TheAnalyticsTeam

Sprocket Central Pty Ltd

Data analytics approach

Prim Hansakul

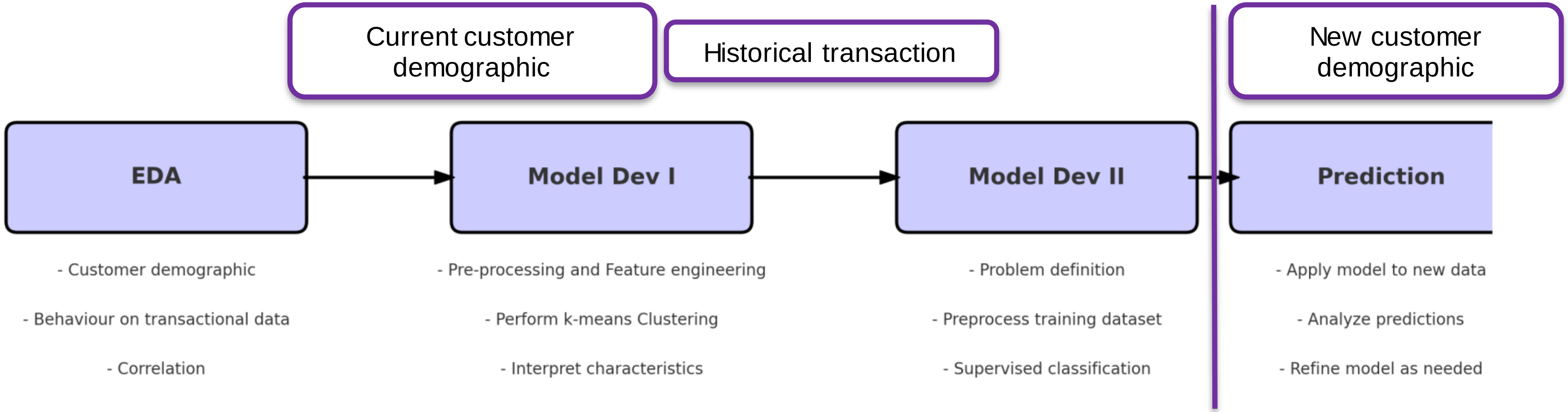
Agenda

1. Introduction
2. Data Exploration
3. Model Development
4. Interpretation

1.Introduction

After identifying data quality issues in three essential datasets related to current customers, the subsequent task involves a statistical analysis and data modeling.

The **objective is to discern target customers from a new cohort of 1,000, distinguished by their lack of historical transactional data.** This step is pivotal for ensuring data integrity and for the strategic identification of key new customers to engage with moving forward.

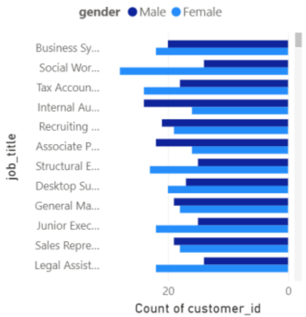


2.Data Exploration

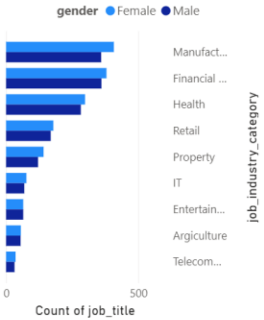
2.1 Data Exploration (EDA) on customer demographic

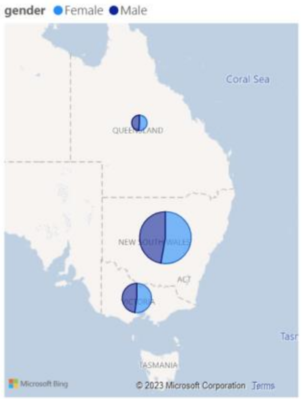
Gender: 52.2% Female, 47.7% Male
Age: Majority 40-50 years
Wealth: Mass market dominant, high net worth and affluent less common
Location: NSW > VIC > QLD
Property Value: 9> 8> 10> 7
Industries: predominantly in the industries of manufacturing, finance, health, retail, and property.

