KPMG Virtual Internship Data Analysis Presentation

Arthur Lam

Agenda

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

Introduction

Identify and Recommend Top 1000 Customer to target from Datasets

Outline of Problem

1. Sprocket Central is a company that specializes in high-quality bikes and cycling accessories
2. Their marketing team is looking to boost business sales by analyzing provided data sets
3. Using the 3 data sets provided the aim is to analyze and recommend 1000 customers that Sprocket Central should target to dive higher value for the company

This will be done with the three phases of: Data Exploration, Model Development, and Interpretation

Contents of Data Analysis

- ‘New’ and ‘Old’ Customer Age Distributions
- Bike related purchases over the last 3 years by gender
- Job industry distributions
- Wealth segmentation by age category
- Number of cars owned and not owned by state
- 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

An in-depth analysis has been sent via email

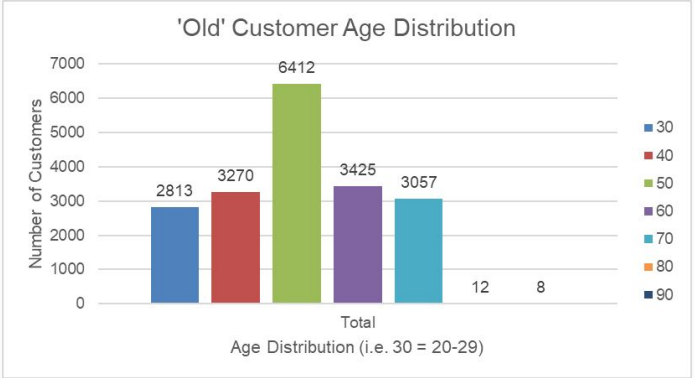
Summary Table

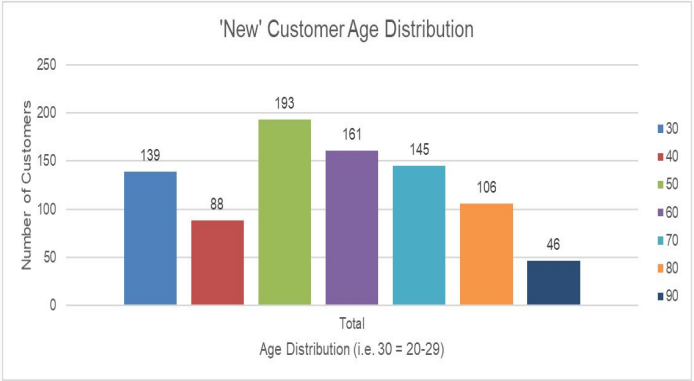
	Accuracy	Completeness	Consistency	Currency	Relevancy	Validity
Customer Demographic	- DOB: inaccurate - Age: missing	- Job title: blanks - Customer id: incomplete	-Gender: inconsistency	- Deceased customers: filter out	- Default column: deleted	
Customer Address		- Customer id: incomplete	- States: inconsistency			
Transactions	- Profit: missing	- Customer id: incomplete - Online order: blanks - Brand: blanks			- Canceled status order: filter out	- List price: format - Product sold date: format

Data Exploration

‘New’ and ‘Old’ Customer Age Distributions

- Most customers are aged between 40-49 for both ‘Old’ and ‘New’ customers
- Aside from 80+, the lowest distribution age group is under 30 for ‘Old’ and 30-40 for ‘New’
- The ‘New’ customer list suggests a rise in customers between age 50 - 70
- There is a steep drop of customers in the 30-39 age group in ‘New’

