TheAnalyticsTeam

Sprocket Central Pty Ltd

Data analytics approach

Tooba Khullat

Agenda

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

Introduction

Identify and Recommend Top 1000 Customers to Target

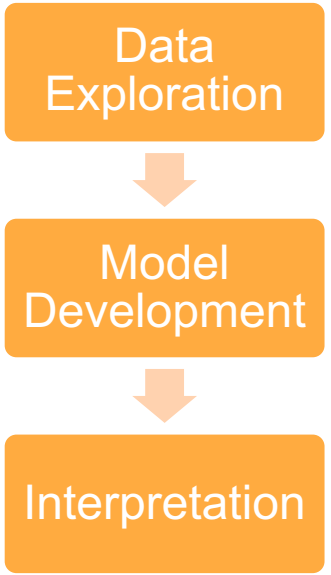
Problem Outline

- Sprocket Central is a company that specialises in high-quality bikes and cycling accessories.
- The marketing team is looking to boost business sales to analyse provided datasets.

Aim

- To analyse the three provided datasets and recommend 1000 customers that Sprocket Central should target to drive higher value for the company.

This will be done using the 3 phases of:



Introduction

Contents of Data Analysis

1. ‘New’ and ‘Old’ Customer Age Distributions
2. Bike-related purchases over the last 3 years by gender and age
3. Job industry distributions
4. Wealth segment by age category
5. Number of cars owned and not owned by state
6. RFM analysis and customer classification

Data Exploration

Data Quality Assessment and 'Clean Up'

Key Issues for Data Quality Assessment

- Accuracy: Correct Values
- Completeness: Data Fields with Values
- Consistency: Values Free from Contradiction
- Currency: Values up to date
- Relevancy: Data items with Value meta-data
- Validity: Data Containing Allowable Values
- Uniqueness: Records that are Duplicated

Summary Table

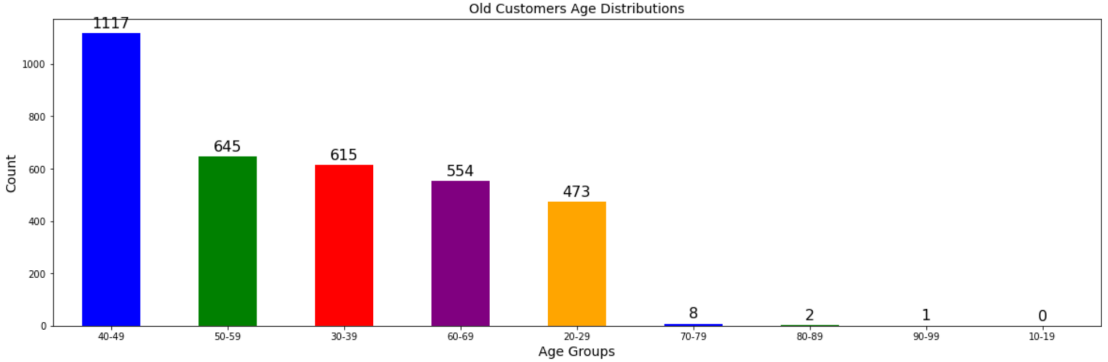
	Accuracy	Completeness	Consistency	Currency	Relevancy	Validity
Customer Demographics	DOB: inaccurate Age: missing	Job title: blanks Customer id: incomplete	Gender: inconsistent	Deceased customers: filter out	Default column: delete	
Customer Address		Customer id: incomplete	States: inconsistent			
Transactions	Profit: missing	Customer id: incomplete Online order: blanks Brands: blank			Cancelled status order: filter out	List price: format Product sold date: format