Count of customer_id by job_title and gender

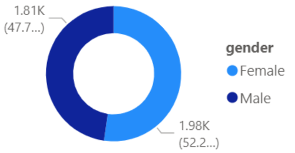


Count of job_title by job_industry_category and gender

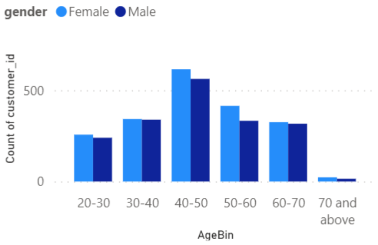




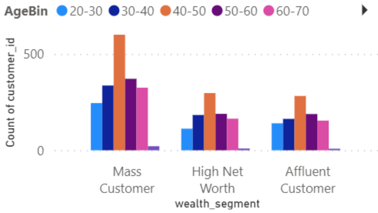
Count of customer_id by gender



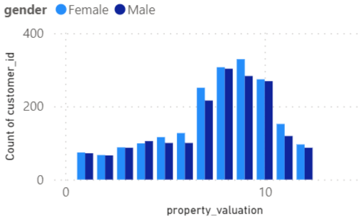
Count of customer_id by AgeBin and gender



Count of customer_id by wealth_segment and AgeBin



Count of customer_id by property_valuation and gender



2.Data Exploration

2.2.1 Data Exploration (EDA) on transaction dataset

Transactions: 19.45k total, 48.6% by women, 45.85% by men

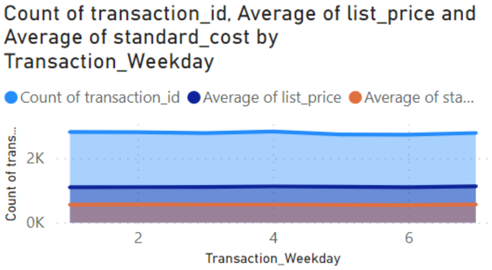
Tranaction Peak: Peak in October and August, followed by July, May, and January.

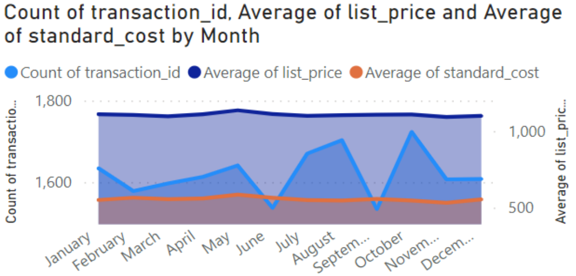
Online/offline: equally

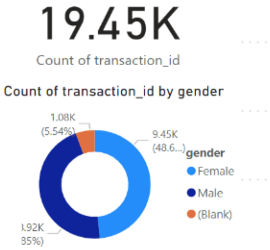
- online Tranaction high on Tuesdays/Fridays,
- offline high on Thursdays/Mondays

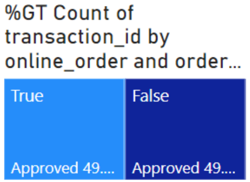
Seasonality:

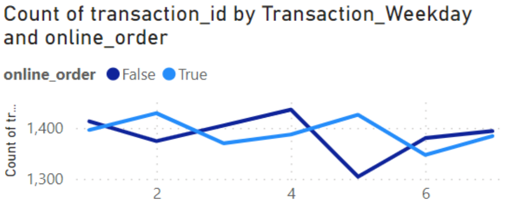
- online high on Aug, Oct, Jan
- offline high on July, Oct, March

