Customer Analytics for Targeted Marketing

**1. Executive Summary**

For this project, I acted as a consultant analyst for a medium-sized bike & cycling accessories company on how to best optimize its marketing strategy with transactional and customer data.

With the provided datasets provided by Sprocket Central Pty Ltd, I managed to:

1. Identify and give recommendations on ways to clean the underlying data and mitigate these issues.
2. Develop Machine Learning models to target high value customers for the next marketing campaign, based on existing customer and transactions data.
3. Build a dashboard to present findings and model development.

I have built three machine learning models and found multinomial logistic regression to perform the best. Thus, this model was used to identify high value customers in the New Customer List.

These chosen customers are defined as those who will bring in the most profit, spend frequently, and visit the business often.

**2. Summary of Learning**

I learned how to translate a business problem into technical requirements in this project. I defined what a “high value” customer is, based on the Recency=Frequency-Monetary analysis in marketing, and used it to label the existing customer data. I also learned how to incorporate additional data such as median salary and latitude/longitude coordinates, which tell us more than just each customer’s postcode. A lot of missing values were handled by using regression imputation, which I will need to investigate more. I also handled high cardinality in job titles by grouping jobs into a few categories based on their common characteristics.

A few interesting findings were:

* 508 customers who never made a single purchase in the past year but have had multiple purchases in the past three years.
* The Mass Customer segment represents most of our customer base.
* The top three industries are: Manufacturing, Financial services, and Health.
* Although Manufacturing is the industry with the most bike related purchases, customers who hold technical positions (developers, programmers, engineers, etc.) are the ones buying the most.
* Mass Customers and customers that work in the Entertainment industry are less likely to own a car, so these two groups have potential for targeted marketing.

**3. Dataset**

*Describe briefly (1-2 paragraphs) the dataset you used in this project, including the data sources, how did you get the data (download the data or scrap data through web pages or API) and what the data is about.*

The data consists of four data sets: CustomerDemographic, CustomerAddress, Transactions, and NewCustomerList. The source is from KPMG Australia as they created the content to simulate their analysts’ work. The data can be found on [Kaggle](https://www.kaggle.com/datasets/emmanuelavogo/kpmg-virtual-internship-project-task-1/data).

CustomerDemographic set contains existing customer information such as name, age, gender, job, tenure, etc.

CustomerAddress set contains customers’ addresses, which were used to add more features.

Transactions set contains transactional data in the past year.

NewCustomerList is a list of new customers who have never had any transactional history with the business, and will be filtered for targeted marketing purposes.

**4. Exploratory Data Analysis**

A graph showing a number of different colored squares

Description automatically generated

The Count of Low-Value customers is high because out of 3993 customers (that are used for fitting models), there are 508 that do not have transaction data in the past year but have demographic and address data and have made purchases in the past three years.

However, including them will lead to a potential class imbalance problem. Thus, we will omit these customers from training.

A graph showing a distribution of target variable

Description automatically generated

This is the target distribution after removing those outlier customers.

A graph of a line graph

Description automatically generated

The 508 with missing 2017 transactions seems to be two populations of customers: those who have made less than average bike purchases, and those who have made more than the average.

We recommend checking in with the existing customers in the second population to see if they are still interested in our products.

**A graph of a number of groups

Description automatically generated**

Our customer base is mostly made up of middle-aged customers, with the 45-54 age group being the largest.

A graph of a graph showing the state

Description automatically generated with medium confidence

New South Wales customers love our bike products.

**A graph of a number of different types of data

Description automatically generated with medium confidence**

Our target variable depends on three variables:

* The **recency score** is determined by sorting the values of the most recent transaction date in ascending order and then grouping these values into four bins. The bin with the oldest dates is assigned a recency score of 1, and so on.
* The **frequency score** is determined by sorting the values of the number of transactions in ascending order and then grouping these values into four bins. The bin with the smallest number of transactions is assigned a frequency score of 1, and so on.
* The **monetary score** is determined by sorting the values of the total amount of the transaction in ascending order and then grouping these values into four bins. The bin with the smallest amount is assigned a monetary score of 1, and so on.

A graph of a graph

Description automatically generated with medium confidence

Tenure is the number of years the customer has been with the company. Its distribution indicates there are three groups of customers, which each may have a different target value.

The median salary is roughly normally distributed, with the mean around 50,000 AUD.

A graph of different colored bars

Description automatically generated

Our customer base is primarily Mass Customer segment.