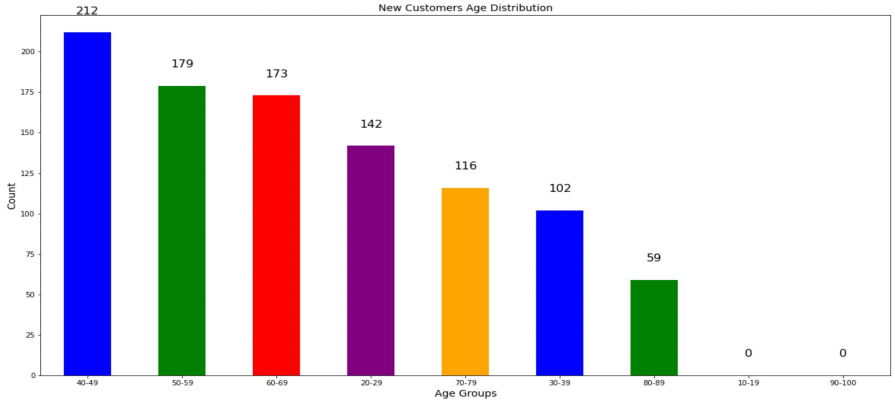
An in-depth email has been sent via email.

Data Exploration

‘New’ and ‘Old’ Customer Age Distributions

- Most customers are aged between 40-49 in ‘New’ and ‘Old’.
- The lowest age groups are 20-29 for ‘New’ and ‘Old’ and 80+ for ‘New’ and 70+ for ‘Old’.
- The ‘New’ customer list suggests that 40-69 and 20-2 are the most populated.
- The ‘Old’ customer list suggests 20-69.
- There is a steep drop of customers in the 30-39 and 80-89 age group in ‘New'

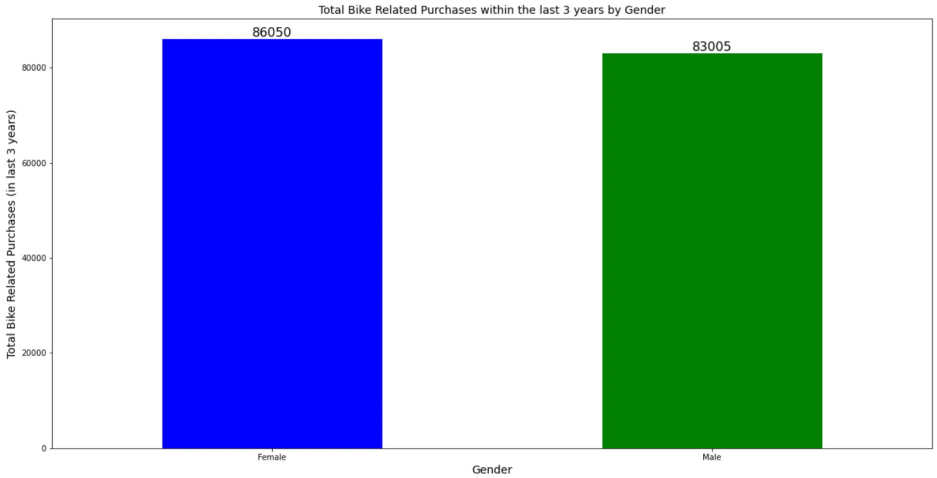




Data Exploration

Bike related purchases over last 3 years by gender

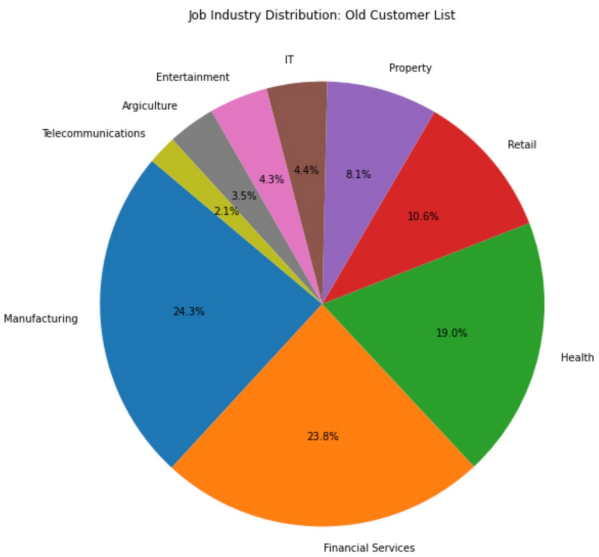
- Over the last 3 years, about 50% of bike related purchases were made by females to 49% of purchases made by males.
- Numerically, females purchases almost 3000 more than males.
- Females make majority of bike related sales.