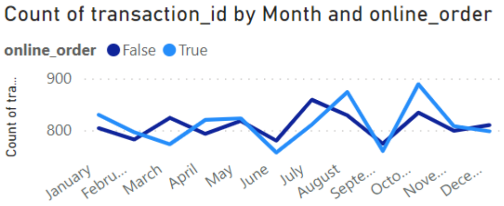








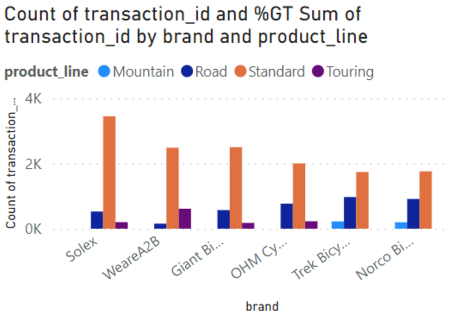


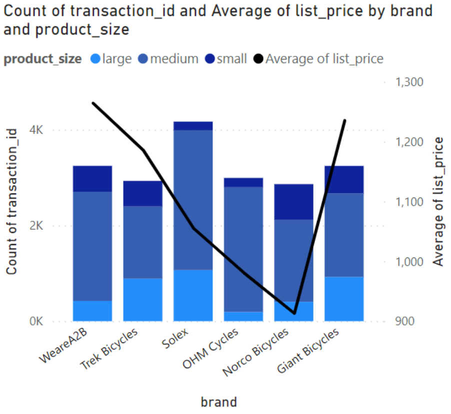


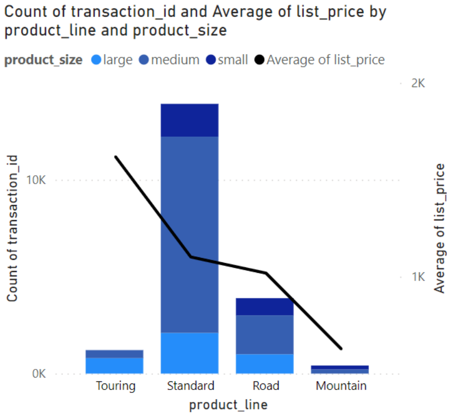
2.Data Exploration

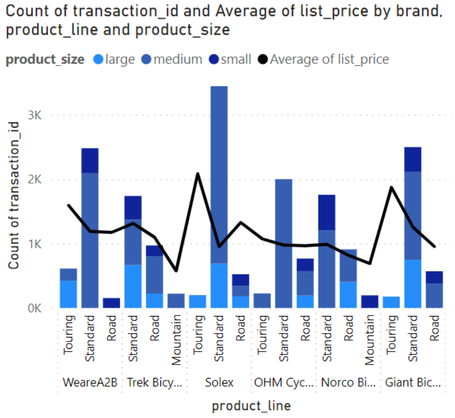
2.2.2 Data Exploration (EDA) on transaction dataset

Brands: Popular - Solex, Wearea2b, Giant, OHM; Least - Trek, Norco
Product Lines: Standard, Road, Touring; Least - Mountain
Sizes: Medium most popular, then large, small
Prices: Highest - WeareA2B; Lowest - Norco
Types: Expensive - Touring, then Road, Mountain









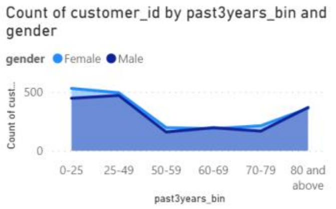
2.Data Exploration

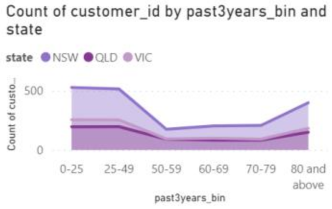
2.2.3 Data Exploration (EDA) on transaction dataset

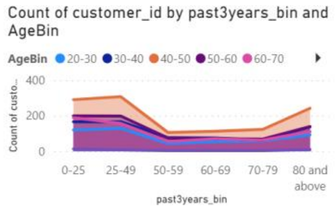
3-Year Purchase Patterns: Majority in 0-25 purchase-bin with more female purchases high, no significant difference on Age , State, Wealth