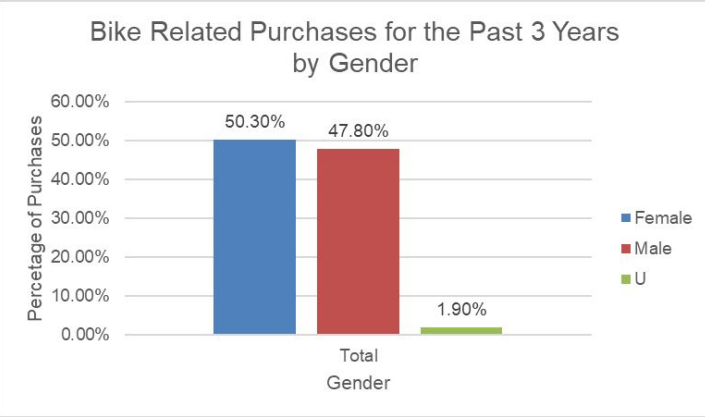


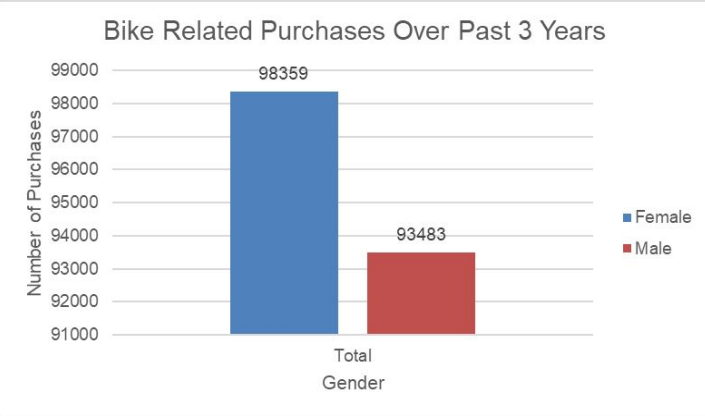


Data Exploration

Bike related purchases over last 3 years by gender

- Over the last three years ~50% of bike related purchases were made by females to ~48% of purchases made by males
- Approximately 2% were made by unknown gender
- Numerically, Females make 5000 more transactions than Males
- Females make up majority of bike related sales

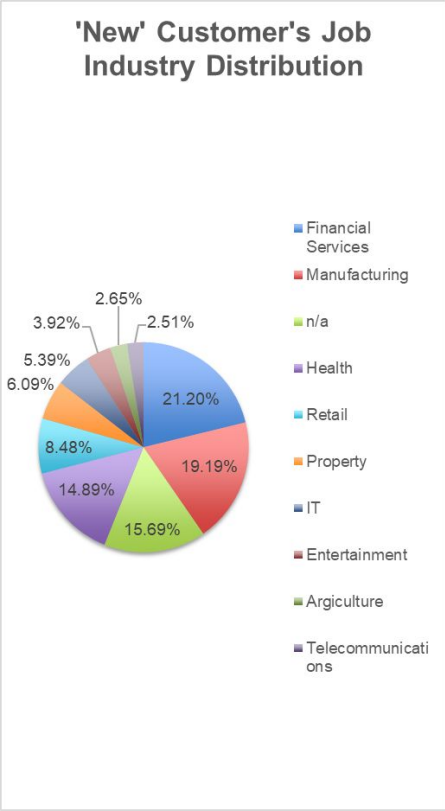


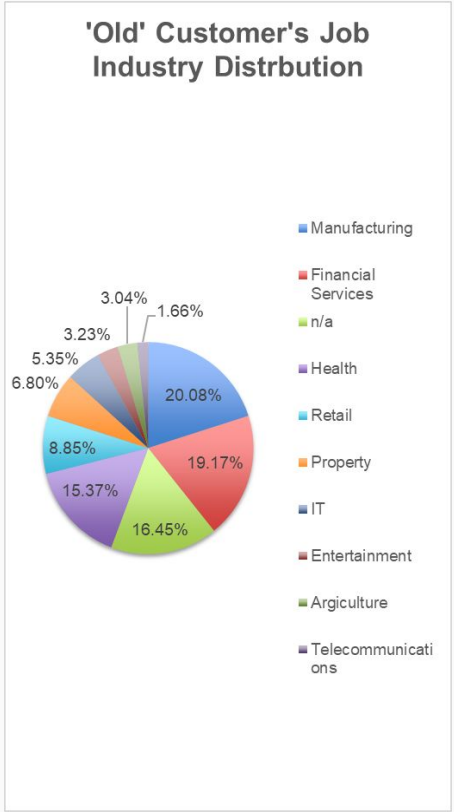


Data Exploration

Job Industry Distribution

- ~40% of ‘New’ and ‘Old’ Customers are in Manufacturing and financial Services.
- The smallest number of both types of customers are in Agriculture and Telecommunications at 3%
- Similar pattern in ‘Old’ customer list and ‘New’ customer

