Machine Learning in Business

MIS710 – A2 (Part A)
Case Study Report (Technical Report)

Great Ocean Banking Group

Victoria, Australia

Sachin Bhat, 218676233

DEAKIN BUSINESS SCHOOL

CONTENTS

Executive Summary

This report is presented to Kathy Hoang, Head of Data Analytics, Great Ocean Banking Group and prepared by business analyst Sachin Bhat. It addresses marketing difficulties faced by the bank, intending to improve their sales campaigns' effectiveness. Through customer data analysis and using machine learning models, we identified key predictors of successful sales outcomes. We developed and evaluated two predictive models: kNN and Decision Trees. Our discoveries revealed that post-pruned DT outperformed kNN, providing higher performance results. We recommend deploying this DT model because of its superior performance and easy interpretation. We also applied k-Means clustering to segment the customer base for effective targeted marketing strategies. Optimal number of clusters= 3 were determined based on the metrics like WCSS, Davies-Bouldin Index, and Silhouette Score enabling more targeted and effective marketing efforts. Value gains: enhanced campaign efficiency and customer experience, great word-of-mouth, improved customer acquisition, retention, and revenues.

Introduction

Great Ocean Bank Victoria, Australia caters to over one million customers and offers a wide range of financial services like personal loans, home mortgages, savings accounts, and credit card accounts. To promote these services among existing as well as prospective customers, the bank engages in periodic sales campaigns. They want to understand the factors influencing campaign success. We are to analyse the provided dataset which includes customer information, details of their banking relationships, records of previous marketing contacts, and several economic indicators. We want to aid the bank in improving effectiveness of their marketing campaigns. The objective is to predict potential “Sale” or “No Sale” outcome for the customers who were contacted during bank’s marketing campaigns through the use of data analytics and machine learning. We want to find the insights that can predict the success of future campaigns thereby allowing the bank to target the right customers with the right offers. The bank will be able to tailor its marketing efforts leading to an increase in customer satisfaction and conversion rates.

Value Proposition

1. **Campaign Efficiency will increase** with a decrease in wasted efforts and costs. Predicting possibility of a sale can help the bank to focus its resources on customers which are probable to respond positively.

2. **Customer Experience will improve** through targeted campaigns as they ensure that bank’s customers receive personalised and relevant offers. This will enhance overall customer satisfaction and loyalty to the bank.

3. Based on insights gained from data analysis and machine learning models, the bank can perform solid strategic **Data-driven Decision Making**. They will be able to optimise their marketing strategies based on verifiable evidence.

4. Utilising advanced analytics and machine learning, the bank will have a **Competitive Advantage** over its competitors while offering more effective and efficient financial solutions to their customers.

Approach

Data Preparation and EDA

We performed data cleaning by handling missing values ensuring data consistency. We did data exploration to understand the distribution and relationships between variables through statistical analysis and visualisations. We transformed and encoded the categorical variables, scaled the numerical variables, selected relevant features, and split the dataset: training and testing.

Supervised Machine Learning

This is a binary classification task where our **target variable** is “**Sale Outcome**” (**Sale or No Sale**). We employed two predictive model classifiers: **kNN** and **Decision Trees**. We trained the models on subset of the data and evaluated their performance using accuracy, precision, recall, f1-score, and ROC-AUC.

Unsupervised Machine Learning

We put the customers into clusters to **identify distinct segments**. We employed **k-Means** clustering to cluster customers based on similarities in their attributes. We evaluated the clustering results using Within-Cluster Sum of Squares, Davies-Bouldin Index, and Silhouette Score.

Model Optimisation and Validation

We did hyperparameter tuning for optimising model parameters to improve the performance. We used Cross-Validation techniques to evaluate model stability and generalisation.

Then, we will interpret the model outputs and provide actionable insights/recommendations to the bank for enhancing their marketing strategies. We aim to predict the sale outcome for marketing campaigns and identify distinct customer segments to target them with tailored marketing offers.

Data preparation and EDA

Data Sources: Primary dataset is ‘GOBank.csv’ which includes various customer information, marketing campaign details, and banking relationships. The metadata dataset ‘GOBank_metadata.csv’ helps to understand the variables. Both have been provided by Great Ocean Bank.

Data Size: The dataset has 22940 rows and 19 columns (one is CustomerID).

Data Types: We have datatypes: object, integer, float.

Categorical variables: ‘Qualification’, ‘Occupation’, ‘Marital Status’, ‘Home Mortgage’, ‘Personal Loan’, ‘Has Other Bank Account’, ‘Last Contact Direction’, ‘Last Contact Month’, ‘Last Contact Weekday’, ‘Previous Campaign Outcome’, ‘Sale Outcome’.

Numerical variables: ‘Age’, ‘Last Contact Duration’, ‘Number of Current Campaign Calls’, ‘Number of Previous Campaign Calls’, ‘RBA Cash Rate’, ‘Employment Variation Rate’, ‘Consumer Confidence Index’.

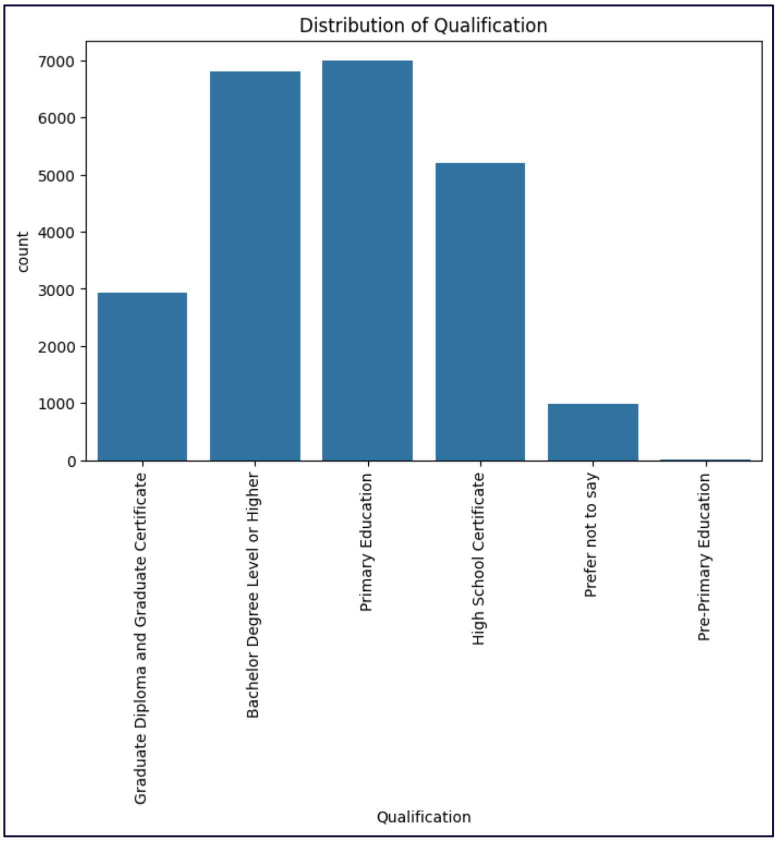
Data Quality and Cleansing: We dealt with missing values across several features using imputation. Imputed missing values for 'Qualification' (162), 'Last Contact Direction' (47), and 'Previous Campaign Outcome' (206) with mode values. Imputed missing values for 'Last Contact Duration' (141) and 'Number of Previous Campaign Calls' (26) with median values.

Pre-Processing: We used label encoding to convert categorical variables, including our target ‘Sale Outcome’ (No sale=0, Sale=1) to numerical format, suitable for ML models. We normalised the data by scaling numerical variables using StandardScaler.

Exploratory Data Analysis (EDA)