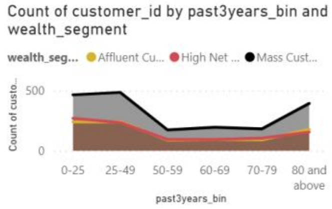
3-Year Purchase Correlations:

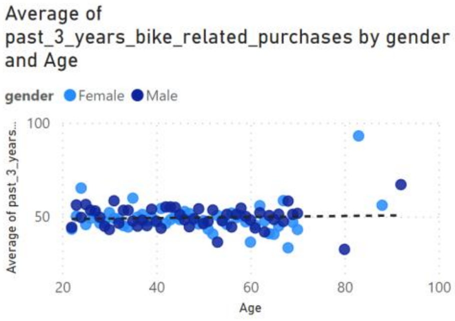
- Positive relationship with Age and Property Valuation. Outliers appear in 80+ age, low property value outliers among females
- Negative relationship with Tenure

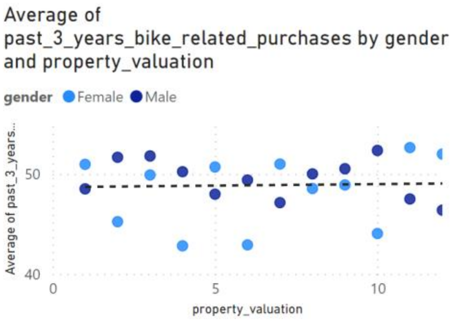


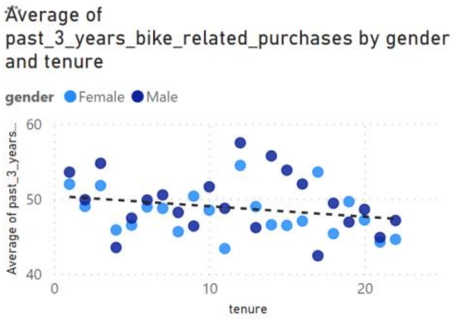






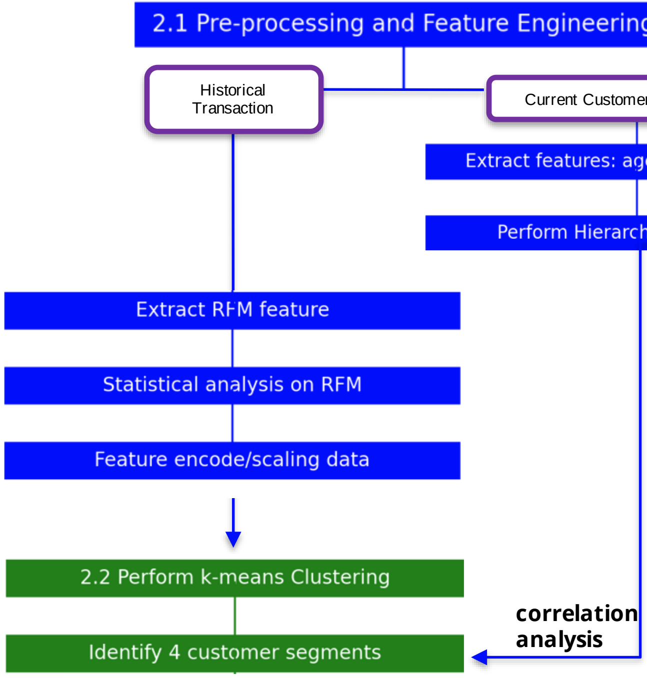






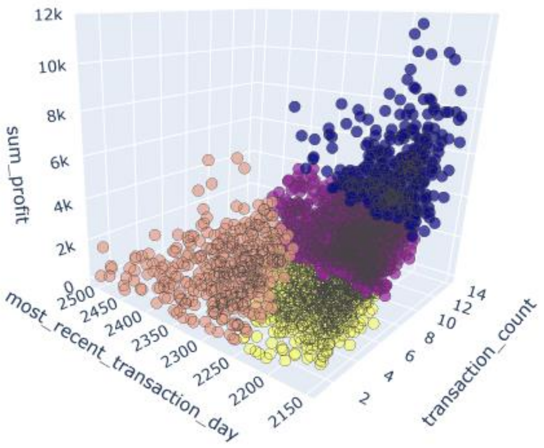
Note: The data and information in this document is reflective of a hypothetical situation and client. This document is to be used for KPMG Virtual Internship purposes only.

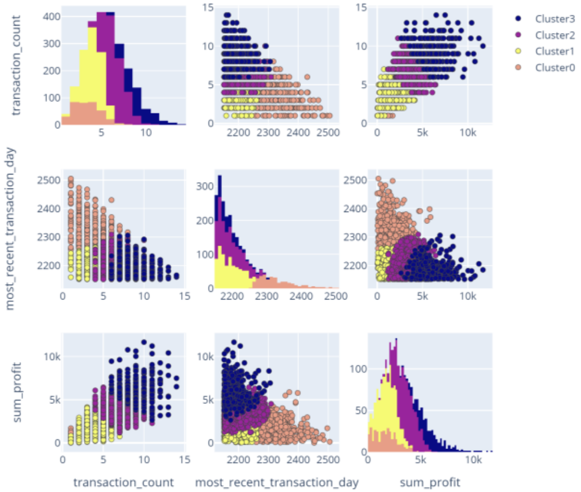
3.1 Model Development (I) - RFM Model / Cluster Analysis



