



Sprocket Central Pty Ltd

The Analytic Team
Data analytics approach

Analyst Blessing. B.

Outline

1. Introduction

- ★ Brief Overview of Problem
- ★ Aims and Objectives

2. Data Exploration

- ★ Data Pre-processing
- ★ Feature Engineering

3. Exploratory Data Analysis

- ★ Tables
- ★ Graphs

4. Model Development

- ★ Pareto Principle
- ★ Modeling Techniques.

5. Conclusion and Recommendation



Introduction

Brief Overview: Sprocket Central Pty Ltd specialises in high-quality bikes and accessible cycling accessories to riders. Their marketing team is looking to boost business by analysing their existing customer dataset to determine customer trends and behaviour. Three datasets (Customer demographic, customer address and transactions) as a labelled dataset, were given.

Aims and Objective: 1000 new customers set was also given separately and the objective of this analysis was to recommend which of these 1000 new customers should be targeted to drive the most value for the organisation. In building this recommendation, the approach is detailed in subsequent slides.



Data Exploration

Data pre-processing: The three data set has a common column called the customer_ID, hence, merging was done using that column. After merging, we had a total of 19968 rows and 30 column. With variables such as name, gender, list price, standard cost, job title, e.t.c. The data consist of both numerical and categorical variables. Checking for missing data, it was discovered that, the maximum missing entry was 3,222 in variable “job industry category”. We had the option of handling that by filling it up with the mode of that column, however, it might lead to bias for the other levels in that column, hence, we opted for the option of deleting those rows entirely, since we have much data left.

Feature Engineering:

- ★ Converting the DOB to age
- ★ Created a new column called “transaction_day” and “transaction_month” from the original column “transaction_date”
- ★ Created a new column called “profit” from the difference between “list_price” and “standard_cost”

Feature Engineering:

- ★ Created a column called “cus_count”, this is the frequency count of each customer visit to the organization.
- ★ Created a new column called “focus”, this are 2 categories of people, the column identifies customers that contributes to the biggest profits of Sprocket Central Pty Ltd. details would be given shortly.

Exploratory Data Analysis (Tables)



Table 1, is the count of customers wealth segment as well as their list price. From the table, we see that the mass customers has the highest count as well as the highest sum of list price. This is worth looking into by Sprocket Central Pty Ltd.

Table 2, on the hand shows the count of customers wealth segment, brands with mean of their property_valuation. From the table, we can deduce that the affluent customers has the highest property valuation across all the brands. The High Net Worth had the least property valuation.

Table 1: The count of customers wealth segment and list price

Wealth Segment	Count	List price
Affluent Customer	3237	3591802.42
High Net Worth	3285	3715639.90
Mass Customer	6324	7077771.89

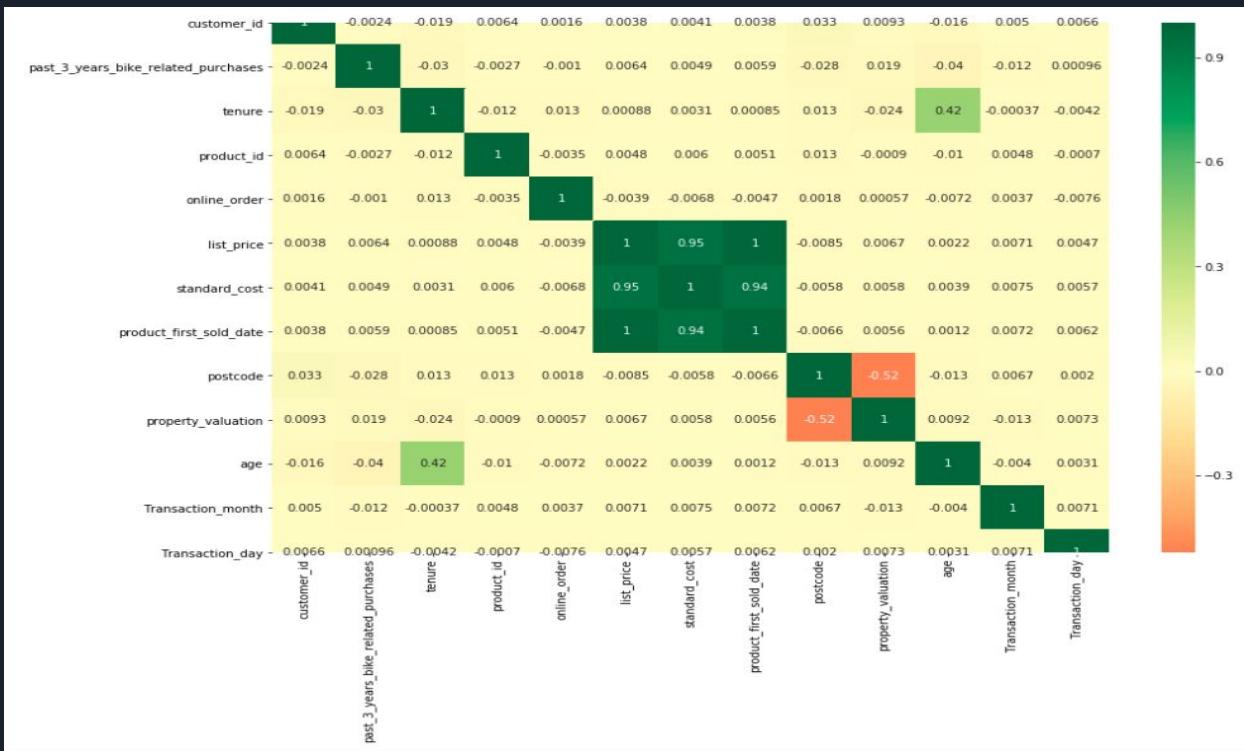
Table 2: The count of customers wealth segment, brands with mean of the property_valuation