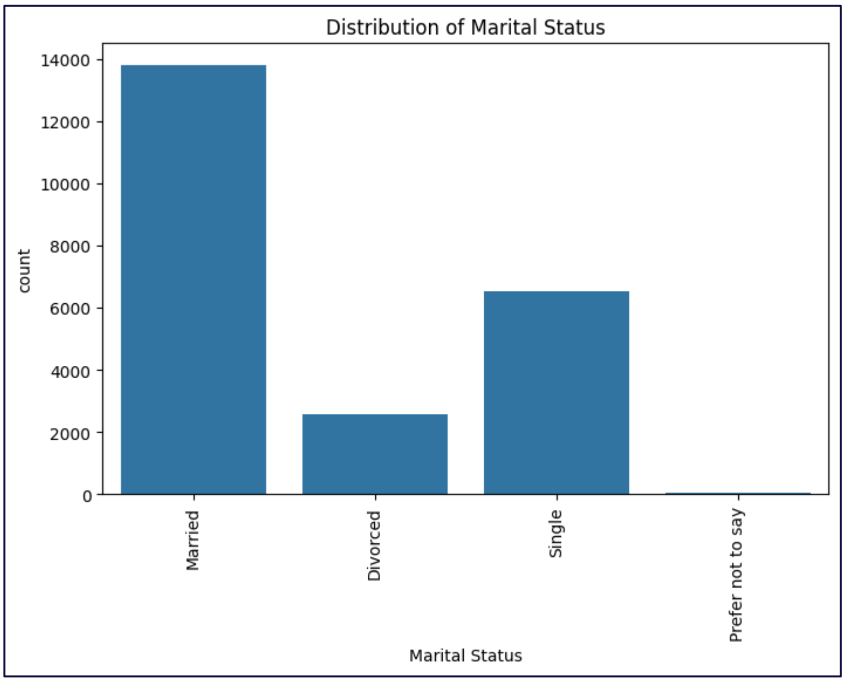
Univariate Analysis

Qualification



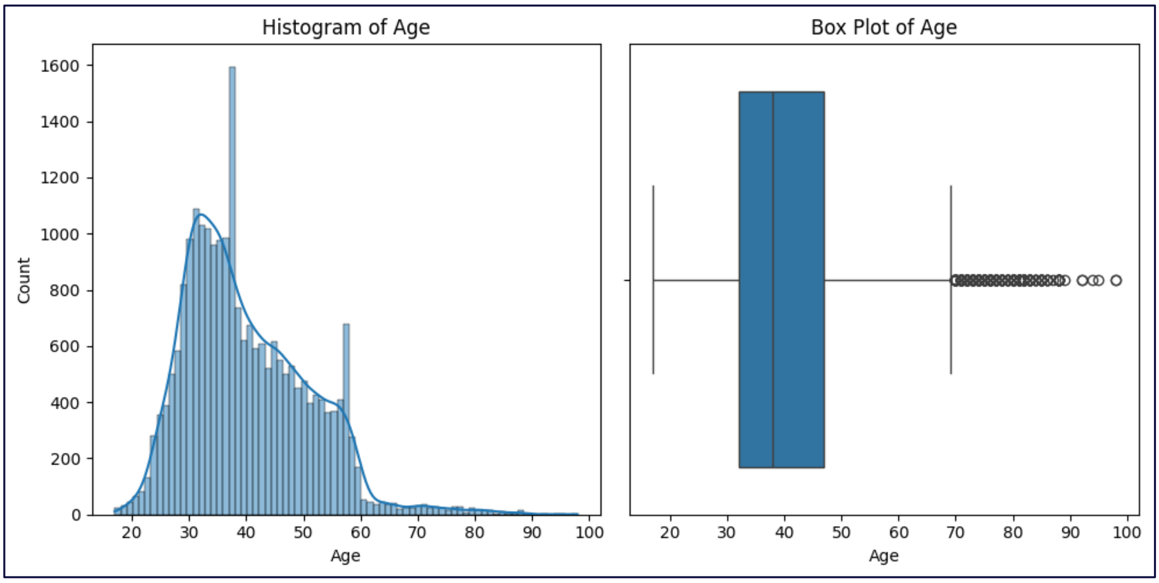
Most customers (6996) have Primary Education (30.50%), 6815 have Bachelor Degree Level or Higher (29.70%), 5198 have High School Certificate (22.66%), 2934 have Graduate Diploma and Graduate Certificate (12.79%), 15 have Pre-Primary Education (0.06%). 982 customers Prefer not to say their qualification (4.28%).

Marital Status



13804 customers are Married (60.17%), 6526 are Single (28.45%), 2259 are Divorced (11.16%), and 51 Prefer not to say their marital status (0.22%).

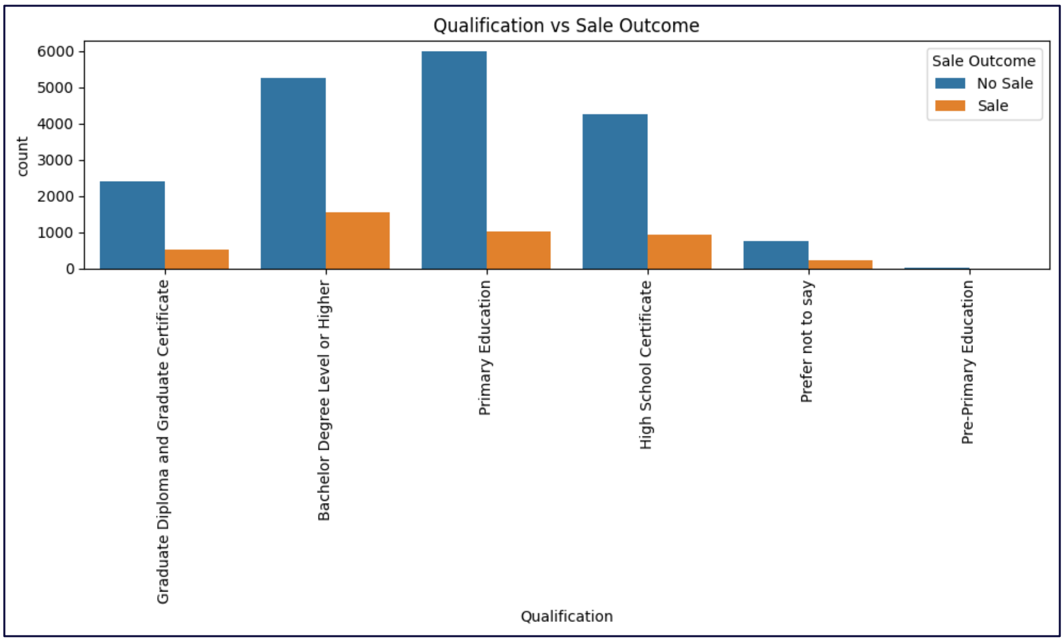
Age



The average age is 40.10 years, and median age is 38 years, depicting the relatively young customer base. Standard deviation of 10.80 years depicts moderate viability in the ages of customers. There are several older customers beyond the age of 70 and are considered outliers. Right skewness depicts that while most of the customers are young to middle-age, there are fewer older customers. Minimum age is 17 years, the maximum is 98 years.

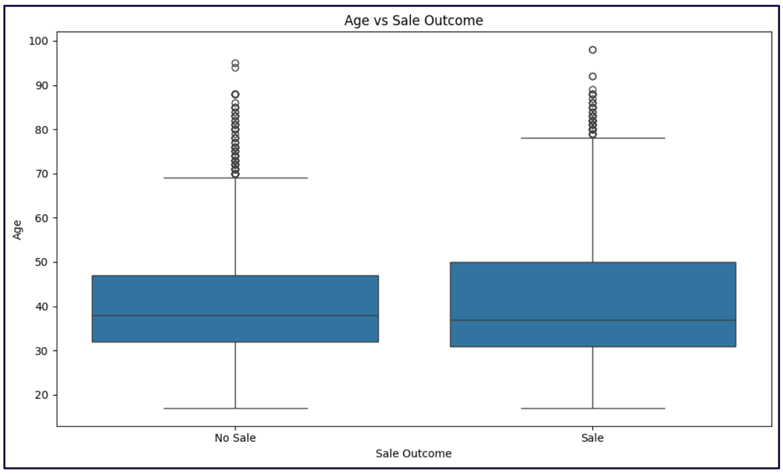
Bivariate Analysis

Qualification with Sale Outcome



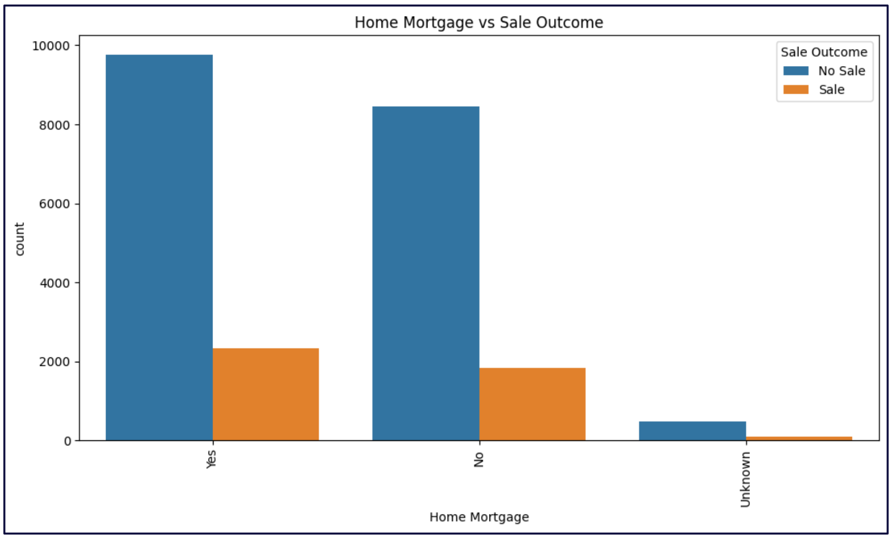
Graduate Diploma and Graduate Certificate: 553 (18.17%) customers show Sale, while 2401 (81.83%) show No Sale. **Bachelor Degree Level or Higher:** 1542 (22.63%) customers show Sale while, 5273 (77.37%) show No Sale. **Primary Education**: 1014 (14.50%) customers show Sale, while 5982 (85.50%) show No Sale. **High School Certificate**: 936 (18.01%) customers show Sale, while 4262 (81.99%) show No Sale. **Prefer not to say**: 234 (23.83%) customers show Sale, while 748 (76.17%) show No Sale. **Pre-Primary Education**: 4 (26.67%) customers show Sale, while 11 (73.33%) show No Sale.