The optimal model identifies **four customer behavior segments based on recency, frequency, and monetary** value of purchases:

- **Non-active Customers:** No recent purchases or activity.
- **Low-Value Customers:** New customers with either low profitability or recent but infrequent purchases.
- **Medium-Value Customers:** Customers with medium profit levels who purchase frequently.
- **High-Value Customers:** Customers generating high profits with the most recent and frequent purchases.





3.1 Model Development (I) - RFM Model / Cluster Analysis

2.3 Interpretation characteristics of the Target RFM cluster

Identify characteristics of high-value customers

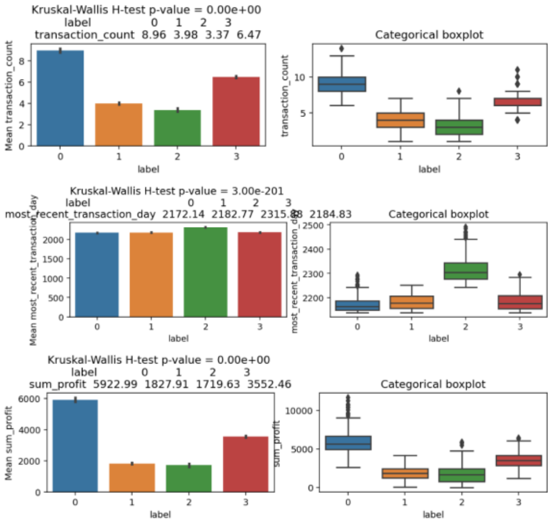
Low correlation with other features / No linear relationship