Wealth Segment\Brand	Giant Bicycles	Norco Bicycles	OHM Cycles	Solex	Trek Bicycles	Weare A2B
Affluent Customer	8	8	8	8	8	8
High Net Worth	7	7	7	7	7	7
Mass Customer	7	8	7	8	7	7

Exploratory Data Analysis (Graphs)

The chart beside is a **heat map** showing the correlation (of all the numerical variables in the data set). From the graph is can be seen that most of the variables are not correlated with each other, however, standard cost and list price has a very strong positive correlation of 0.95. Also, age and tenure are moderately correlated with value 0.42. Post code and property evaluation has a moderately negative correlation of -0.52.

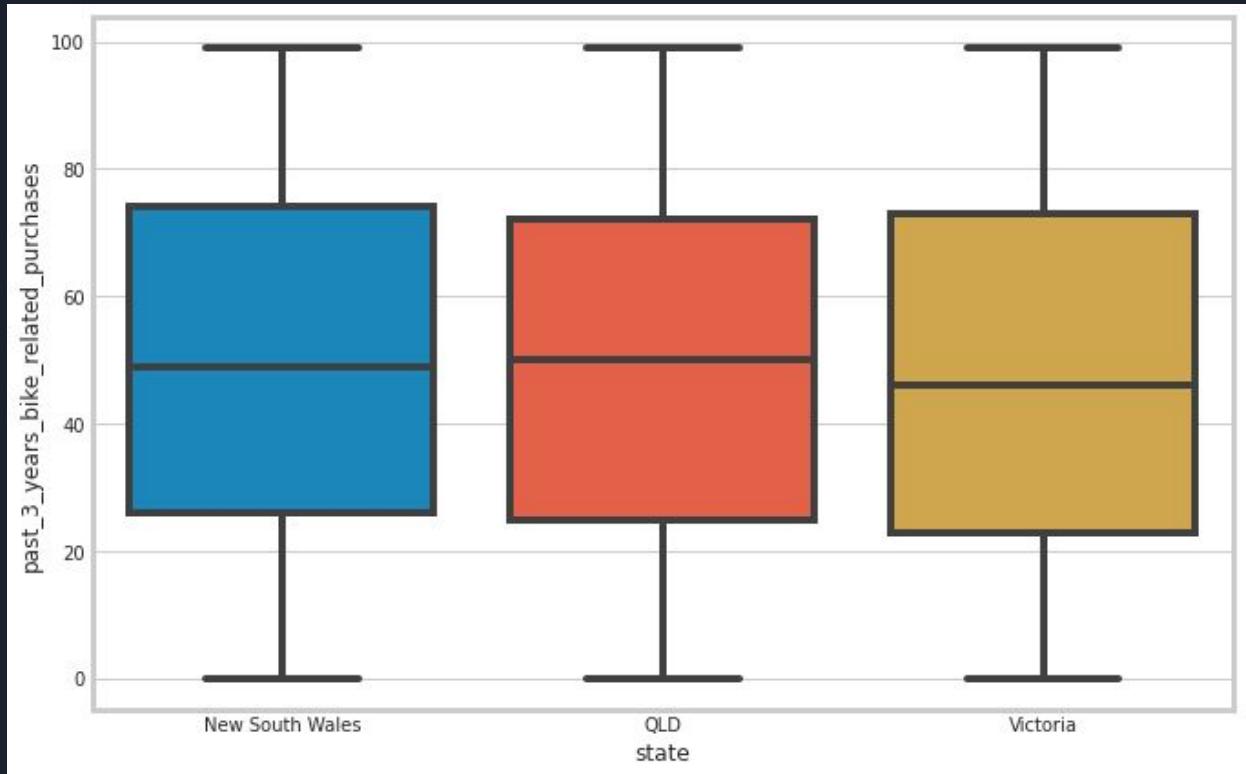
Figure 1: Correlation plot of all numerical variables