Age with Sale Outcome



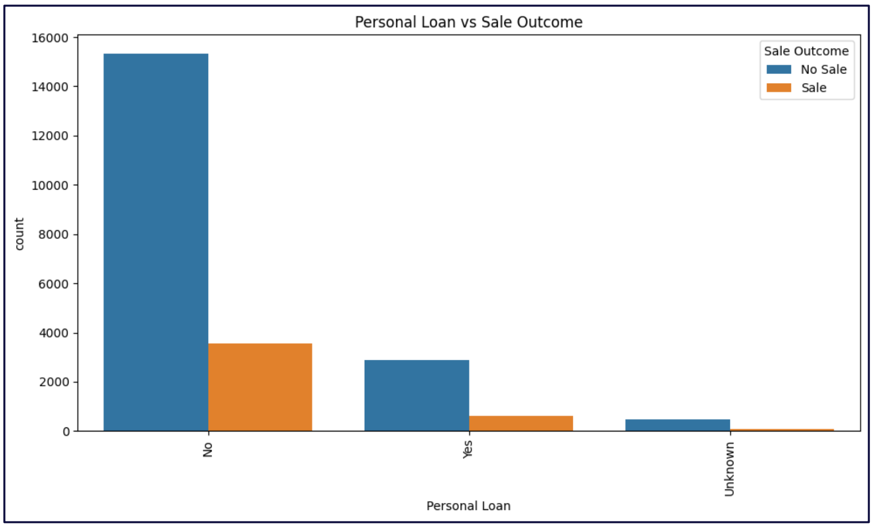
Mean age of customers showing Sale: 40.91 years, and for No Sale: 39.91 years. Median age for both groups is close to the mean with median for Sale: 37 years, and for No Sale: 38 years. The standard deviation for ‘Sale’ group: 13.79, while for the ‘No Sale' group: 9.98 represents lesser variability in age. The IQR for ‘Sale’ group (19 years) is wider as compared to the ‘No Sale’ group (15 years), representing greater age diversity among ‘Sale’ group. Both groups have several outliers in the higher age range with a maximum age of 98 years in the ‘Sale’ group.

Home Mortgage with Sale Outcome



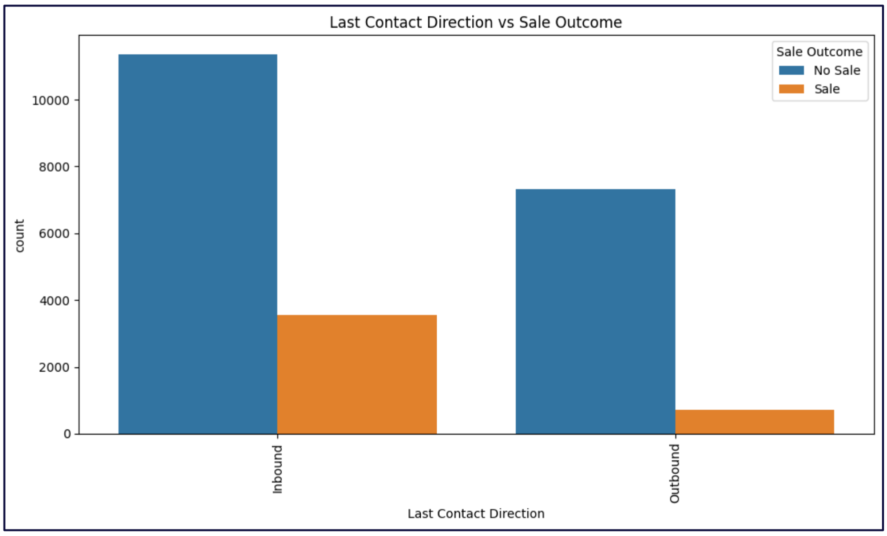
Home Mortgage ‘**No**’ customers- Sale: 1841 (17.90%) and No Sale: 8443 (82.10%). Home Mortgage ‘**Unknown**’ customers- Sale: 99 (17.37%) and No Sale: 471 (82.63%). Home Mortgage ‘**Yes**’ customers- Sale: 2323 (19.22%) and No Sale: 9763 (80.78%).

Personal Loan with Sale Outcome



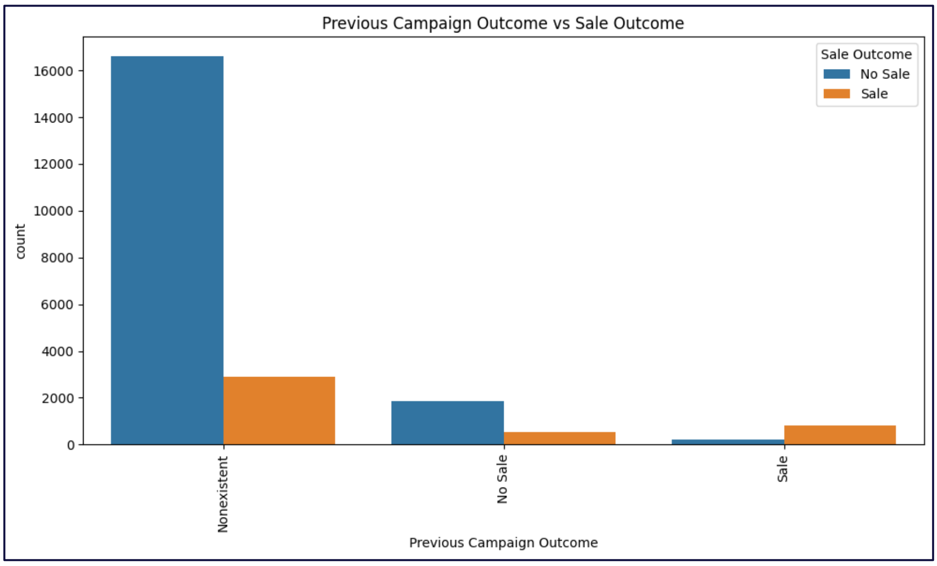
Personal Loan ‘**No**’ customers- Sale: 3547 (18.79%) and No Sale: 15330 (81.21%). Personal Loan ‘**Unknown**’ customers- Sale: 99 (17.37%) and No Sale: 471 (82.63%). Personal Loan ‘**Yes**’ customers- Sale: 617 (17.66%) and No Sale: 2876 (82.34%).

Last Contact Direction with Sale Outcome



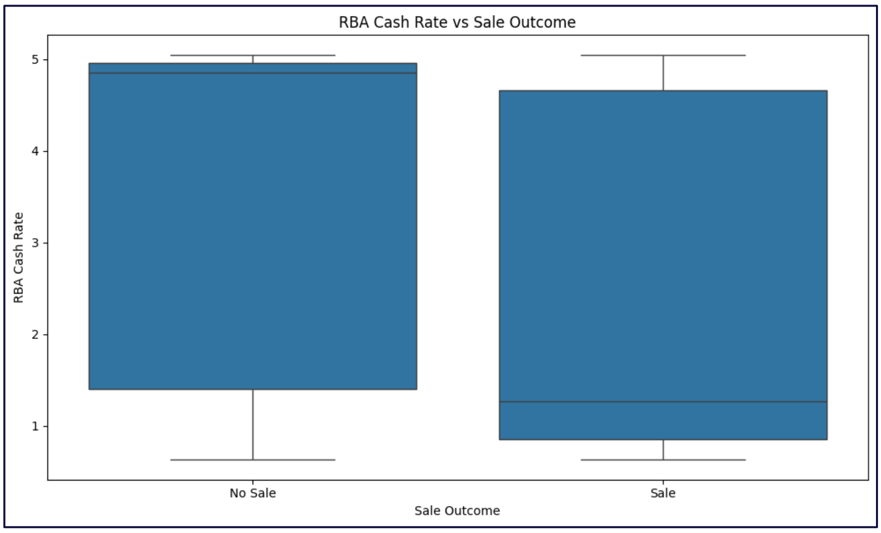
Last Contact Direction ‘**Inbound**’ customers- Sale: 3540 (23.75%) and No Sale: 11364 (76.25%). Last Contact Direction ‘**Outbound**’ customers- Sale: 723 (8.99%) and No Sale: 7313 (91.01%).

Previous Campaign Outcome with Sale Outcome



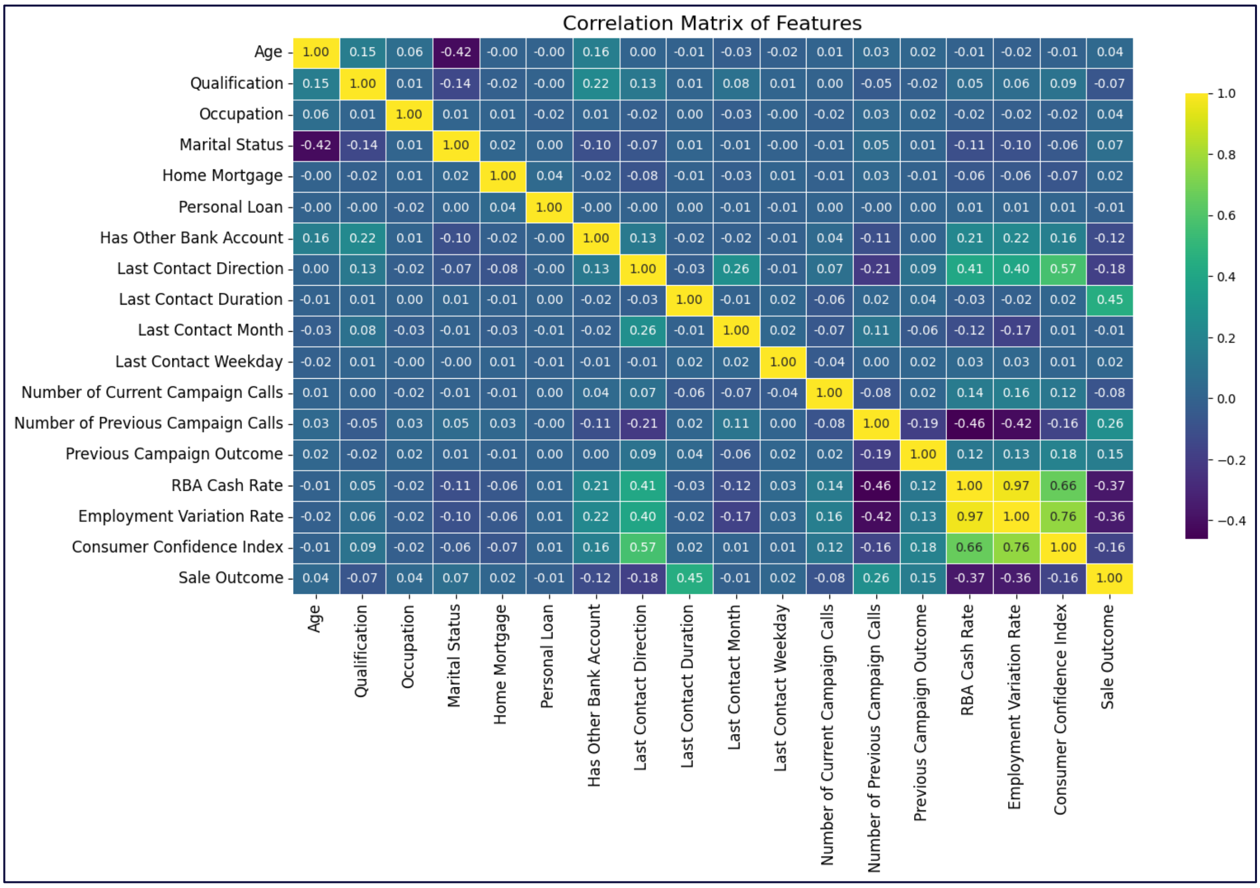
Previous Campaign Outcome ‘**No Sale**’ customers- Sale: 542 (22.76%) and No Sale: 1839 (77.24%). Previous Campaign Outcome ‘**Nonexistent**’ customers- Sale: 2900 (14.87%) and No Sale: 16605 (85.13%). Previous Campaign Outcome ‘**Sale**’ customers- Sale: 821 (77.90%) and No Sale: 233 (22.10%).