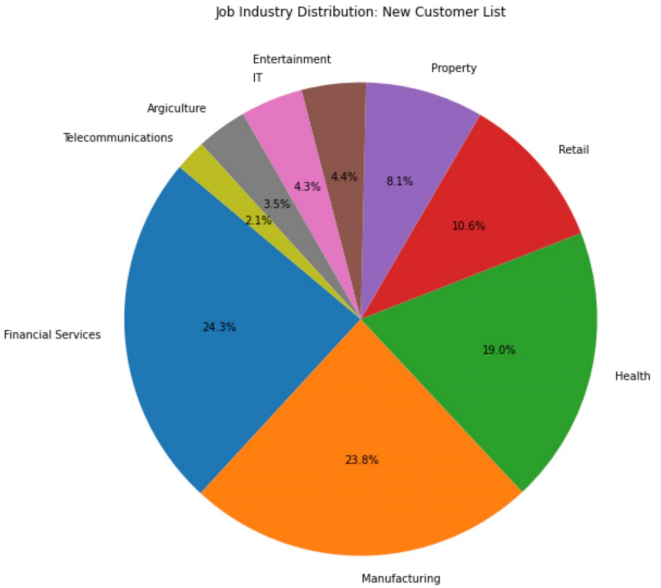


Data Exploration

Job Industry Distribution

- 24% of 'New' customers are in Manufacturing and Financial Services.
- The smallest number of customers are in Telecommunications and Agriculture.
- Similar pattern in both Customers List.

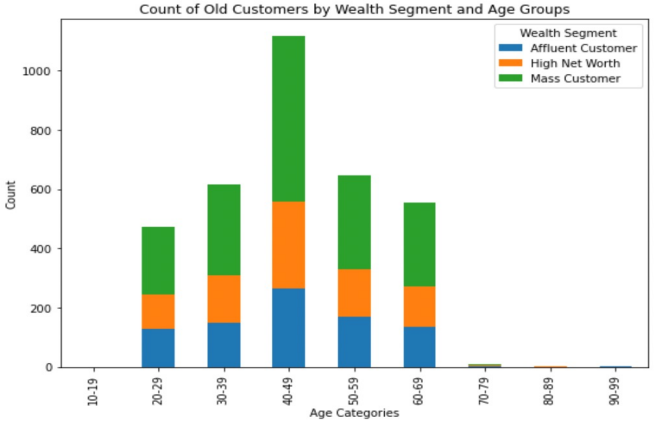




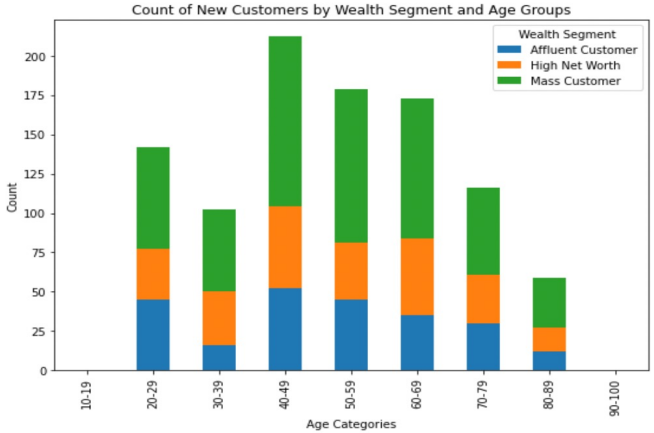
Data Exploration

Wealth Segment by Age Category

- In all age categories the largest number of customers are classified as 'Mass Customer', with 'High Net Worth' customers.
- The 'Affluent Customers' outperforms the 'High Net Worth' customers in the 40-49 age group.