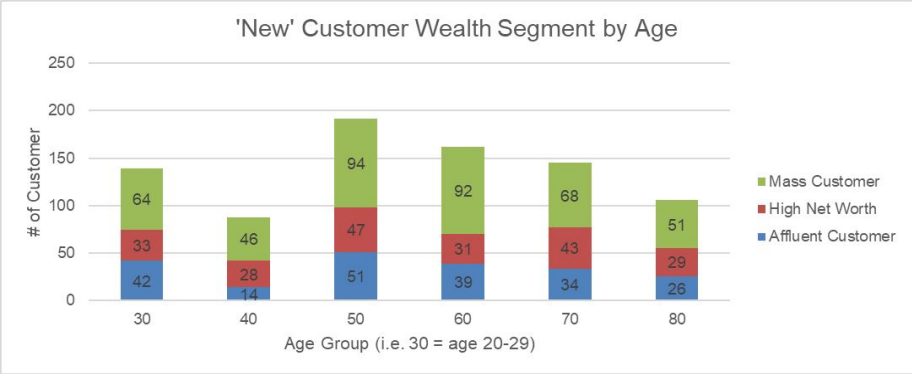


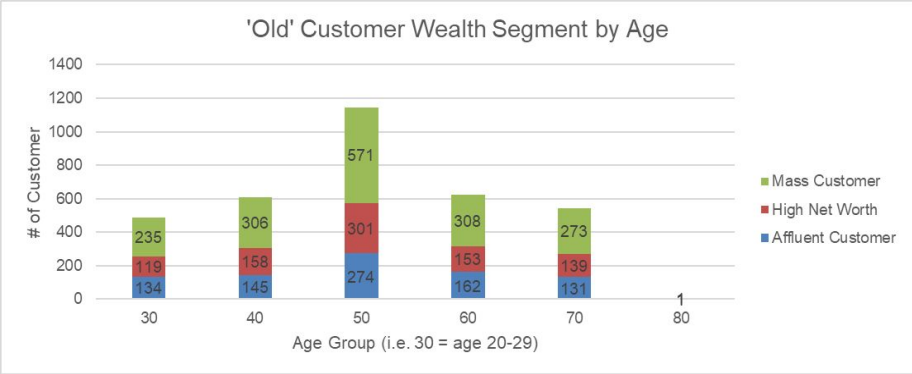


Data Exploration

Wealth Segmentation by Age Category

- In all age categories the largest number of customers are classified as ‘Mass Customer’
- The next category is the ‘High Net Worth’ customers
- Distribution patterns are consistent throughout age groups in both customer types
- A rise of age 70-79 customers according ‘New’ customer data

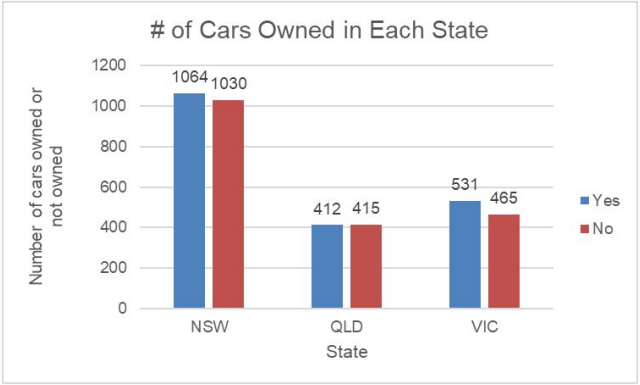


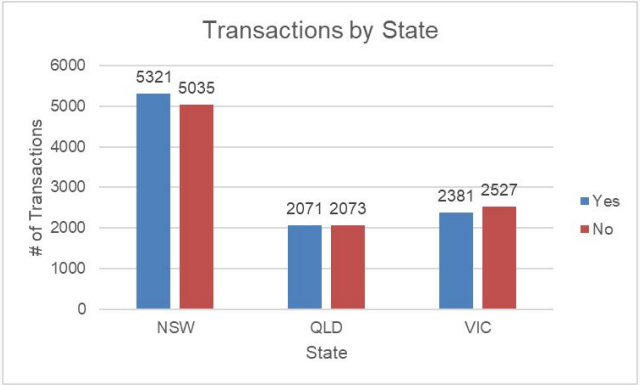


Data Exploration

Number of cars owned and not owned by state

- New South Wale (NSW) has the largest amount of customers that do not own a car. NSW seems to have a higher number of customers from which data was collected
- Victoria (VIC) is also split quite evenly. But both numbers are significantly lower than those of NSW
- Queensland (QLD) has a relatively high number of customers that own a car