Let's see if this majority are also the ones who bring in the most profit.

A graph of a number of people

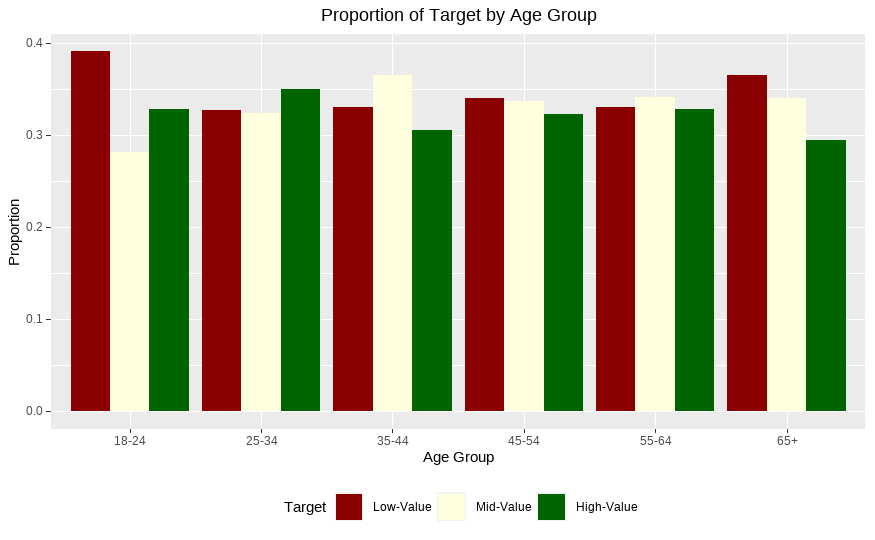
Description automatically generated with medium confidence

Across age groups, the Mass Customer segment brings in the most profit, except in the 25-34 age group, where the Affluent Customer segment takes the lead. We recommend looking into this 25-34 age group to see if we have maximized our effort for this group.

**A graph of a distribution of a group

Description automatically generated with medium confidence**

The Mass Customers in the 18-24 age group appear to spend more widely than in the other wealth segments. However, they are also the smallest age group.

****

Across all age groups, 18-24 year olds are the most likely to be low-value customers.

**A graph of a bicycle purchase

Description automatically generated with medium confidence**

The top 3 industries are Manufacturing, Financial Services, and Health.

**A graph of a graph showing a number of people

Description automatically generated with medium confidence**

However, customers with technical professions (programmers, developer, engineers, etc.) have made more purchases than any other job types despite the IT industry not being in the top 3.

**A map of australia with different colored dots

Description automatically generated**

Rather unsurprisingly, our customer base consists of three clusters: Melbourne, Sydney, and Brisbane.

**A chart of a graph

Description automatically generated with medium confidence**

The Health industry has the most proportion of High-Value customers. Meanwhile, Financial Services has the most proportion of Low-Value customers.

**A screenshot of a computer screen

Description automatically generated**

It makes sense that median salary and property valuation are correlated. We will omit the property valuation variable since both give similar information.

**5. Model Performance**

*Or you can report the models and results here; and include your* ***rmd or r*** *file in the final submission.*

*You can choose to build supervised models or unsupervised models.* ***Build at least three different models****. For each model you built, add several sentences to describe what you find, how to evaluate the model and what is the model performance.*

Three models were built: Multinomial Logistic Regression, Random Forest, K-nearest Neighbors.

The results are below:

**A screenshot of a computer

Description automatically generated**

Multinomial Logistic Regression’s performance

**A screenshot of a computer

Description automatically generated**

K-nearest Neighbors’ performance.

A screenshot of a computer

Description automatically generated

Random Forest’s performance.

**7. Customer Predictions**

**A screenshot of a computer

Description automatically generated**

Here is a subset of the high-value customers for targeted marketing.

**A chart with red green and yellow squares

Description automatically generated**

**6. Limitations**

* Interaction terms were not considered.
* Only the customers in CustomerDemographic were included for training.
* Model accuracies were poor (~35%)
  + - This is possibly due to there not being many differences in the data by Target.
    - Perhaps there’s no clear demographic pattern by customer value segment.
    - Or the New Customers have some different characteristics not accounted for.
    - Alternatively, I could also try only removal (ie. no imputation) of all observations with missing values and train the models to see if performance improves.
* Given more time, I would bring in more Australian Bureau of Statistics data at different geographic levels and create more external features for the model. I would also come up with a set of hypotheses for statistical testing.