Statistical analysis and hypothesis testing

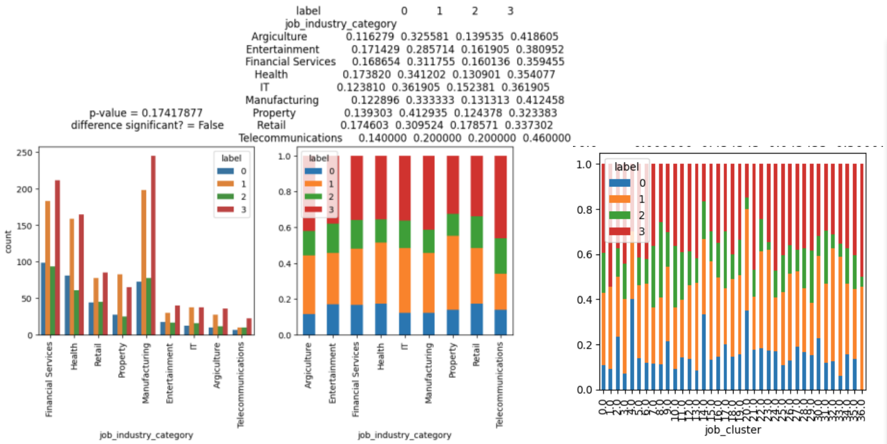
One-way ANOVA

Chi-squared test

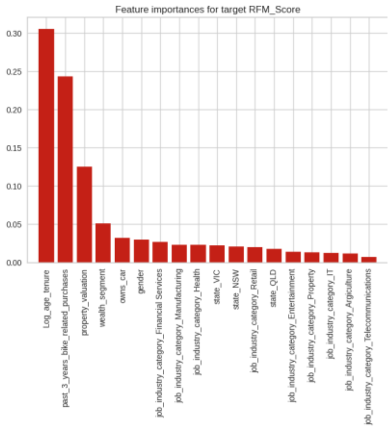
- The study aims to define characteristics of high-value customers but finds **low correlations between RFM clusters and other features**, suggesting **non-linear relationships**.
- Statistical analysis reveals significance only within RFM features. However, **visual data analysis shows distinct distributions in age, tenure, property valuation, job industry, and job title clusters.**



RFM Features



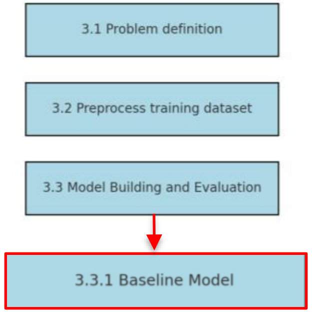
Job Features



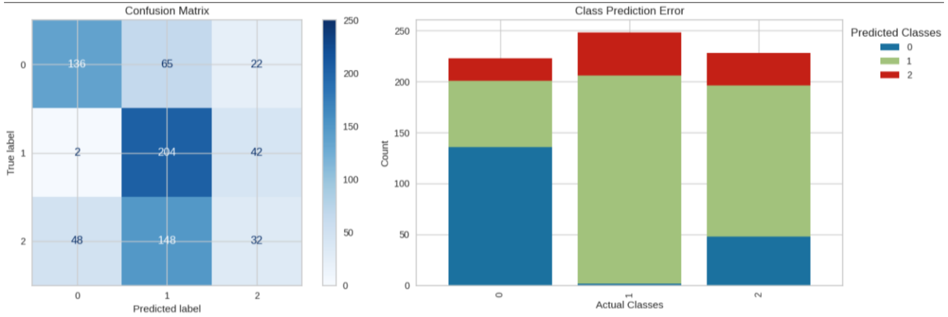
Features Ranks

3.2 Model Development (II) - Supervised learning classification

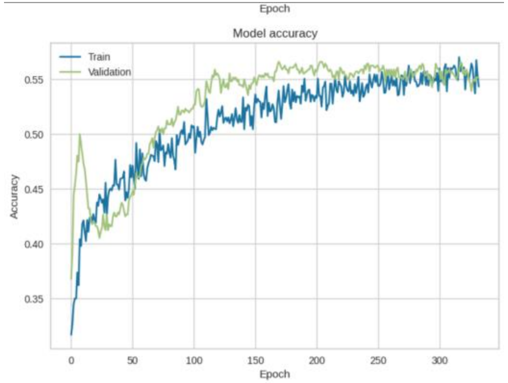
Baseline Model: Multiclass Classification

