TheAnalyticsTeam

Sprocket Central Pty Ltd

Data Analytics Approach

[Division Name] - [Engagement Manager], [Senior Consultant], [Junior Consultant]

This presentation has live translations.

Presented by:
Joaquim Bolós Fernández

Date Submitted:
July 31th, 2023

Agenda

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

Introduction

SPROCKET CENTRAL

Company Description

Sprocket Central Pty Ltd, a long-standing KPMG client, is a company that specializes in high-quality bikes and cycling accessories.

Overall Objective

Their marketing team aims to boost business sales by analyzing their existing customer dataset to identify customer trends and behaviours.

Project Aim

Utilizing the labelled dataset of three existing datasets (Customer demographic, Customer address, and Transactions), the project's goal is to recommend the most valuable new customers, out of the 1000, to be targeted by the marketing team for effective marketing campaigns and improved business growth.

Problem Statement

Back to Agenda

01

Target New Customers for the marketing Team to drive higher value for the company

Strategic Objective

02

Capacity to determine which clients should be targeted by the marketing Team

Tactical Objectives

03

Perform EDA and develop a classification model to determine relevant clients

Operational Objectives

Project Outline / Operational Objectives

Perform EDA and develop a classification model to determine relevant clients

Exploratory Data Analysis

Data Quality Assessment

Data Exploration

- Age Distributions
- Purchases Over the last 3 Years
- Job Industry Distribution
- Wealth Segmentation by Age
- Number of Cars by State

RFM Analysis and Customer Classification

- Scatter Plots of RFM Analysis
  - Recency vs Monetary
  - Frequency vs Monetary
  - Recency vs Frequency
- Customer Title Definition List with RFM values
- Customer Title Distributions in Dataset

Model Development

- Multi Classification Sci-Kit Learn Model
  - Data Preparation
  - Models
  - Results

Back to Agenda

Data Exploration - Data Quality Assessment

SPROCKET CENTRAL

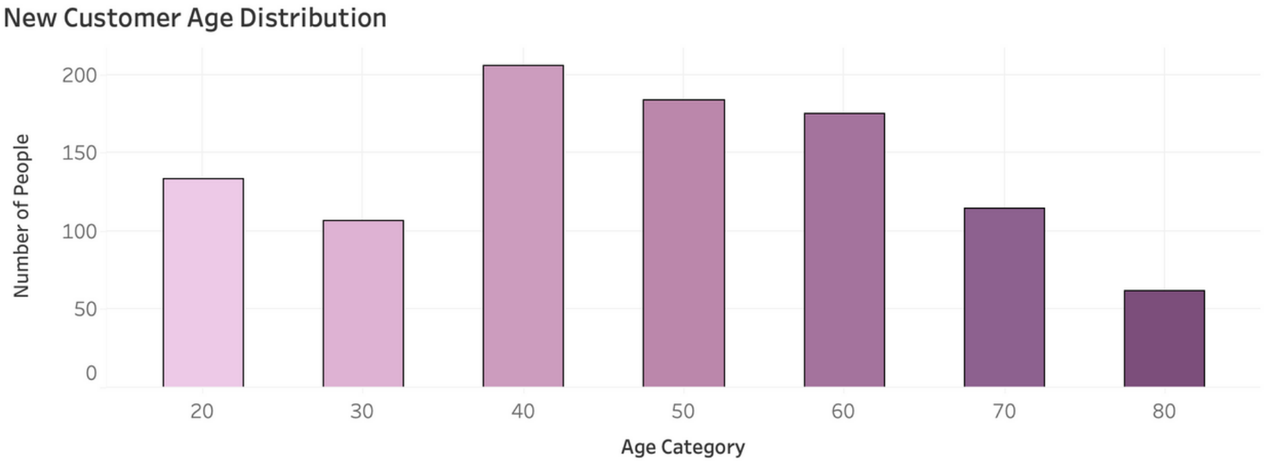
	Accuracy Correct Values	**Completeness** Data Fields with Values	**Consistency** Values Free From Contradiction	**Currency** Values Up to Date	**Relevancy** Data Items with Value Meta-Data	**Validity** Data Containing Allowable Values	**Uniqueness** Values that are Duplicated
Customer Demographic	One value in the Date of Birth (DOB) column is incorrect, showing the year as 1843.	Data fields have missing values, including: last_name, DOB, job_title, job_industry_category, default, and tenure.	Inconsistencies between the job_title and job_industry_category columns, and gender column.	Columns such as deceased_indicator and owns_car may not be up to date.		The default column does not contain any allowable values.	
Customer Address		Customers with IDs [3, 10, 22, 23] do not have address data.	Inconsistencies in State naming	The address data may not be up to date.	Country column irrelevant.		
Transactions	Profit is missing	The online_order column has missing data. The product_id 0 has missing data in multiple columns				The product_first_sold_date values and list price are in the wrong format.	

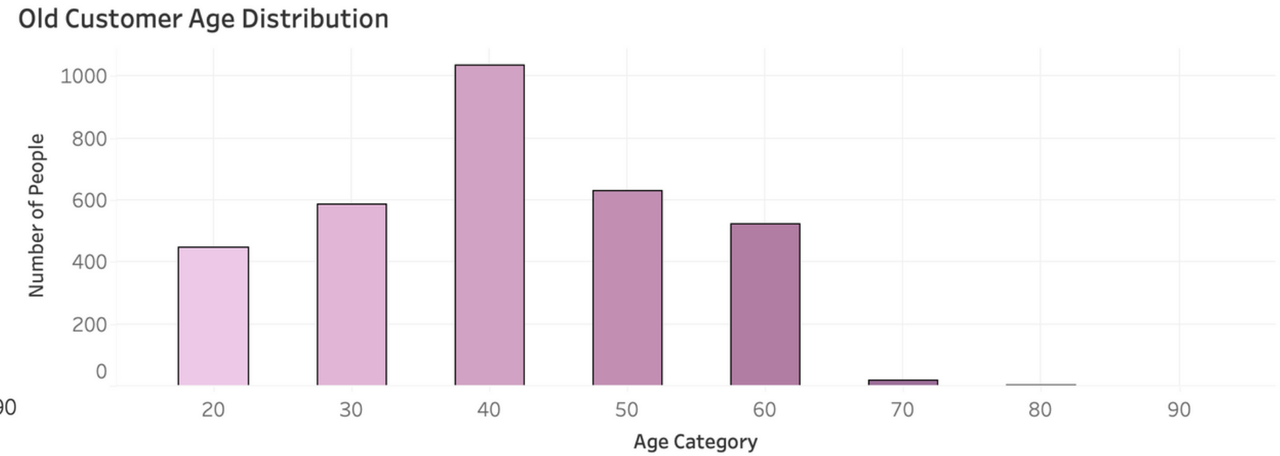
An in-depth email about this matter has been sent with more information