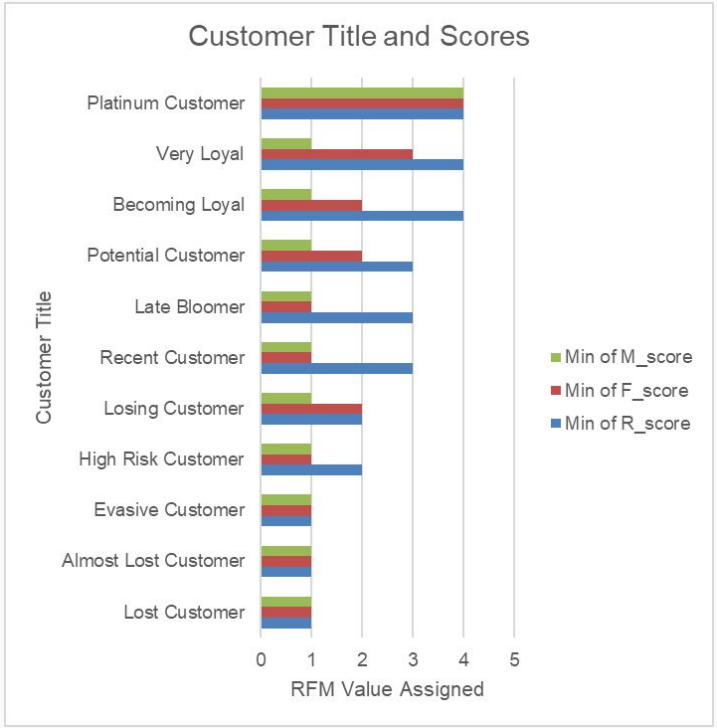




Model Development

RFM Analysis and Customer Classification

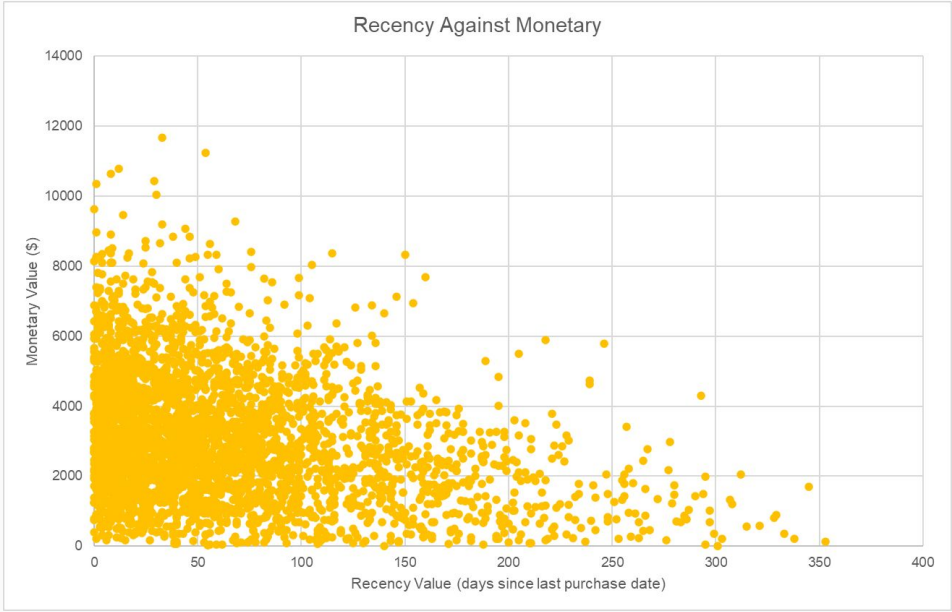
- RFM analysis is used to determine which customers a business should target to increase its revenue and value
- The RFM (Recency > Frequency > Monetary) model shows customers that have displayed high levels of engagement with the business in the three categories mentioned



Model Development

Scatter-Plot based off RFM Analysis

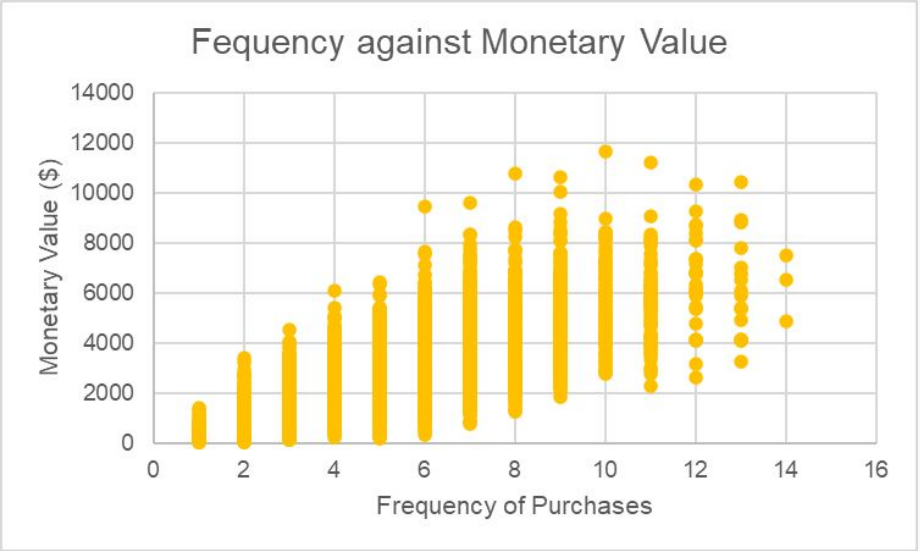
- The chart shows that customers who purchased more recently have generated more revenue, than customer who visited a while ago
- Customers from recent past (50-100 days) also show to generate a moderate amount of revenue
- Those who visited more than 200 days ago generated low revenue