RBA Cash Rate with Sale Outcome



Mean RBA Cash Rate for ‘No Sale’ group (3.81) is significantly higher than ‘Sale’ group (2.12). The median also depicts quite a difference: 4.857 for ‘No Sale’ vs. 1.266 for ‘Sale’. The standard deviation for ‘Sale’ (1.75) is slightly higher than ‘No Sale’ (1.64) depicting a wider spread of cash rates among ‘Sale’ group. The IQR for ‘No Sale’: 3.557 and for ‘Sale’: 3.814.

Multivariate Analysis



Moderate positive correlation between Number of Previous Campaign Calls and Sale Outcome depicts that follow-up calls are beneficial. High positive correlation between RBA Cash Rate and Last Contact Duration (0.97) amplifies their individual positive impacts on Sale Outcome (customers’ willingness to engage in longer conversations is influenced by their financial circumstances).

Feature Selection and Data Split

We have set Sale Outcome as the label. Relevant variables influencing the label: Age, Qualification, Occupation, Marital Status, Home Mortgage, Personal Loan, Has Other Bank Account, Last Contact Direction, Last Contact Duration, Last Contact Month, Last Contact Weekday, Number of Current Campaign Calls, Number of Previous Campaign Calls, Previous Campaign Outcome, RBA Cash Rate, Employment Variation Rate, Consumer Confidence Index.

We have split the data: 70% for training and 30% for testing. Random state= 2024 to maintain the same training and testing sets through different executions important for debugging, reproducibility, and comparing the results.

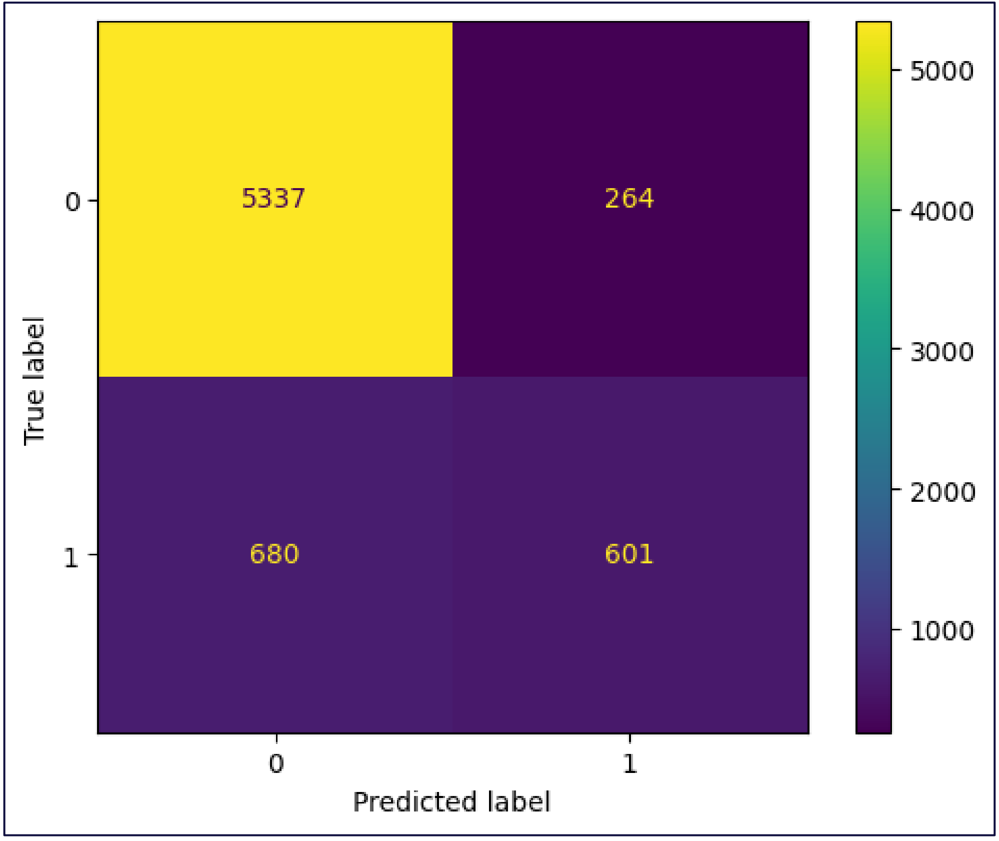
Model Development and Evaluation

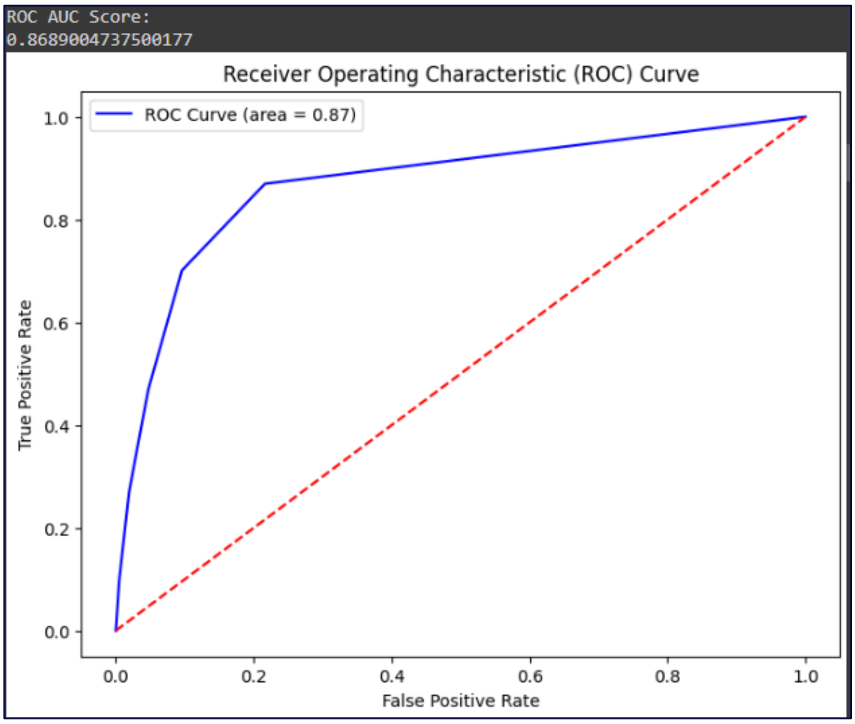
Supervised Machine Learning

1. kNN Classifier: k=5.