Back to Agenda

Data Exploration - Age Distributions

- Most customers are in their 40's.
- The lowest age group is over 70 for the old customer list.
- The new customer list suggests that people in their 20's and 40-60's are the most populated.
- Most of the old customer list clients are between 20-70 years old.
- There is a steep drop in customers in the age 30-40 for the new customers.





Age Category
20 30 40 50 60 70 80 90

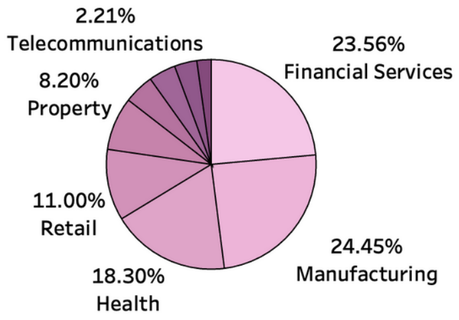
Source: Tableau

Back to Agenda

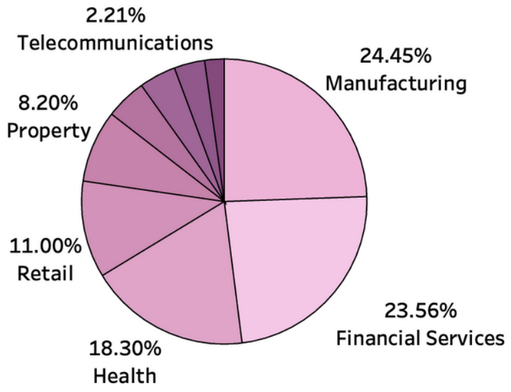
Data Exploration - Job Industry Distribution

- Both New and Old Customers datasets have similar Job Industry Distribution.
- Almost 50% of the customers are in manufacturing and financial services.
- The next big group (37.5%) of customers works in Health, Retail and Property.
- Finally, the smallest percentage of customers work in Telecommunications, Agriculture, Entertainment and IT.

New Customers Job Industry Distribution



Old Customers Job Industry Distribution



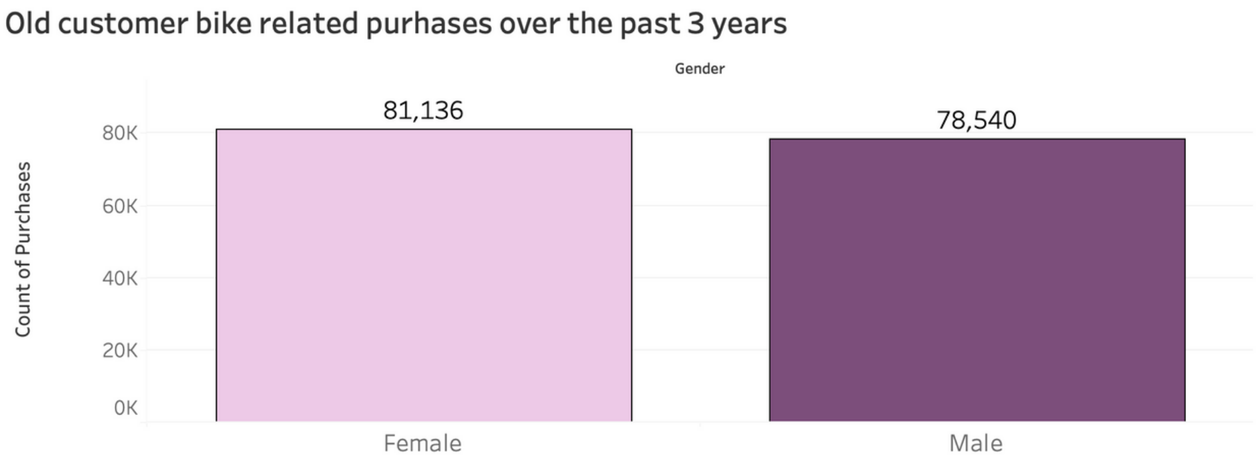
Job Industry Category
- Financial Services
- Manufacturing
- Health
- Retail
- Property
- IT
- Entertainment
- Argiculture
- Telecommunications