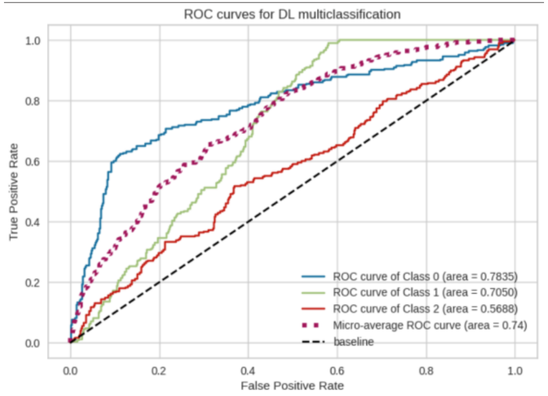


Facing challenges in identifying the underlying characteristics of the target RFM cluster, we implemented supervised learning algorithms to discern patterns and classify the target clusters.



Task	**Multiclass classification** to categorize customers into - non-active (0) - low /medium-value (1) - high-value (2)
Model	A neural network with automated hyperparameter tuning
Evaluation	The model's performance, assessed through loss and accuracy on training and validation set
Result	**Imbalanced accuracy across classes**, as indicated by the ROC/AUC curves and classification report. **Overall 0.74 AUC with 0.55 accuracy score**

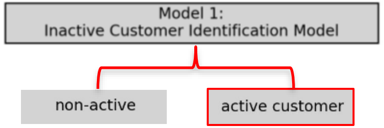




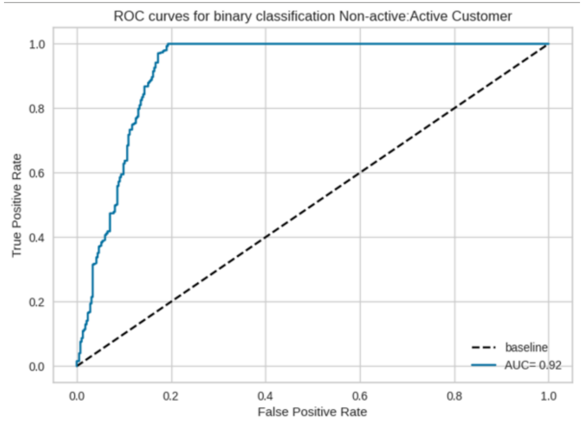
3.2 Model Development (II) - Supervised learning classification

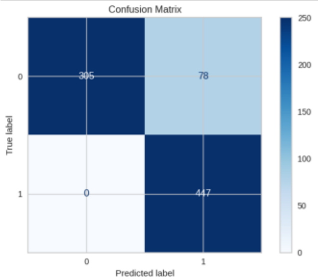
Two-Step Binary Classification

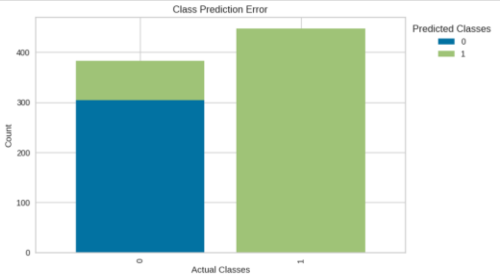
Instead of multi-classification task, we will breaking it down into two-step binary classification model, as there's a significant class imbalance or a lack of sufficient information in the dataset.



Task I	**Binary classification Model :** Inactive Customer Identification - non-active (0) - active customer (1)
Model	A neural network with automated hyperparameter tuning
Result	effectively differentiates between non-active and active customers, demonstrating **high accuracy (0.91) and AUC scores (0.92)**