Accuracy 0.86:

	Precision	**Recall**	**f1-score**
Class 0:	0.89	0.95	0.92
Class 1:	0.69	0.47	0.56

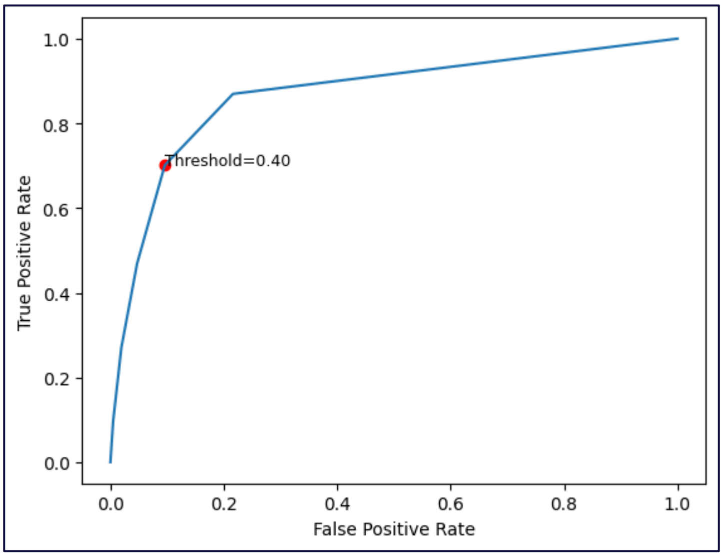




```
ROC AUC Score:
0.8689004737500177
```

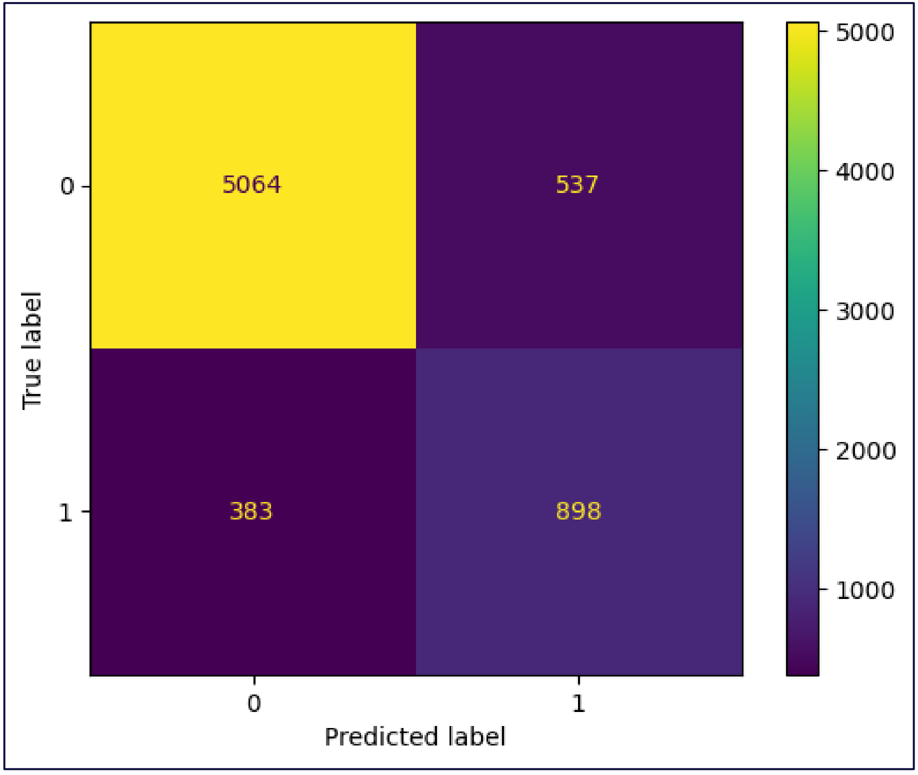
After finding best threshold based on Accuracy and F1 Score = 0.4:

AUC=0.87



Accuracy 0.87:

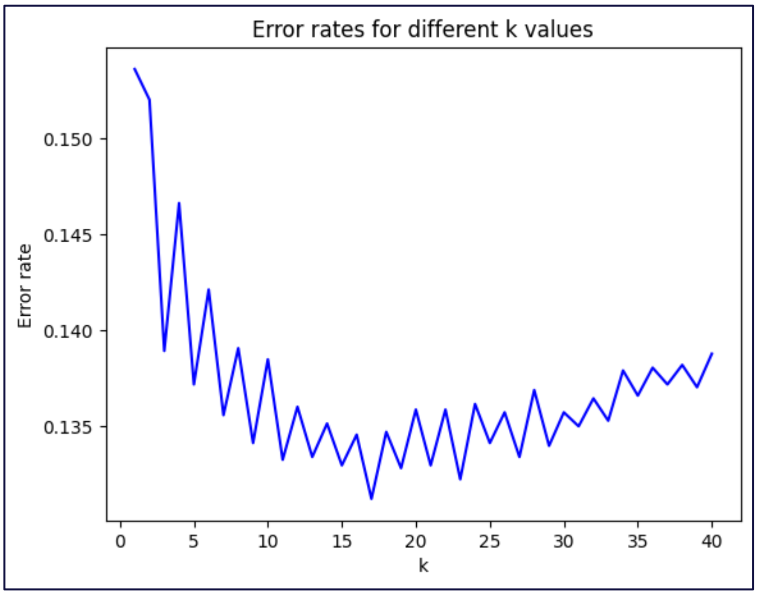
	Precision	**Recall**	**f1-score**
Class 0:	0.93	0.90	0.92
Class 1:	0.63	0.70	0.66



Optimising k

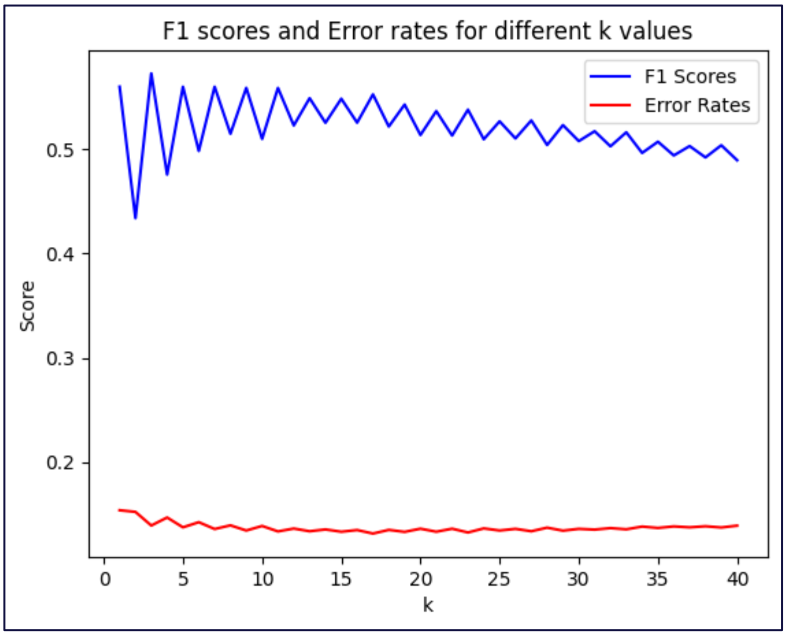
1. On Accuracy:

Best k= 17, best accuracy= 0.869, F1 score for best accuracy= 0.553.



2. On F1 Score:

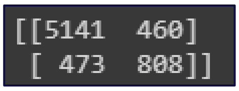
Best k= 3, best F1 score= 0.573, Accuracy for best F1 score= 0.861.



2. Decision Tree Classifier:

Base DT:

Accuracy=0.86

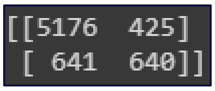


```
[[5141  460]
 [ 473  808]]
```

	Precision	**Recall**	**f1-score**
Class 0:	0.92	0.92	0.92
Class 1:	0.64	0.63	0.63

Pre-Prune DT:

Accuracy=0.85

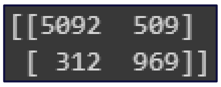


```
[[5176  425]
 [ 641  640]]
```

	Precision	**Recall**	**f1-score**
Class 0:	0.89	0.92	0.91
Class 1:	0.60	0.50	0.55

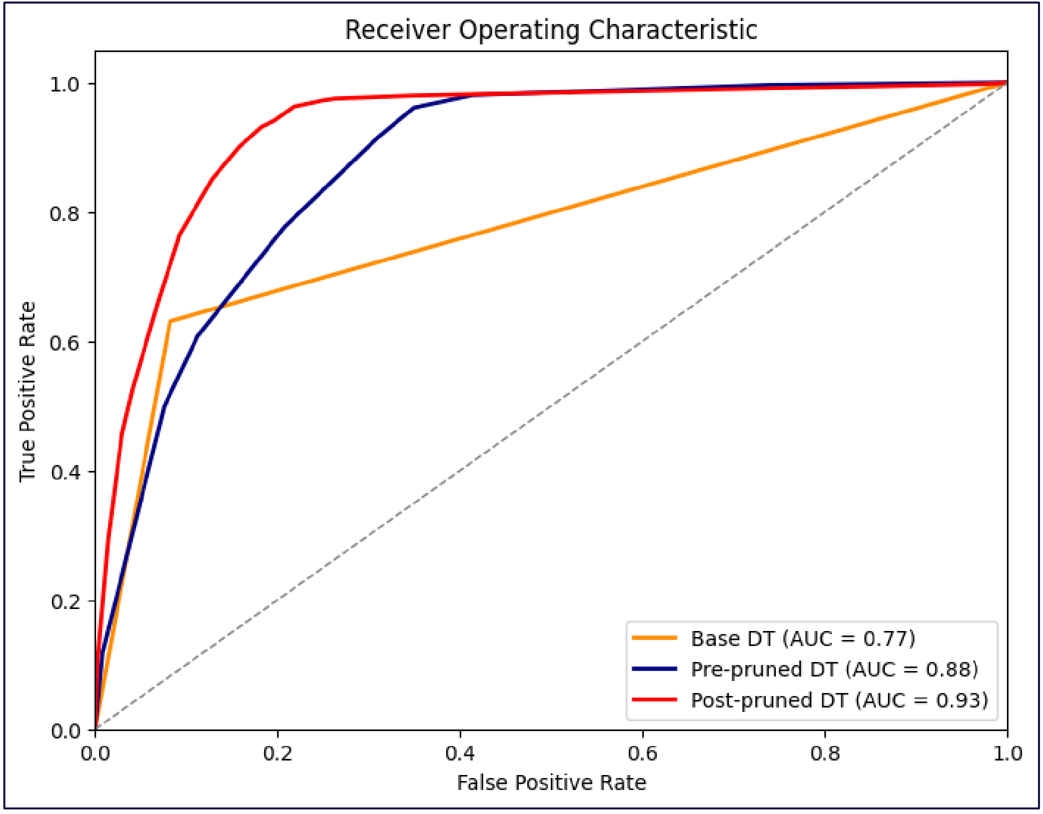
Post-Prune DT:

Accuracy=0.88



```
[[5092  509]
 [ 312  969]]
```

	Precision	**Recall**	**f1-score**
Class 0:	0.94	0.91	0.93
Class 1:	0.66	0.76	0.70

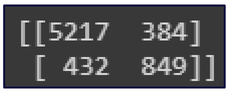


Base DT AUC: 0.77, Pre-pruned DT AUC: 0.88, Post-pruned DT AUC: 0.93.