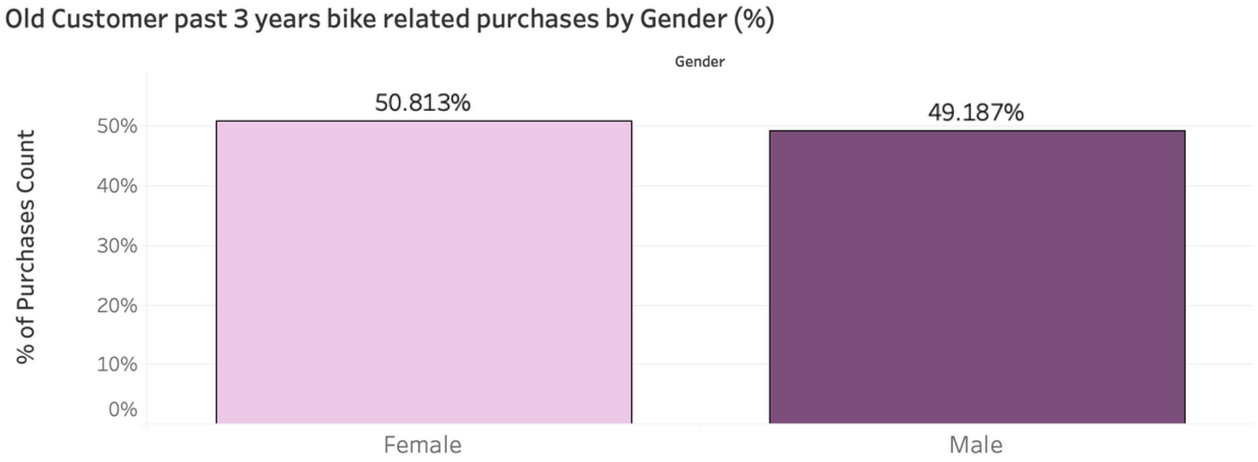
Source: Tableau

Back to Agenda

Data Exploration - Purchases Over the last 3 Years by Gender

- Over the last 3 years, about 51% of purchases were made by females, to 49% of purchases were made by males.
- Numerically, females almost have 2000 purchases more than males
- Females make up the majority of sales.





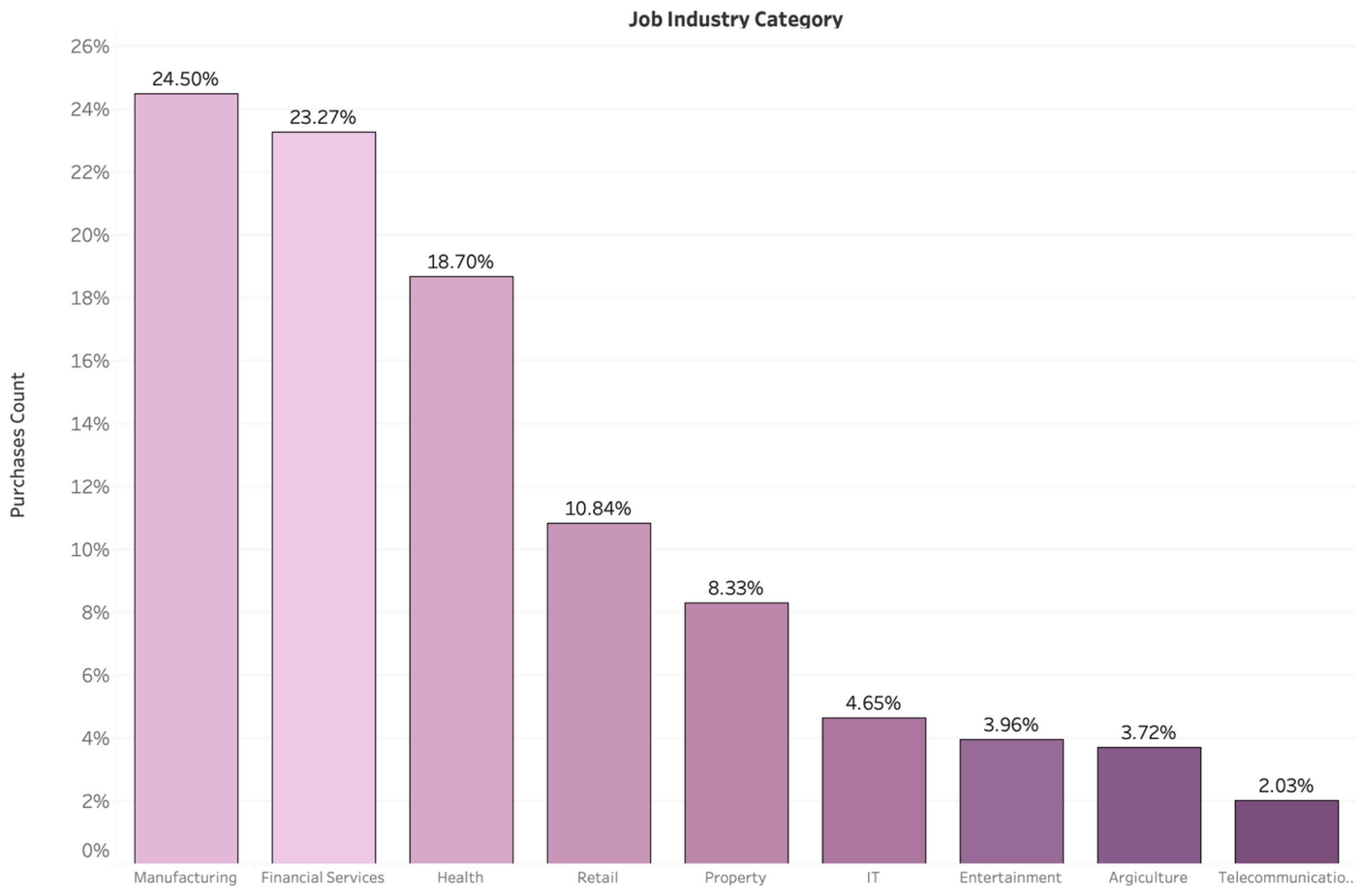
Source: Tableau

Back to Agenda

Data Exploration - Purchases Over the last 3 Years by Job Industry

- The amount of sales by job industry corresponds to the job industry distribution.

- Consequently, the job industry from new customers provably gives us little indication of the potential as a customer.



Source: Tableau

Back to Agenda

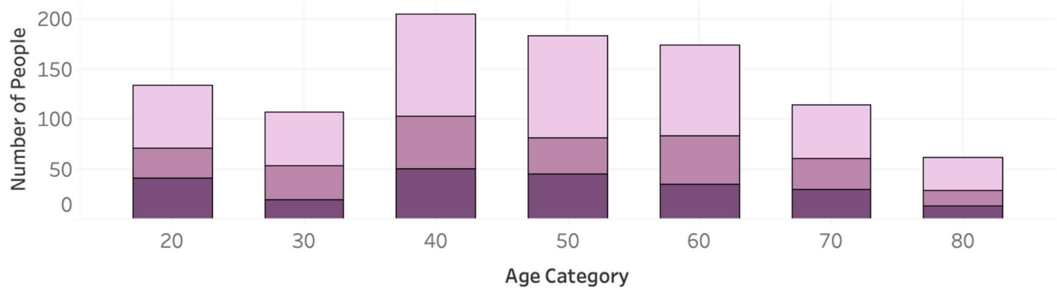
Data Exploration - Wealth Segmentation by Age

- In all age categories, the most significant number of customers are classified as "Mass customers".

