data should also be a consideration in the event they choose to opt-out; If possible, it should be disclosed whether data will be deleted, including from backup servers; Otherwise, there should be a process in place to de-personalize data, so they are not easily linked to an individual. Depending on the scale of data usage at the retail company, it could also be useful to set up a committee to have oversight of data governance and guide employees who handle data as part of a data office. Finally, it is important that regular audits are conducted to ensure that the company meets its expectations around data.

5. Conclusion

The Forest Model, Cluster Analysis, and Regression were employed to better understand the retail store's performance across products, regions, categories, and customer segments. All models provided satisfactory metrics which justifies the credibility of the insights derived during analysis. One major theme was that discounting, especially for products in the Furniture category and in the Central region, significantly reduced overall and profit percentage made by the company. The retail company barely covered their costs within the Furniture category with only a 2% profit percentage, and the Central region underperformed with a profit percentage that was approximately half that of the three other regions.

Improvements discussed to improve Furniture category performance were centred on different ways of retaining customers without providing excessive discounts, such as bundling and financing options, which will possibly also help to increase the Central region's performance. The South region was found to be an area where there was an opportunity for growth as, though the retail company's market share was not large, consumers in the region were willing to spend money. It is therefore recommended that marketing resources could focus on expanding business in the South, while maintaining the status quo with the West and East regions.

The data used in this analysis was managed with the utmost care upon engagement with the retail company, in accordance with the Privacy Act (2020). An audit of the current structure was completed, and recommendations were suggested for the retail company for best practice, including guidelines on communicating with customers if the company decides to pursue targeted marketing, as suggested.

References

CAN-SPAM Act (U.S.A)

Privacy Act 2020

The Institute of Internal Auditors. (2020). *Data governance: Providing assurance regarding data risk management.* https://www.theiia.org/globalassets/site/content/articles/industry-knowledge-brief/2020/data-governance/data-governance.pdf