Affluent Customer	High Net Worth	Mass Customer
0	0	0
129	115	229
149	159	307
263	293	561
170	161	314
134	139	281
3	3	2
0	1	1
1	0	0

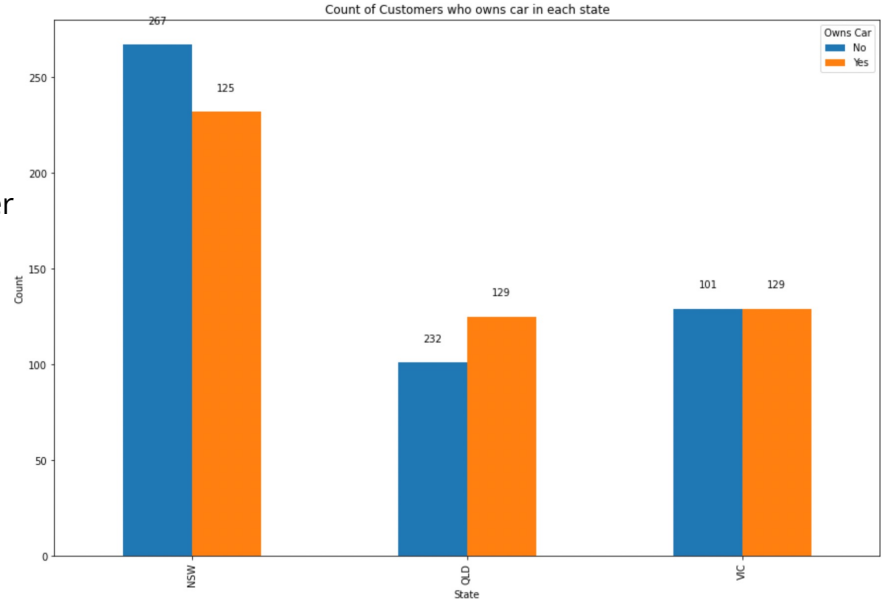


Affluent Customer	High Net Worth	Mass Customer
0	0	0
45	32	65
16	34	52
52	52	108
45	36	98
35	49	89
30	31	55
12	15	32
0	0	0

Data Exploration

Number of cars owned and not owned by state

- NSW has the highest number of customers **do not** own a car whilst QLD has the lowest. NSW seems to have a higher number of people from which data was collected.
- Victoria is split evenly but numbers are significantly lower than NSW.
- QLD has a relatively high number of customers who own a car.



Model Development

RFM Analysis

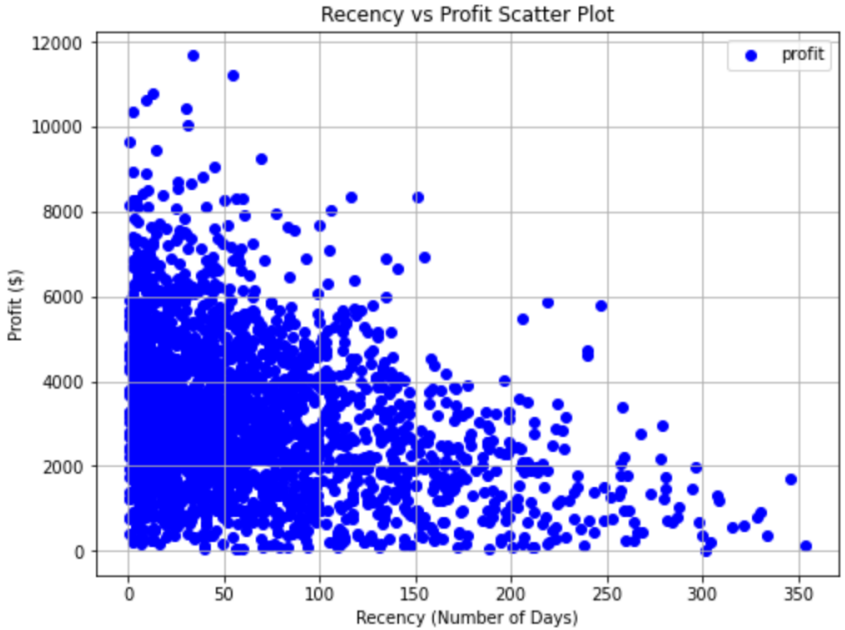
- RFM analysis is used to determine which customers a business should target to increase its revenue and value.
- The RFM (Recency, Frequency, and Monetary) model shows customers that have displayed high levels of engagement with the business in the categories mentioned.