- The "High Net Worth" Customers comprise the second highest number of customers in the dataset and are generally more than "Affluent Customers" for each age category.

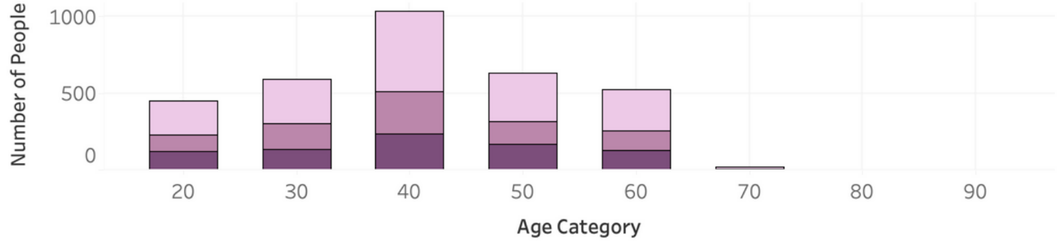
- The "Affluent Customer" outperform the "High Net Worth" Customers in the 50's age group for the New Customers dataset and the 20's and 60's age group for the Old Customers dataset.

New Customer Wealth Segment by Age



	20	30	40	50	60	70	80
Mass Customer	63	53	103	102	91	54	33
High Net Worth	30	34	52	37	49	31	16
Affluent Customer	41	20	51	45	35	30	13

Old Customer Wealth Segment by Age



	20	30	40	50	60	70	80	90
Mass Customer	221	289	529	316	270	9	2	
High Net Worth	105	163	270	151	131	5	1	
Affluent Customer	122	137	238	165	126	5		1

Wealth Segment
Mass Customer High Net Worth Affluent Customer

Source: Tableau

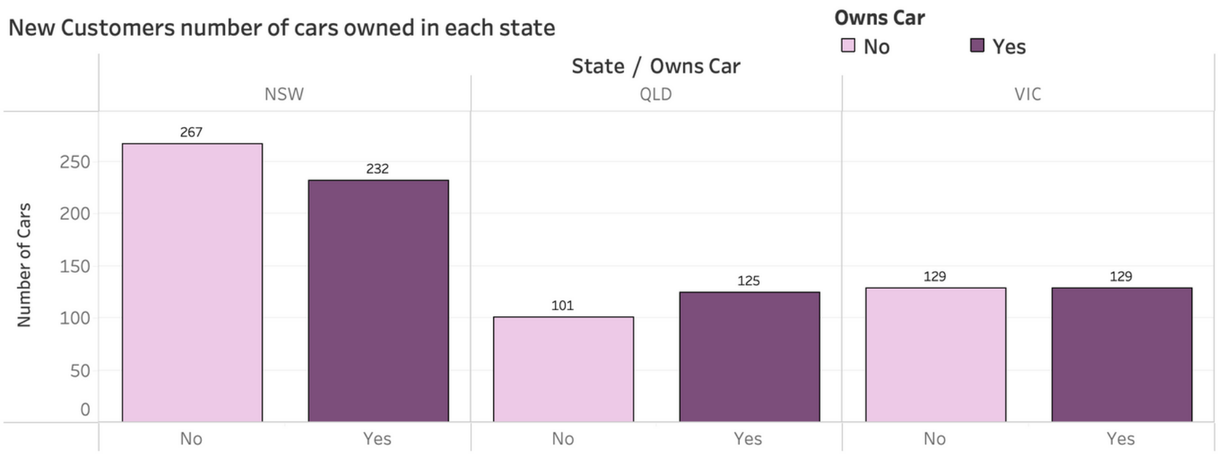
Back to Agenda

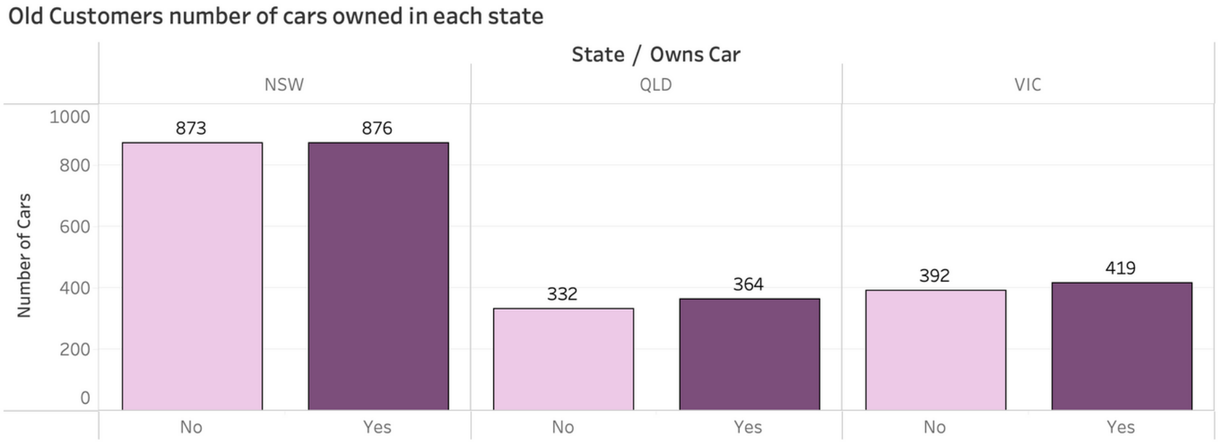
Data Exploration - Number of Cars by State

- New South Wales has more customers in both New and Old datasets. It also has the largest number of people that do not own a car in numbers and percentages.

- Victoria is split relatively evenly, but the number of customers in this state is significantly lower than in New South Wales.

- Queensland has even fewer customers than Victoria, but a relatively high number of customers own a car.