Model Development

Scatter-Plot based off RFM Analysis

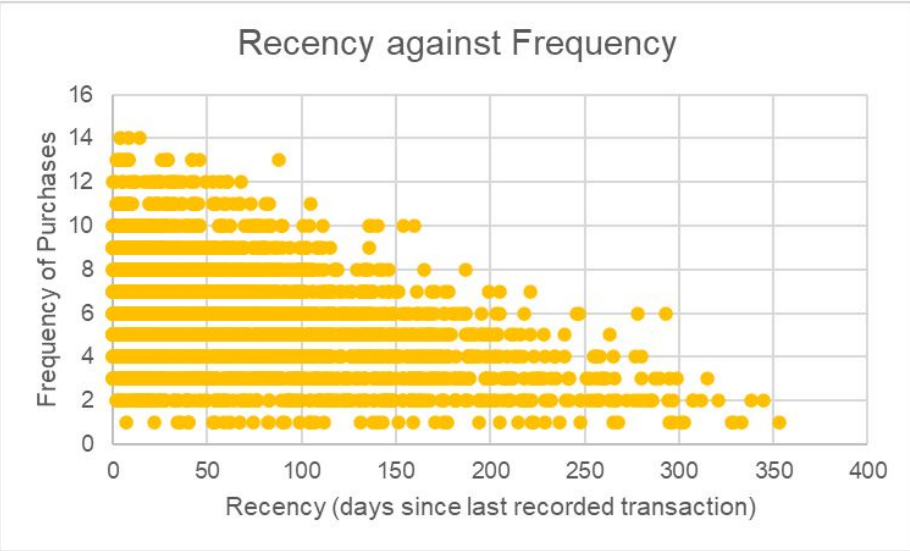
- Customer classified as “Platinum Customer,” “Very Loyal,” and “Becoming Loyal” visit frequently, which correlated with increased revenue for the business
- Naturally, there is a positive relationship between frequency and monetary gain for the business



Model Development

Scatter-PLot based off RFM Analysis

- Very low frequency of 0-2 correlated with high recency values. i.e. More than 250 days ago
- Customers that 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. Such that very recent customers are also frequency customers