Customer Title	RFM Score
Platinum customer	>8
Very Loyal customer	7
Becoming Loyal	6
Recent customer	5
Potential customer	4
Losing customer	3
High risk customer	2
Lost customer	1

Model Development

Scatter-Plot based off RFM analysis

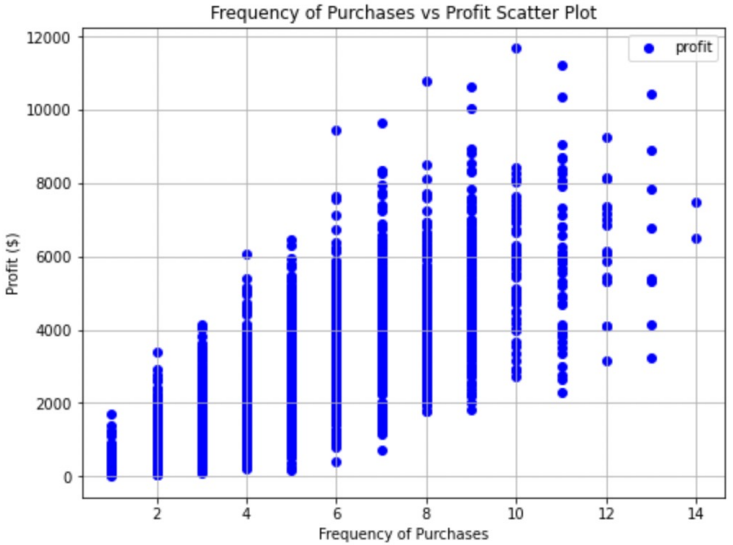
- The chart shows that customers who purchased more recently generated more revenue, than customers who purchased a while ago.
- Customers from recent past (50-100 days) also show to generate a moderate amount of revenue.
- Those who visited more than 200 days ago generated low revenue.



Model Development

Scatter-Plot based off RFM analysis

- Customer classified as “Platinum Customer”, “Very Loyal”, and “Becoming Loyal” visit frequently, which correlated with increased revenue for the business.
- Naturally, there is a positive relationship between frequency and monetary gain for the business.