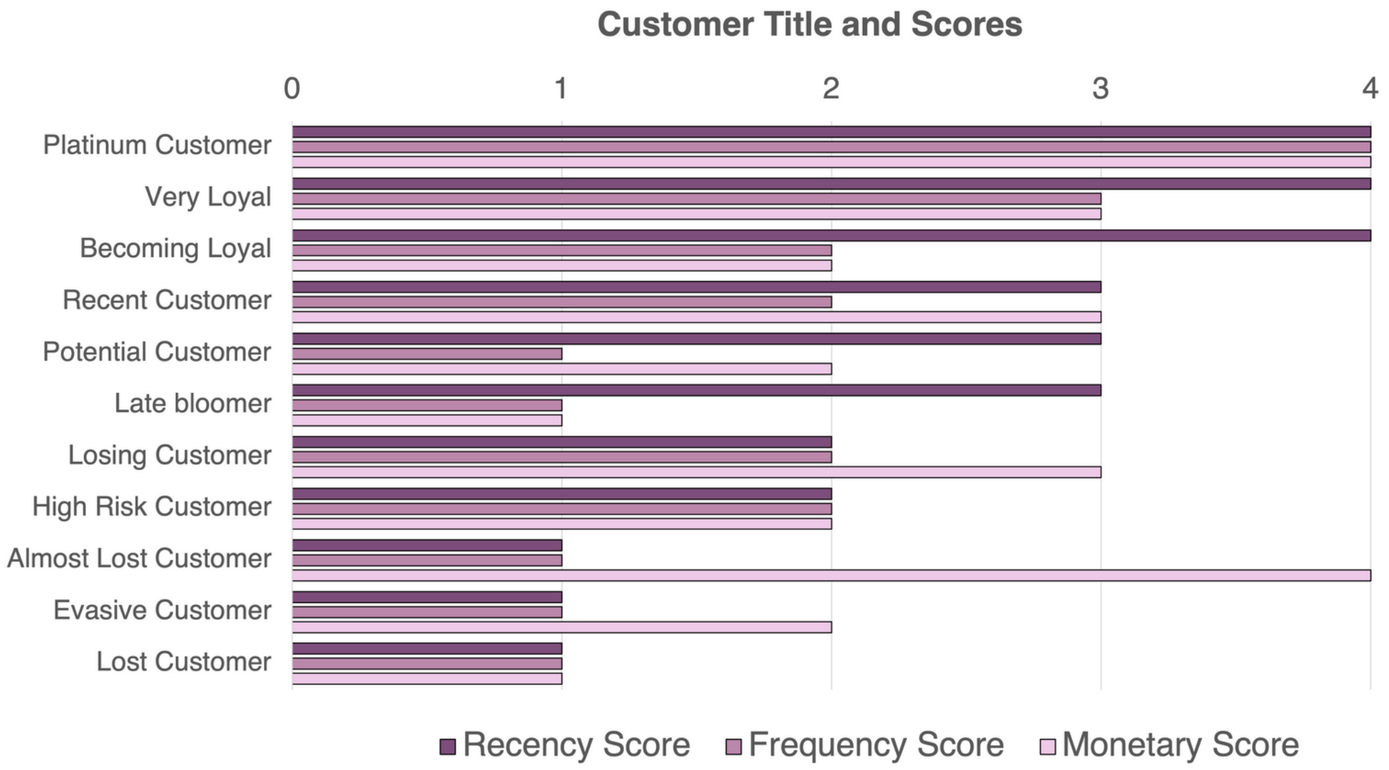
Source: Tableau

Back to Agenda

Model Development - RFM Analysis

- RFM analysis determines which customers a business should target to increase its revenue and value. Consequently, it will be helpful to target the "best customers" in the New Customers DataSet.

- The RFM (Recency, Frequency and Monetary) model shows customers with high levels of engagement within the business in the three categories mentioned.



Back to Agenda

Model development - RFM Customer Title Definition List

SPROCKET CENTRAL

RANK	CUSTOMER TITLE	CUSTOMER SUBTITLE	DESCRIPTION	R SCORE	F SCORE	M SCORE	RFM VALUES
1	Platinum	Platinum Customer	Most recent buy, buys often, most spent	4	4	4	444
2	Platinum	Very Loyal	Most recent, buys often, spends large amount of money	4	3	3	433
3	Platinum	Becoming Loyal	Relatively recent, bought more than once, spends large amount of money	4	2	2	422
4	Gold	Recent Customer	Bought recently, not very often, average money spent	3	2	3	323
5	Gold	Potential Customer	Bought recently, never bought before, spent small amount	3	1	2	312
6	Gold	Late bloomer	No purchase recently, but RFM value is higher than average	3	1	1	311
7	Silver	Losing Customer	Purchase was a while ago, below average RFM value	2	2	3	223
8	Silver	High Risk Customer	Purchase was a long time ago, frequency is quite high, amount spent is high	2	2	2	222
9	Bronze	Almost Lost Customer	Very low recency, low frequency, but high amount spent	1	1	4	114
10	Bronze	Evasive Customer	Very low recency, very low frequency, small amount spent	1	1	2	112
11	Bronze	Lost Customer	Very low RFM	1	1	1	111