Ensemble Learning

Using Bootstrap Aggregation:

Accuracy=0.88



```
[[5217  384]
 [ 432  849]]
```

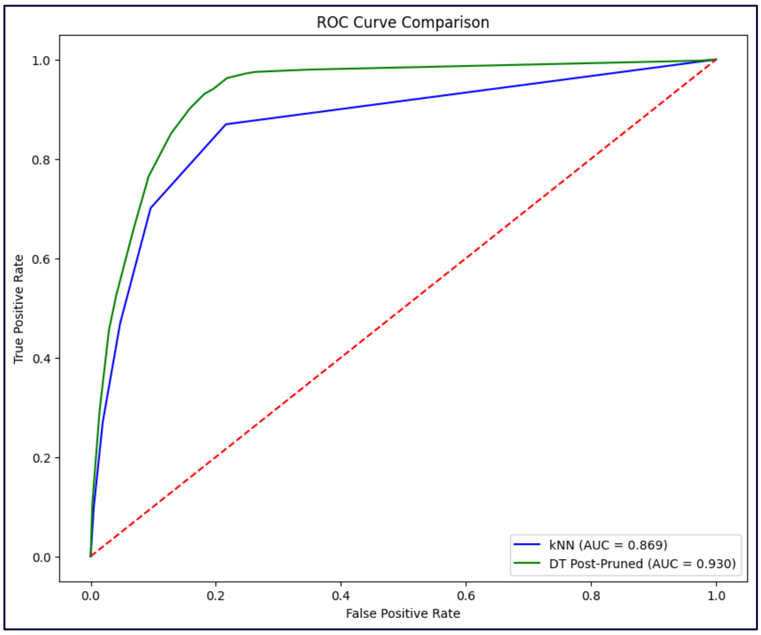
	Precision	**Recall**	**f1-score**
Class 0:	0.92	0.93	0.93
Class 1:	0.69	0.66	0.68

Model Comparison (Between best 2): kNN (Best threshold) and DT (Post-Prune)

kNN- TN: 5064, FP: 537, FN: 383, TP: 898.

DT- TN: 5092, FP: 509, FN: 312, TP: 969.

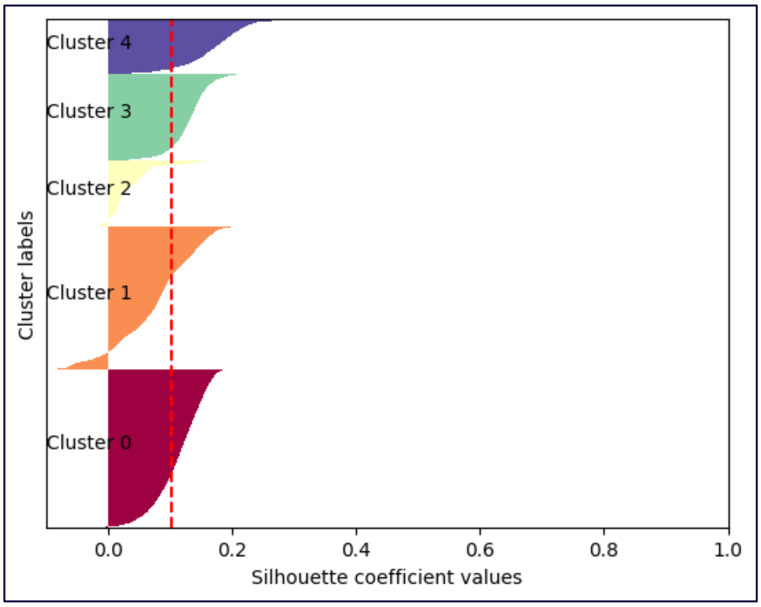
Post-Pruned Decision Tree has higher accuracy, precision, recall, and F1-score, AUC.



Unsupervised Machine Learning

k-Means clustering: k=5.

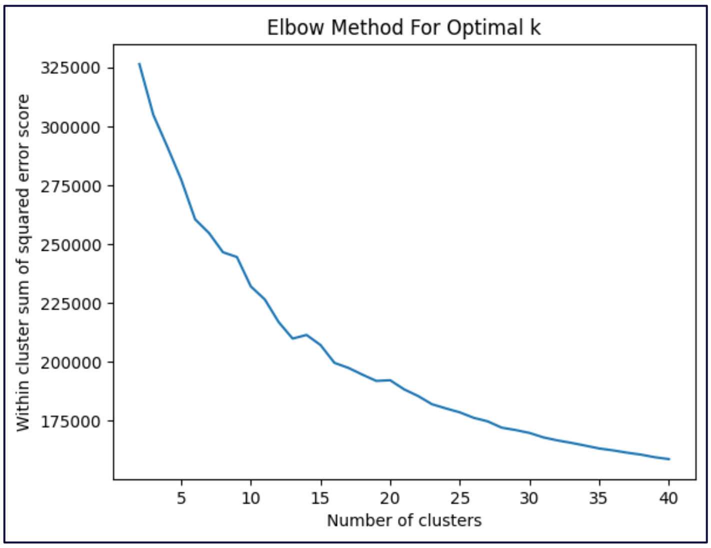
Within-Cluster Sum of Squares: 277358.311, Davies-Bouldin Index: 2.681, Silhouette score: 0.103



Optimising k

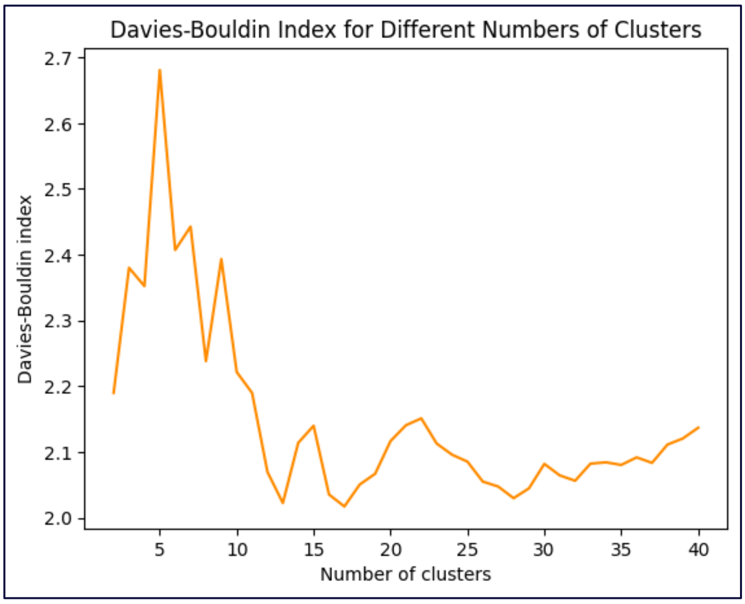
1. Elbow method based on the WCSS Score:

Best k= 2, Best WCSS error score: 4622.613



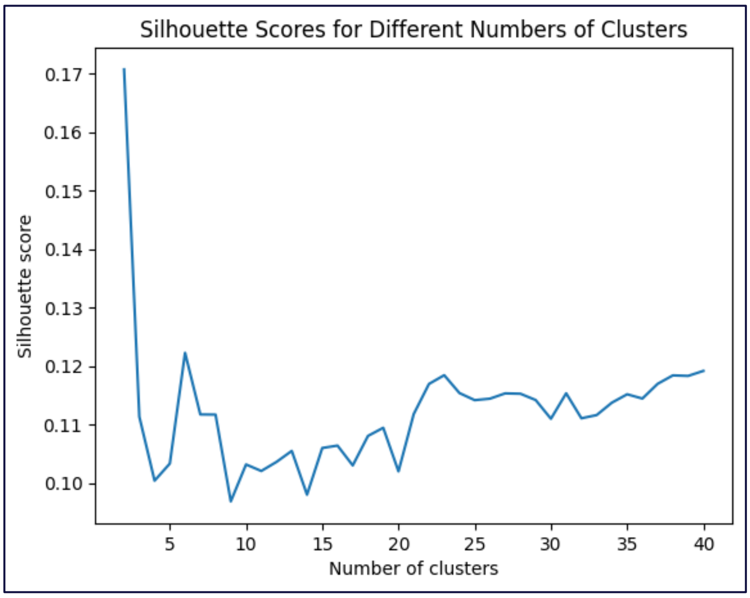
2. Based on Davies-Bouldin Score:

