Customer Analytics for Targeted Marketing

**1. Executive Summary**

*Describe briefly (1-2 paragraphs) what you have accomplished in this project: the research question and the conclusions.*

Advising a medium size bikes & cycling accessories company how to best optimize its marketing strategy with transactional and customer data.

Business question: how to target high value customers for next marketing campaign, based on customer and transactions data.

**2. Summary of Learning**

*Describe briefly (1-2 paragraphs) you have learned in this project– for example, challenges, interesting findings, new approaches and how to learn by yourself.*

Using business question as context to guide the project

Incorporating additional data to existing data

Defining a meaningful target variable

Customer segmentation / classification

Choosing the “best” model based on business case and assumptions

Standardizing date columns across data sets

Choosing imputation or removal

EDA before or after data split

High cardinality of categorical variable

Incorporating more data using address and zipcode

**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.*

**4. Exploratory Data Analysis**

*Describe the methods you have used for exploratory analysis of the data. You may use Rmarkdown file to record you exploratory data analysis process, \*knit\* to generate a pdf file named ‘ExploratoryAnalysis’ and include the* ***rmd and pdf*** *file in the final submission.*

*Or you can put your visualization results here and include your* ***rmd or r*** *file in the final submission. Do not put code in this report.*

*You may include the following steps:*

1. *Statistical summary of the data. Then describe what you find.*
2. *Data visualizations. Add 2-3 sentences to describe what you find for each plot you draw.*
3. *Statistical test. Add 2-3 sentences to describe what you conclude based on the test results.*

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 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 than the other wealth segments.

**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 High-Value customers. Meanwhile, the Financial Services has the most 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 property valuation variable.

**5. Model Development**

Multinomial Logistic Regression

Regularized Logistic Regression

Decision Trees

Random Forest

K-means Clustering

**6. Model Performance**

*Describe the models you built with the data. You may use Rmarkdown file to record all the models you have built, \*knit\* to generate a pdf file named ‘Models’ and include the* ***rmd and pdf*** *file in the final submission.*

*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.*

**7. Model Interpretation**

**8. Limitations**