Model Development

Scatter-Plot based off RFM analysis

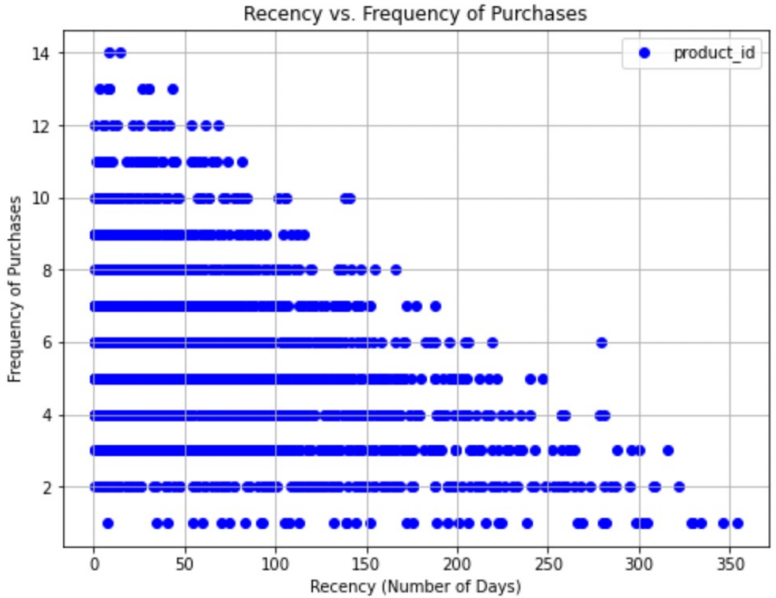
- Very low frequency of 0-2 correlated with high recency values.

  e.g. More than 350 days ago.

- Customers who have visited more recently (0-50 days) have a higher chance of visiting more frequently (6+).

- Higher frequency has a negative relationship with recency values.

  e.g. very recent customers are also frequent customers.