Best k=3, Best Davies-Bouldin Index: 1.000

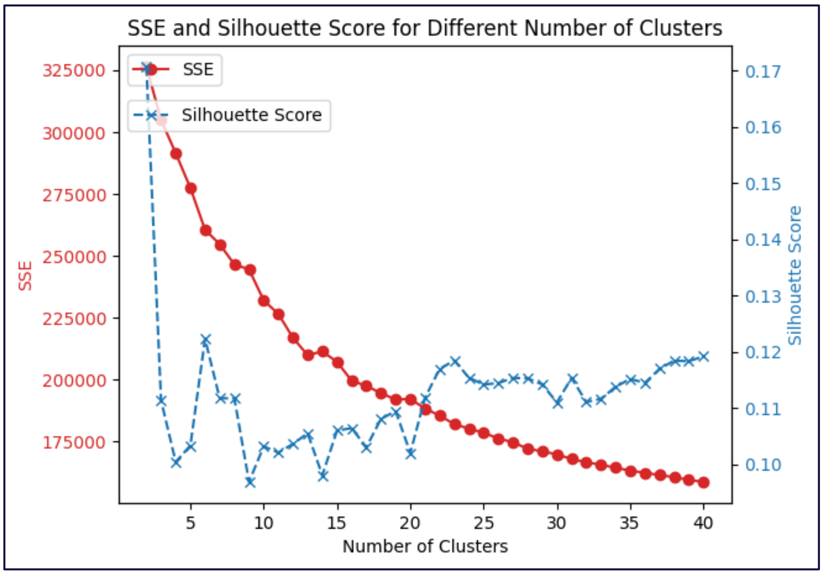


3. Based on Silhouette Score:

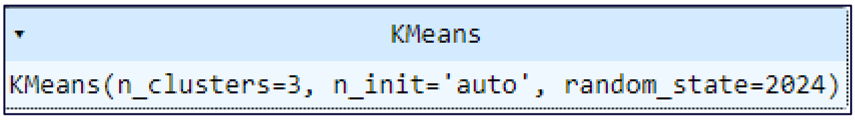
Best k=2, Best Silhouette Score: 0.171



SSE and Silhouette Score:

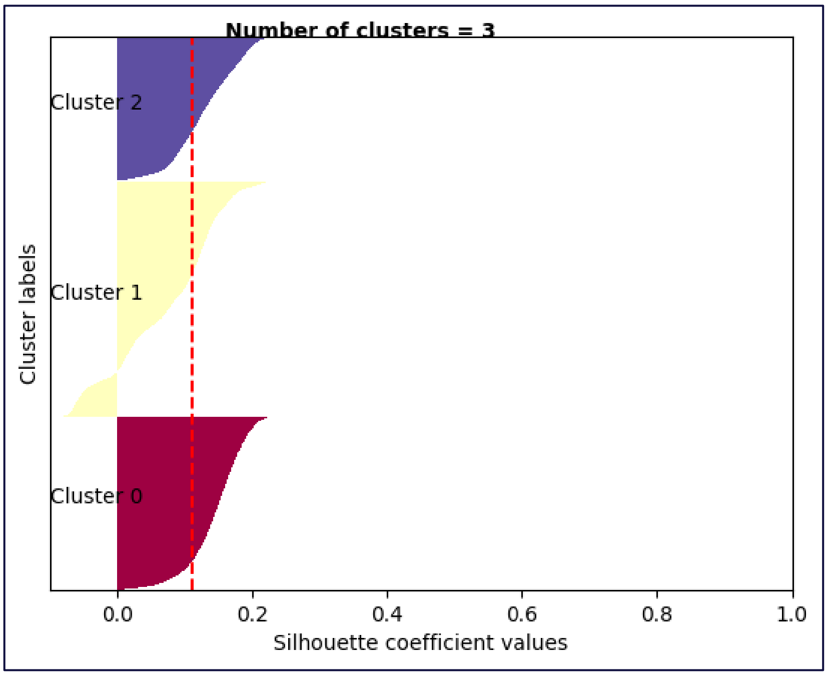


Model Selection: Choosing **k=3** for clustering.

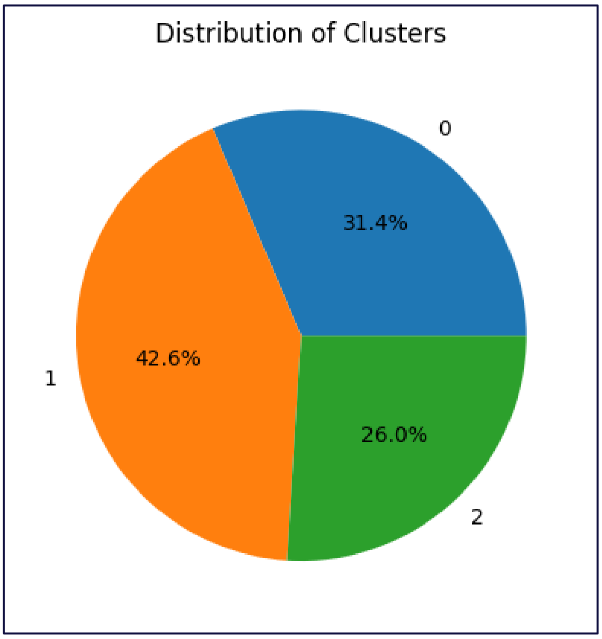


```
KMeans(n_clusters=3, n_init='auto', random_state=2024)
```

Within-Cluster Sum of Squares: 304826.439, Davies-Bouldin Index: 2.380, Silhouette score: 0.111.



Counting Clusters: 0= 7210, 1= 9766, 2= 5964



Even though k=5 has lower WCSS indicating tighter clusters, the difference might not be big enough if other metrics are significantly better for k=3. Hence, we will use k=3 model for clustering; due to **better** Silhouette Score and DBI **suggesting** k=3 forms more distinct and better-defined clusters than k=5.

Solution Recommendation

kNN: Accuracy: 0.866, Precision: 0.626, Recall: 0.701, F1: 0.661.

(10-fold Cross-Validation): Accuracy scores: 0.845 (+/- 0.002). F1 scores: 0.541 (+/- 0.006).

Decision Tree: Accuracy: 0.88, Precision: 0.66, Recall: 0.76, F1 score: 0.70.

(K-fold Cross-Validation): Accuracy scores: 0.881 (+/- 0.005). F1 scores: 0.696 (+/- 0.014).

We are **recommending** Post-Pruned Decision Tree due to higher accuracy, precision, recall, F1-score, and AUC.

Pros: It is easy to understand and interpret, has good visual representation, is non-parametric, needs less effort for data preparation (no normalisation), can work for both numerical and categorical targets, and can handle non-linear relationships.

Cons: Depending on its implementation, data preparation is needed (numerical or categorical), has high probability of overfitting, is sensitive to training datasets and small variations leading to overfitting and instability respectively, has loss of information when predicting continuous targets because of the discretisation of continuous variables, and becomes complex when having multi-class targets.

Future engagements will aid Great Ocean Bank in leveraging the insights gained from data analysis and performance of post-pruned decision tree model.

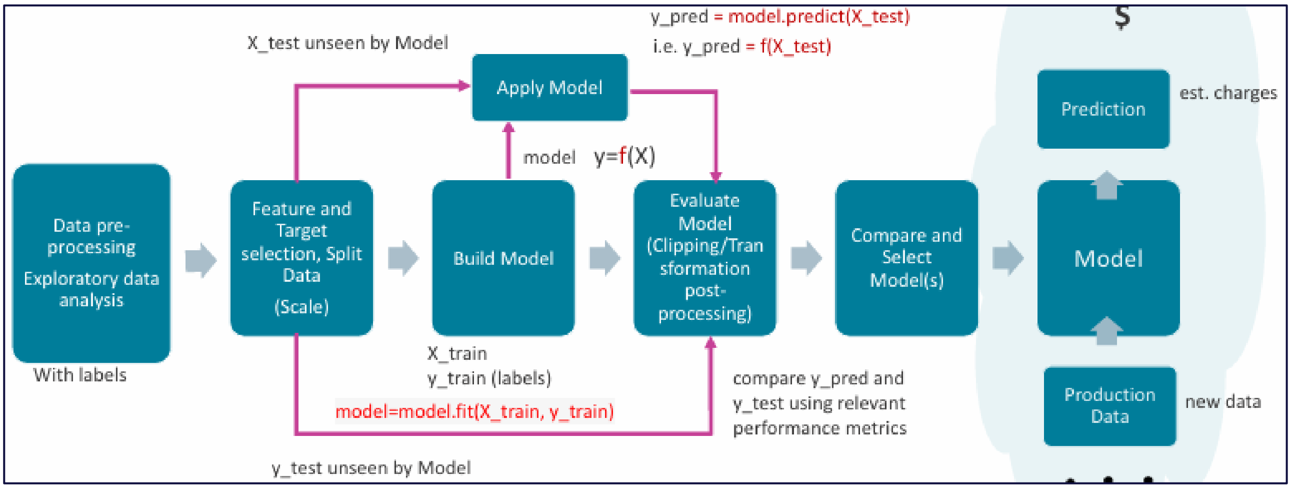
Technical Recommendations

<u>Software Libraries</u>: **Pandas**: Data manipulation and analysis. **NumPy**: Numerical computations and array operations. **Seaborn** and **Matplotlib**: Data visualisation. **Scikit-learn**: Machine Learning algorithms and model evaluation.