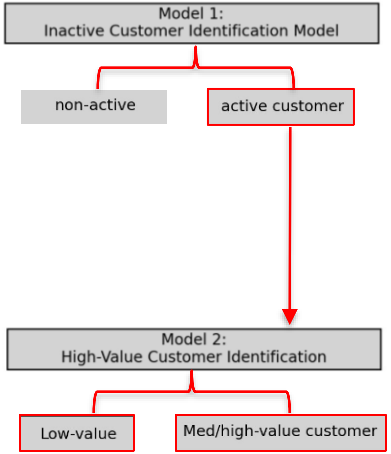




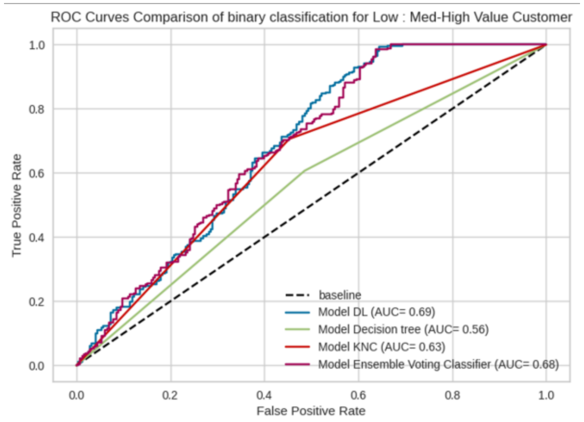
3.2 Model Development (II) - Supervised learning classification

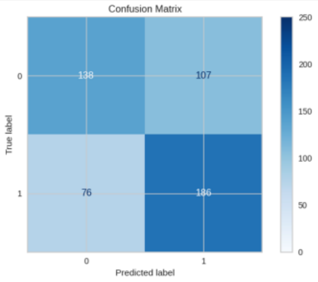
Two-Step Binary Classification

The resulted active customer from the first model are fed into the second training model to distinguish a level of customer value.



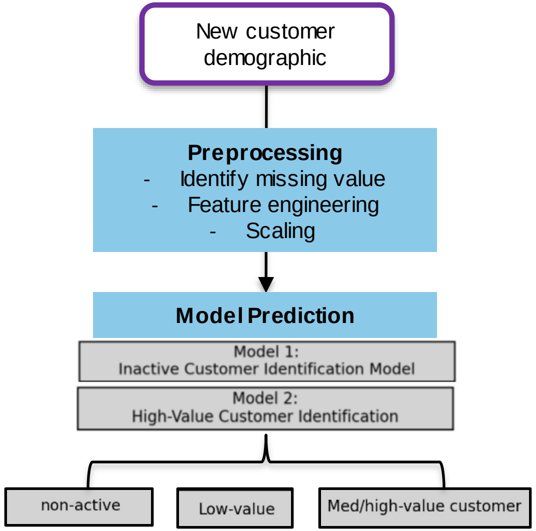
Task II	**Binary classification Model :** High-Value Customer Identification - low-value customer (0) - med/ high-value customer (1)
Model	deep learning model decision tree KNN ensemble voting classification
Result	The deep learning model outperformed baseline and others with an **AUC score of 0.69**

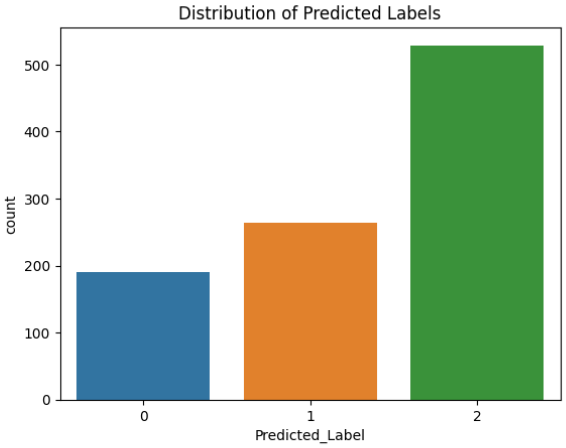




4. Prediction and Interpretation

Finally, we utilize the trained two-stage model to classify new customer dataset into 3 potential groups of non-active (0), low-value (1), med/high-value customers.





Key takeaway:

- This predicted segmentation can lead to more personalized marketing strategies, improved customer engagement.
- Targeted promotions and incentives resonate with each group's behavior and preferences can be used for potentially increasing customer lifetime value and loyalty.