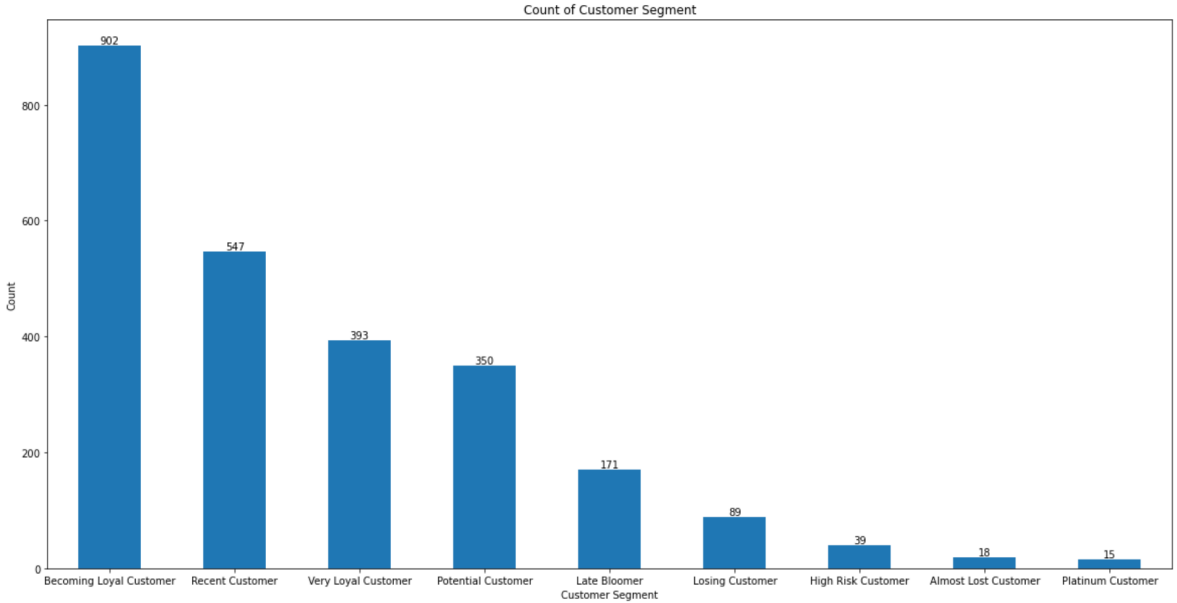
Model Development

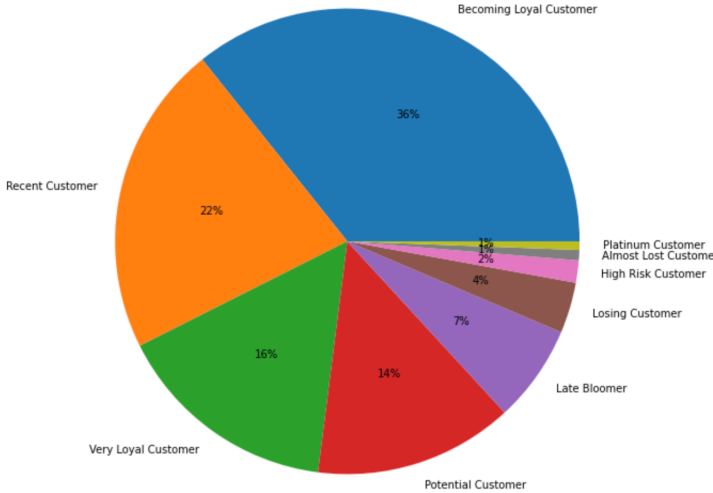
Customer Title Definition List with RFM Values Assigned

Rank	Customer Title	Description	RFM Value
1	Platinum Customer	Most recent buy, buys often, most spent	14+
2	Very Loyal Customer	Most recent buy, buys often, spends large amount of money	11-13
3	Becoming Loyal	Relatively recent, bought more than once, spend large amounts of money	9-10
4	Recent Customer	Bought recently, not very often, average money spent	8
5	Potential Customer	Bought recently, never bought before, spent small amount	7
6	Late Bloomer	No purchases recently, but RFM value is larger than average	6
7	Losing Customer	Purchase was a while ago, below average RFM value	5
8	High Risk Customer	Purchases was long time ago, frequency is quite high, amount spent is high	4
9	Almost Lost Customer	Very low recent, low frequency, but high amount spent	3
10	Evasive Customer	Very low recency, very low frequency, small amount spent	2
11	Lost Customer	Very low RFM	1

Model Development

Distribution of Customer Titles





Interpretation

Summary Table of the Top 1000 customers to target

Rank	Customer Title	Description	Number of Customers	Cumulative	Customer Selection
1	Platinum Customer	Most recent buy, buys often, most spent	15	15	15
2	Very Loyal Customer	Most recent buy, buys often, spends large amount of money	393	408	393
3	Becoming Loyal	Relatively recent, bought more than once, spend large amounts of money	902	1310	592
4	Recent Customer	Bought recently, not very often, average money spent	547	1857	0
5	Potential Customer	Bought recently, never bought before, spent small amount	350	2207	0
6	Late Bloomer	No purchases recently, but RFM value is larger than average	171	2378	0
7	Losing Customer	Purchase was a while ago, below average RFM value	89	2467	0
8	High Risk Customer	Purchases was long time ago, frequency is quite high, amount spent is high	39	2506	0
9	Almost Lost Customer	Very low recent, low frequency, but high amount spent	18	2524	0
10	Evasive Customer	Very low recency, very low frequency, small amount spent	0	2524	0
11	Lost Customer	Very low RFM	0	2524	0

Interpretation

Customer Target and Methodolody

Rank	Customer Title	Description	Number of Customers	Cumulative	Customer Selection
1	Platinum Customer	Most recent buy, buys often, most spent	15	15	15
2	Very Loyal Customer	Most recent buy, buys often, spends large amount of money	393	408	393
3	Becoming Loyal	Relatively recent, bought more than once, spend large amounts of money	902	1310	592

- Filter through the top 1000 customers by assigning the conditions discussed in the table above.
- The 1000 customers discovered would’ve bought very frequently in the past and tend to spend more money than the other customers.

Future Improvement/Recommendation

- Bring in ABS data at different geographical levels and create additional features for the model e.g. geographical remoteness of different postcodes to gain proximity of a customers journey to their work and whether in need of a bike.
- Determine a hypothesis-related question and perform statistical analysis.
- Test the performance of the model using factors relevant for the given model chosen (i.e. residual deviance, AIC, ROC curves, R Squared).
- Document model performance, assumptions and limitations.

Appendix

Appendix

1. http://www.abs.gov.au/browse?opendocument&ref=topBar