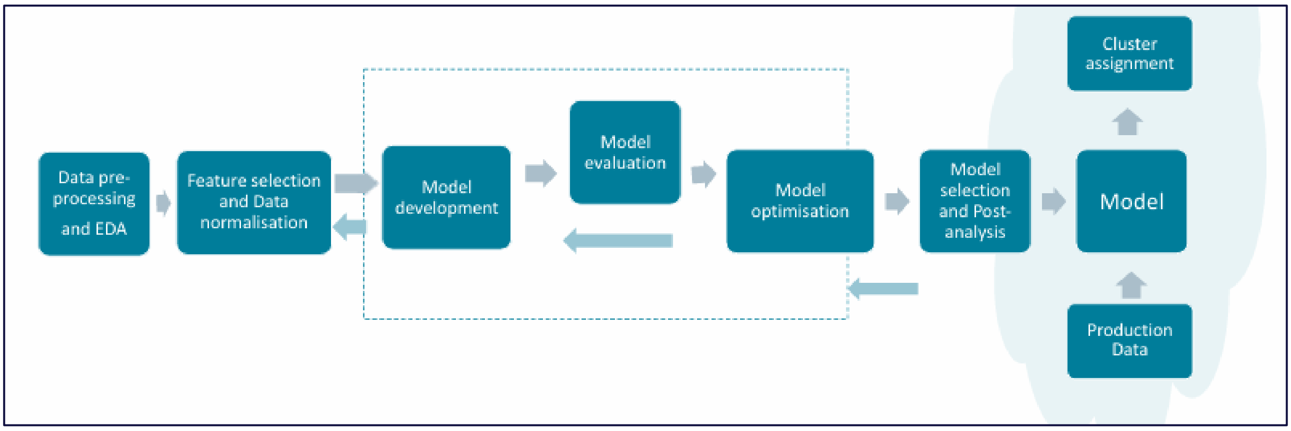
<u>Programming Language</u>: Python.

<u>Computing Environment</u>: Google Colab and Local Development.

<u>Supervised ML Diagram</u>:


y_pred = model.predict(X_test)
i.e. y_pred = f(X_test)
$
X_test unseen by Model
Apply Model
Prediction
est. charges
model
y=f(X)
Data pre-processing
Exploratory data analysis
Feature and Target selection, Split Data
(Scale)
Build Model
Evaluate Model (Clipping/Transformation post-processing)
Compare and Select Model(s)
Model
With labels
X_train
y_train (labels)
model=model.fit(X_train, y_train)
compare y_pred and y_test using relevant performance metrics
Production Data
new data
y_test unseen by Model

<u>Clustering Diagram</u>:


Cluster assignment
Data pre-processing and EDA
Feature selection and Data normalisation
Model development
Model evaluation
Model optimisation
Model selection and Post-analysis
Model
Production Data

For maintaining Accuracy and Relevance Over Time do regular model retraining (e.g. monthly/quarterly), performance monitoring (using tools, alerts, and threshold), data quality management (through data validation and handling drift data), stay in the feedback loop to make iterative improvements, exploration through feature engineering/model experimentation, document and share knowledge.

References

1. scikit-learn . (2019). *sklearn.neighbors.KNeighborsClassifier — scikit-learn 0.22.1 documentation*. Scikit-Learn.org. https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

2. *What is the k-nearest neighbors algorithm? | IBM*. (2024, April 22). Www.ibm.com. https://www.ibm.com/topics/knn#:~:text=The%20k%2Dnearest%20neighbors%20(KNN

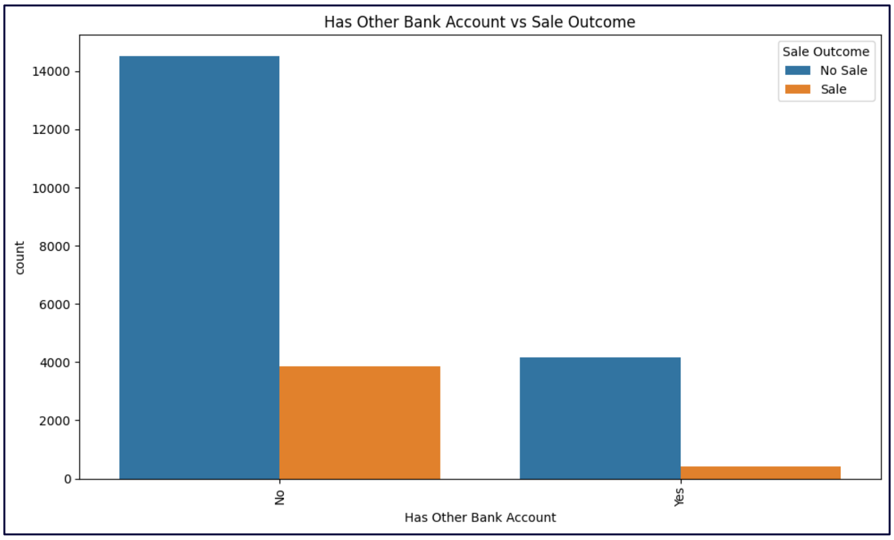
3. GeeksforGeeks. (2018, November 13). K-Nearest Neighbours - GeeksforGeeks. GeeksforGeeks. https://www.geeksforgeeks.org/k-nearest-neighbours/

4. scikit-learn. (2009). *1.10. Decision Trees — scikit-learn 0.22 documentation*. Scikit-Learn.org. https://scikit-learn.org/stable/modules/tree.html

5. GeeksforGeeks. (2019, May 30). *K Means Clustering - Introduction - GeeksforGeeks*. GeeksforGeeks. https://www.geeksforgeeks.org/k-means-clustering-introduction/

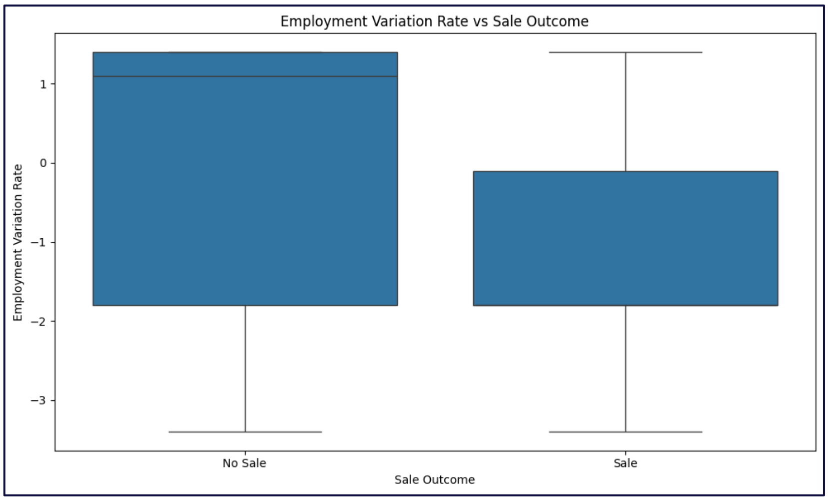
Multivariate EDA

Has Other Bank Account with Sale Outcome



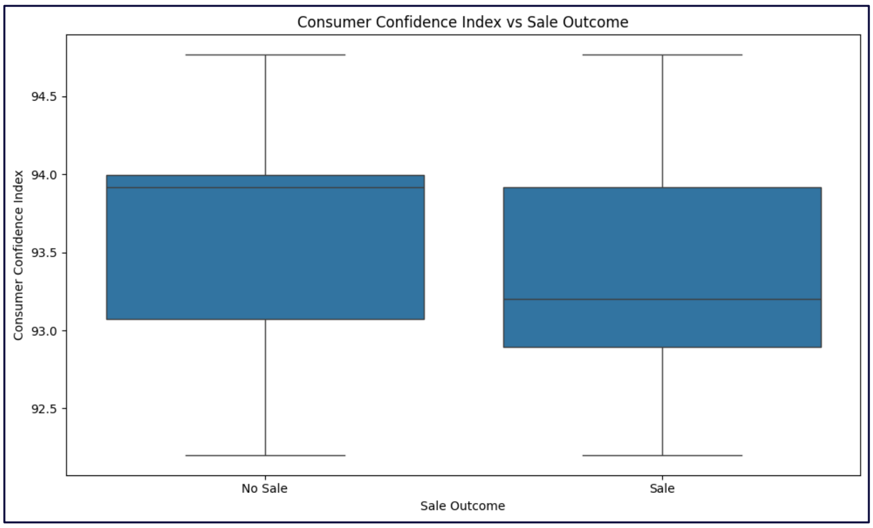
Has Other Bank Account ‘**No**’ customers- Sale: 3858 (20.99%) and No Sale: 14520 (79.01%). Has Other Bank Account ‘**Yes**’ customers- Sale: 405 (8.88%) and No Sale: 4157 (91.12%).

Employment Variation Rate with Sale Outcome



Mean Employment Variation Rate for ‘No Sale’ group (0.247) is significantly higher than ‘Sale’ group (-1.231). The median also shows quite a difference: 1.1 for ‘No Sale’ vs. -1.8 for ‘Sale’. Standard deviation for ‘Sale’ (1.63) is slightly higher than ‘No Sale’ (1.49) representing a wider spread of employment rates among ‘Sale’ group. The IQR for ‘No Sale’: 3.2 and for ‘Sale’: 1.7.

Consumer Confidence Index with Sale Outcome



Mean Consumer Confidence Index for ‘No Sale’ group (93.60) is slightly higher than ‘Sale’ group (93.35). The median is also slightly higher: 93.92 for ‘No Sale’ vs. 93.20 for ‘Sale’. Standard deviation for ‘Sale’ (0.68) is slightly higher than ‘No Sale’ (0.56) depicting a wider spread of confidence index values among ‘Sale’ group. The IQR for ‘No sale’: 0.91 and for ‘Sale’: 1.03.