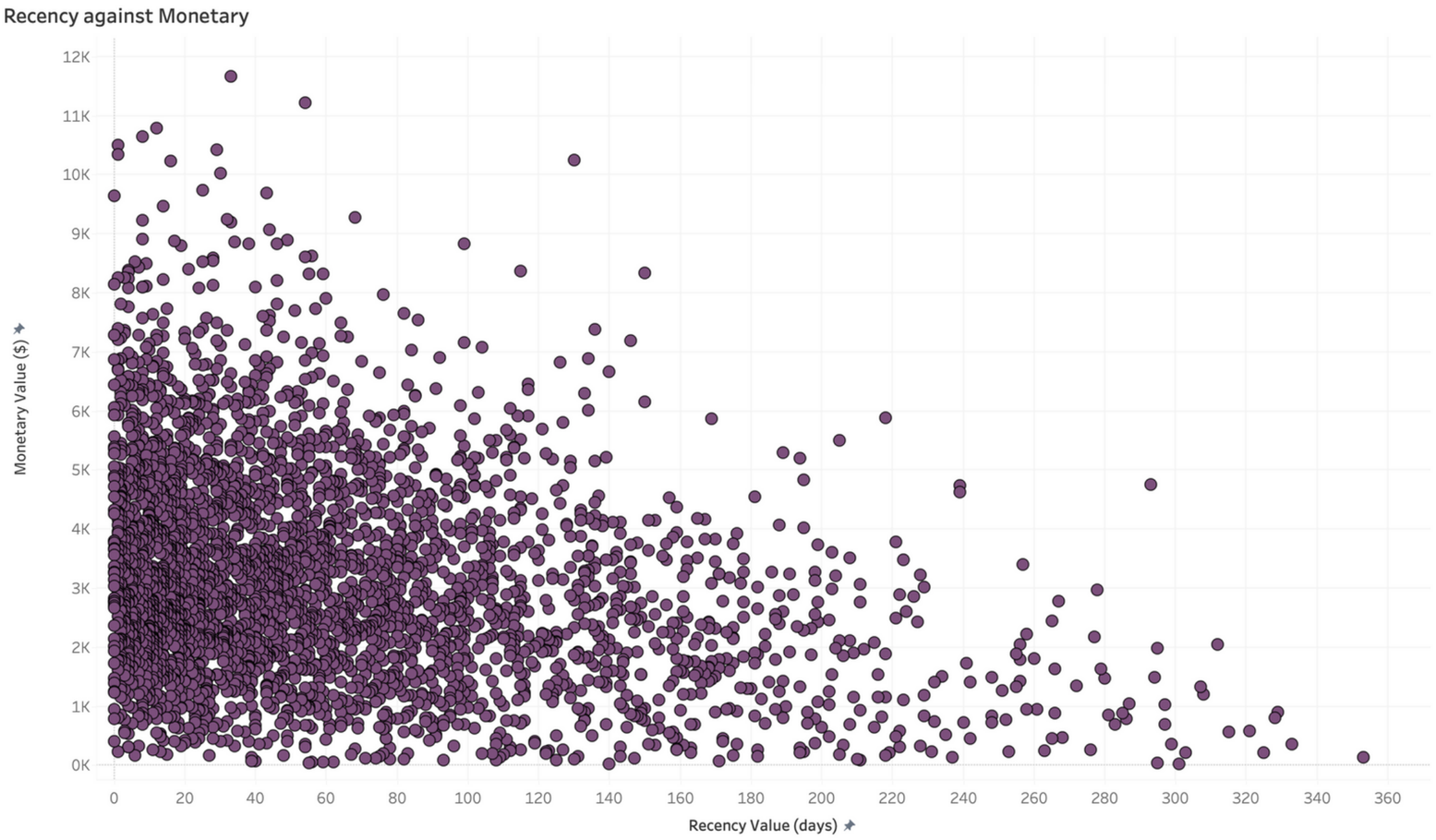
Back to Agenda

Model Development - RFM Analysis - Scatter Plots

- The chart shows that customers who purchased more recently have generated more revenue than those who visited a while ago.

- Customers from the recent past (50-100 days) also generate moderate revenue.

- After 200 days, the customers generate low revenue.



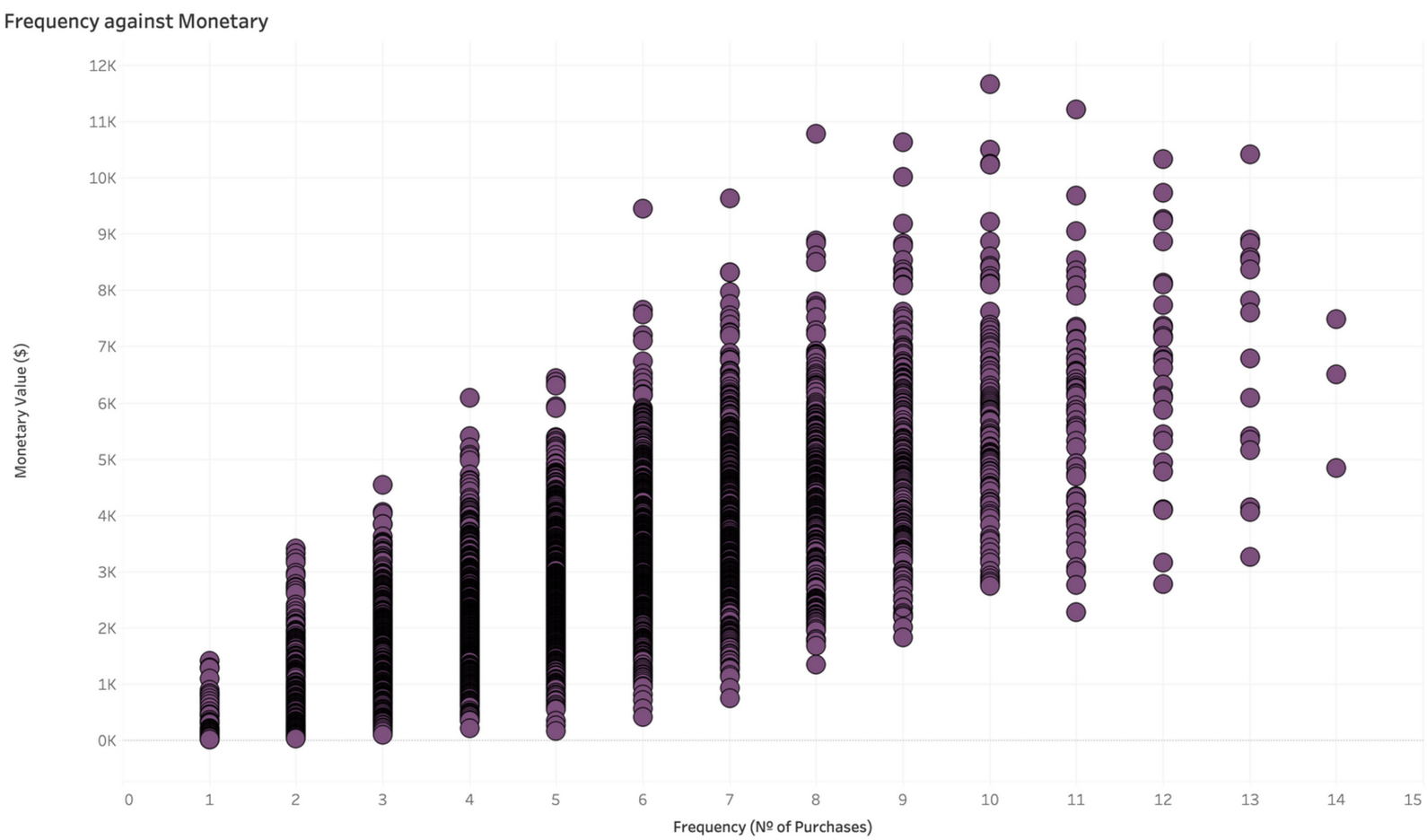
Source: Tableau

Back to Agenda

Model Development - RFM Analysis - Scatter Plots

- Customers classified as Platinum correlate with increased revenue for the business