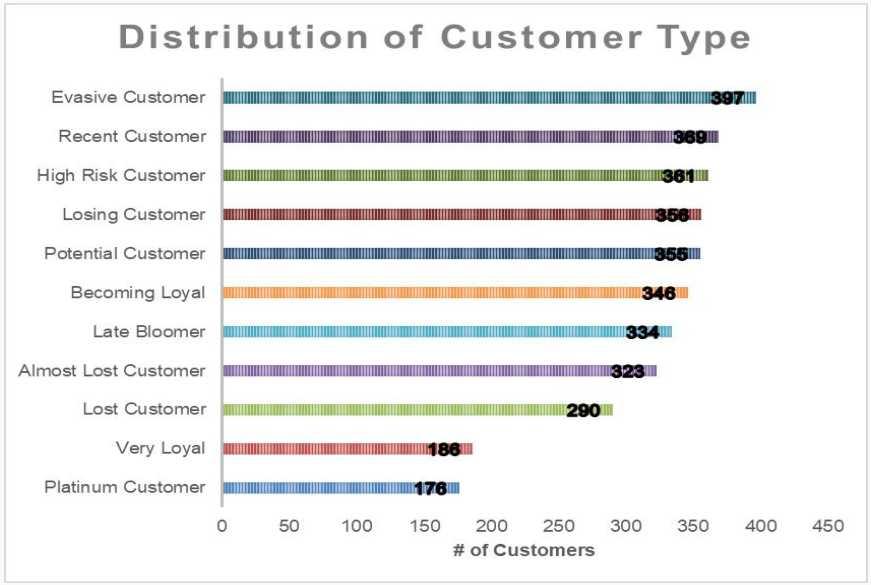
Model Development

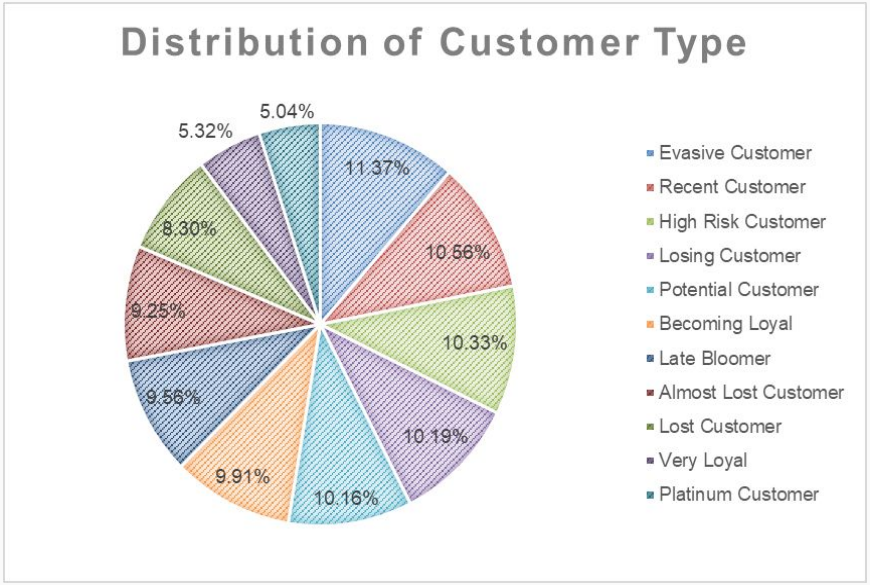
Customer Title Definition List With Rfm Values Assigned

Rank	Customer Title	Description	RFM Value
1	Platinum Customer	Most recent buy, buys often, most spent	444
2	Very Loyal	Most recent, buys often, spends large amount of money	433
3	Becoming Loyal	Relatively recent, bought more than once, spends large amount of money	421
4	Recent Customer	Brought recently, not very often, average money spent	344
5	Potential Customer	Bought recently, never bought before, spent small amount	323
6	Late Bloomer	No purchases recently but RFM value is larger than average	311
7	Losing Customer	Purchase a while ago, below average RFM value	224
8	High Risk Customer	Purchase long time ago, frequency is quite high, amount spent is high	212
9	Almost Lost Customer	Very low recency, very low frequency, high amount spent	124
10	Evasive Customer	Very low recency, very low frequency, small amount spent	112
11	Lost Customer	Very low RFM value	111

Model Development

Customer Title Distributions in Dataset





Model Development

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	176	176	176
2	Very Loyal	Most recent, buys often, spends large amount of money	186	362	186
3	Becoming Loyal	Relatively recent, bought more than once, spends large amount of money	346	708	346
4	Recent Customer	Brought recently, not very often, average money spent	369	1077	292
5	Potential Customer	Bought recently, never bought before, spent small amount	355	1432	0
6	Late Bloomer	No purchases recently but RFM value is larger than average	334	1766	0
7	Losing Customer	Purchase a while ago, below average RFM value	356	2122	0
8	High Risk Customer	Purchase long time ago, frequency is quite high, amount spent is high	361	2483	0
9	Almost Lost Customer	Very low recency, very low frequency, high amount spent	323	2806	0
10	Evasive Customer	Very low recency, very low frequency, small amount spent	397	3203	0
11	Lost Customer	Very low RFM value	290	3493	0

Interpretation

Customer Target and Methodology

Rank	Customer Title	Description	Number of Customers	Cumulative	Customer Selection
1	Platinum Customer	Most recent buy, buys often, most spent	176	176	176
2	Very Loyal	Most recent, buys often, spends large amount of money	186	362	186
3	Becoming Loyal	Relatively recent, bought more than once, spends large amount of money	346	708	346
4	Recent Customer	Brought recently, not very often, average money spent	369	1077	292

- Filter through the top 1000 customers by assigning the conditions discussed in the table above (RFM values & Titles)
- The 1000 customers would have bought recently, frequently and tend to spend more money than other customers
- The top 100 customers' meta data will help us understand our best customer profile

Appendix

- Excel Files / Data Set