- Naturally, there is a positive relationship btw frequency and monetary gain for the business.

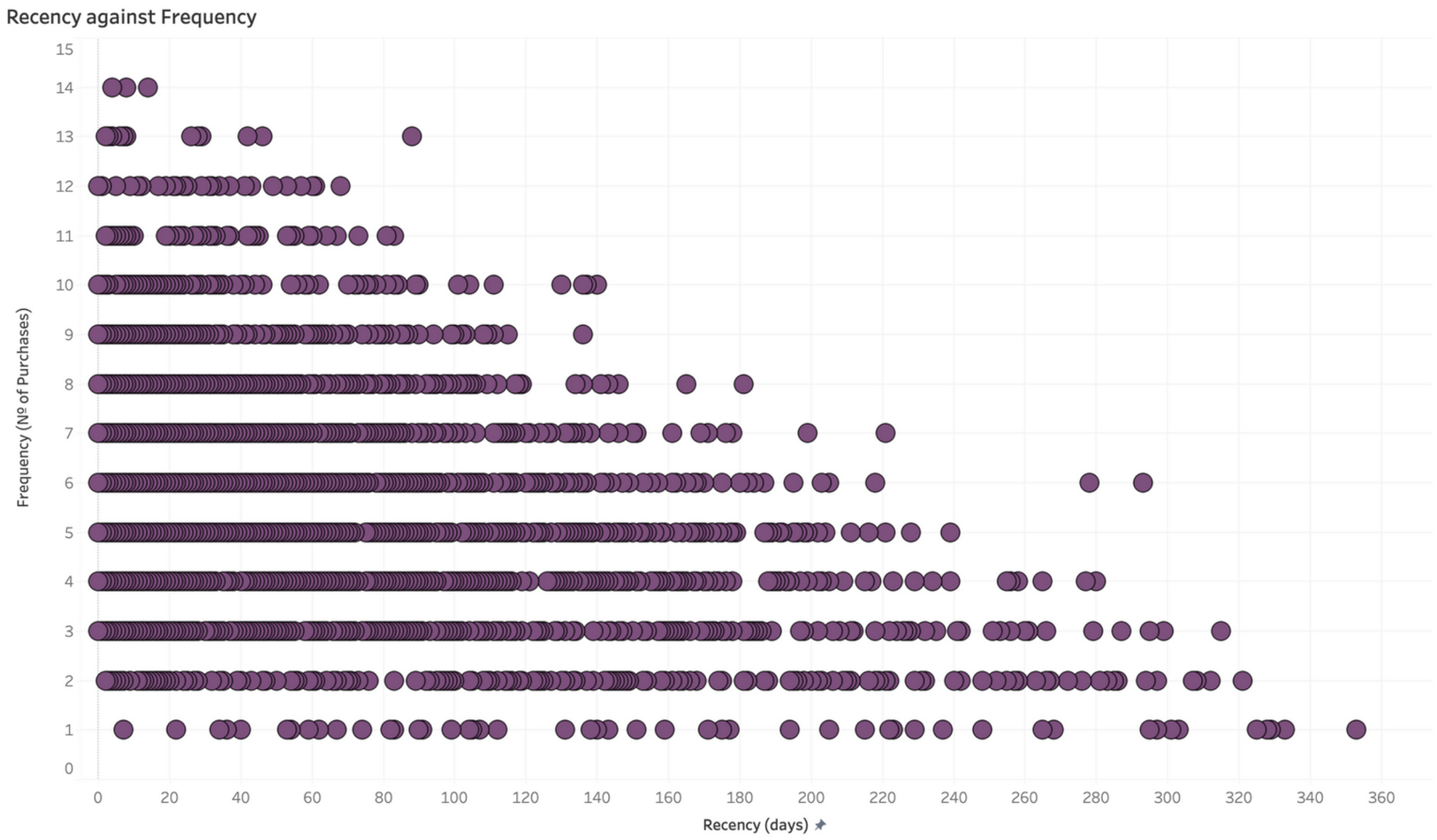


Source: Tableau

Back to Agenda

Model Development - RFM Analysis - Scatter Plots

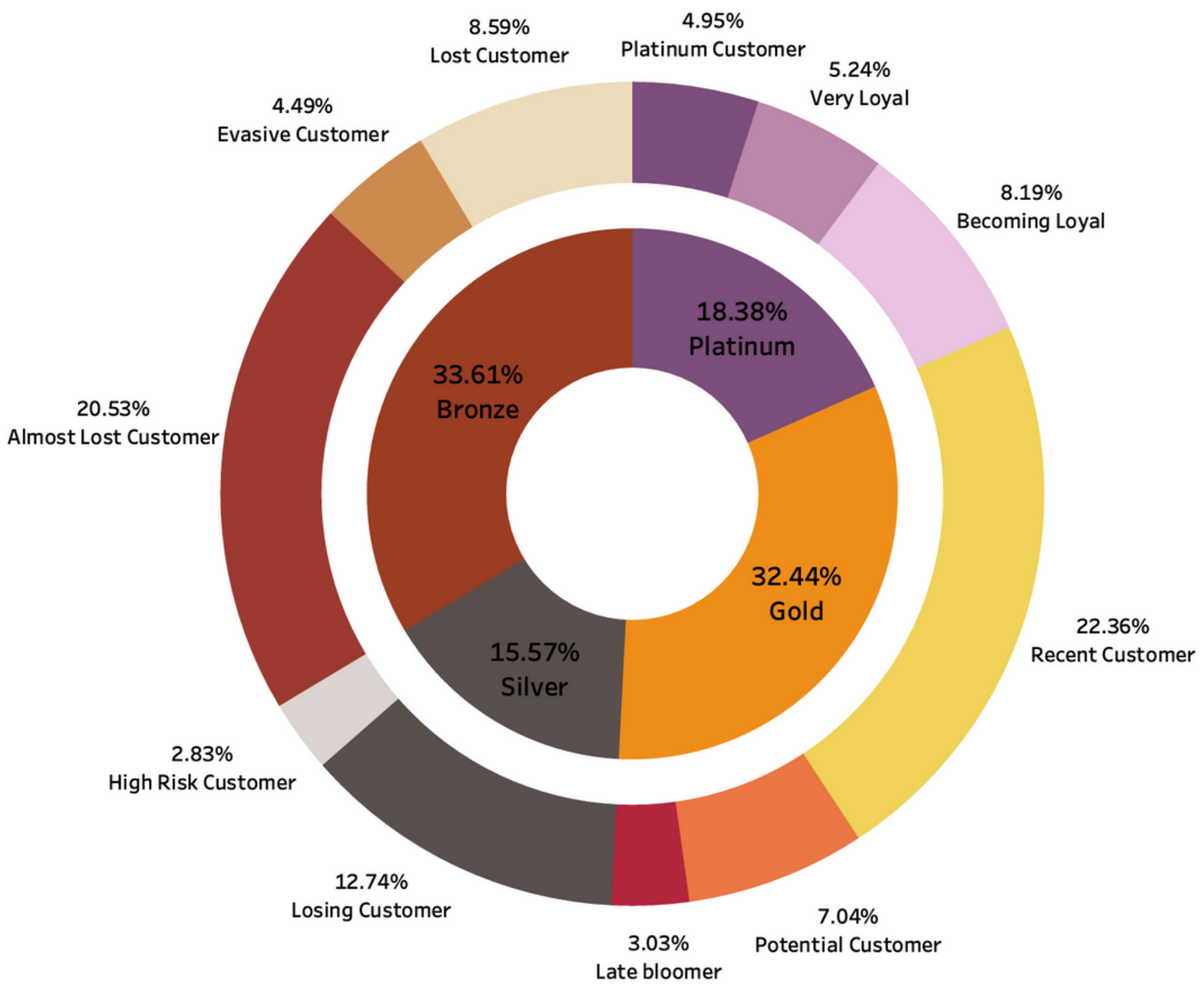
- Shallow frequency (0-2 number of purchases) correlates with high recency values.
- Customers that have visited more recently (0-50 days) have a higher chance of seeing more frequently(6+).
- Higher frequency has a negative relationship with recency values. Such that very recent customers are also frequent customers.



Source: Tableau

Back to Agenda

Model Development - RFM Analysis - Customer Title Distribution



Source: Tableau

Back to Agenda

Predictive Multi Classification Model

Data Preparation

- Format & Filter Data
- Split Data into training and testing datasets
- Standardise Data
- Label Encoding / One Hot Encoding

Sci-Kit Learn ML Models

- **Decision Tree Classifier**
- Random Forest Classifier
- SVC
- K-Neighbors Classifier
- Gaussian Naive Bayes
- Gradient Boosting Classifier

Github Repo: Quimbolos/KPMG_Internship

Decision Tree Classifier

Confusion Matrix

True Label \ Predicted Label	0	1	2	3
0	123	19	42	12
1	108	10	27	4
2	93	12	40	5
3	104	13	30	10

Metrics

F1 score - 0.21
Precision - 0.27
Recall - 0.26
Accuracy - 0.28

GridSearchCV Scoring Method

F1 Score - harmonic mean(average) of the precision and recall.

Best Hyperparameters

criterion': 'gini'
'max_depth': 5
max_features': 'log2'

Cross-Validation Method

Stratified K Fold

Back to Agenda

Findings & Recommendations

The Models perform poorly, currently making it unfeasible to classify new customer data into correct customer titles. However, the performance may improve with more resources allocated to the modelling. Things to look at would be:

1. Feature Selection and Engineering: Analyze and select relevant features for classification. Consider feature importance analysis and PCA for dimensionality reduction. Create new features to improve customer title distinction.

2. Data Encoding: Properly encode data before feeding the model. Use one-hot or label encoding or advanced techniques like embeddings for categorical variables.

3. Data Augmentation: Benefit from data augmentation with limited data. Generate synthetic samples through perturbations to improve model generalization and reduce overfitting.

4. External Data Sources: Incorporate relevant external data to enrich the dataset, providing valuable context and improving model performance.

5. Neural Network Models: Explore different architectures (CNNs, RNNs, Transformers like BERT or GPT-3) for specific data types (images, sequential, NLP).

6. Hyperparameter Tuning: Optimize model performance by fine-tuning hyperparameters using grid search, random search, or Bayesian optimization.

7. Regularization and Dropout: Implement regularization techniques (L1, L2) and dropout layers to prevent overfitting.

8. Cross-Validation: Assess model generalization using cross-validation, identifying overfitting and data leakage issues.

9. Ensemble Methods: Combine predictions from multiple models (bagging, boosting) to create a stronger ensemble model.

10. Hardware and Parallelism: Use powerful hardware and parallelism (multiple GPUs, distributed systems) to accelerate training and experimentation.

11. Data Preprocessing: Correctly preprocess data, handling missing values, and outliers, and scaling numerical features as needed.

Back to Agenda

Interpretation

Once the Model is completed, it can be used to classify customers in the new customers dataset and report it to the marketing team to devise their strategy

Back to Agenda

Appendix

Back to Agenda

Appendix

Back to Agenda

SPROCKET CENTRAL

Get In Touch

Back to Agenda

Email

bolosfernandez@hotmail.es

GitHub

Quimbolos/KPMG_Internship