Exploratory Data Analysis (Graphs)

Figure 2: Customers past 3 years bike related purchases by state

This is a **bar chart** displaying the distribution of the number of bikes customers bought by state. From the plot, we can infer that the most bike purchase came from New South Wales, with the mean purchase of about 50 bikes. The least purchase came from Victoria with a mean of about 45 bikes. We also observe that, the variance across each of the state are almost equal. Well, they may look so much the same, it would be nice to see if there is a statistical significant among the states.



Exploratory Data Analysis (Graphs)

Figure 3: ANOVA test to check for the statistical difference in the number of bikes customers bought among the 3 states.

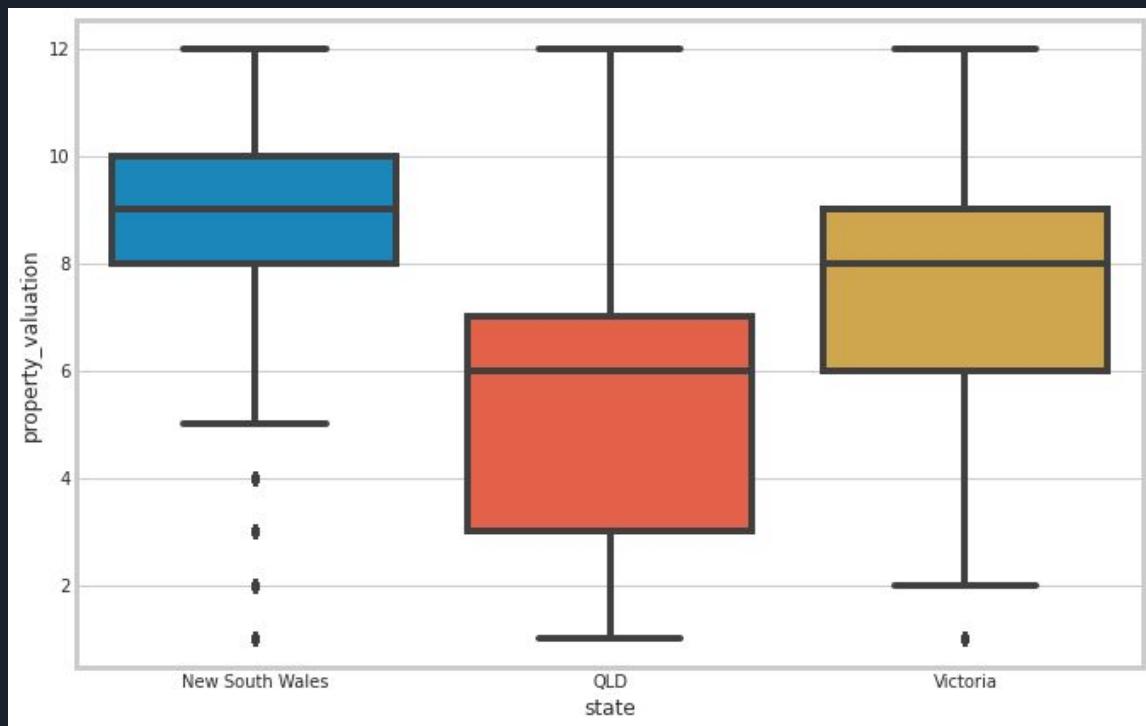
	sum_sq	df	mean_sq	F	PR(>F)	eta_sq	omega_sq
state	7.143678e+03	2.0	3571.839148	4.355871	0.01285	0.000678	0.000522
Residual	1.053133e+07	12843.0	820.005744	NaN	NaN	NaN	NaN

The **ANOVA test** known as “Analysis of Variance” is a parametric test used to test for a statistically significant difference of an outcome between 3 or more groups. From the previous slide, we observed that the number of bike customer bought among the three states were somehow close, we needed to see if this closeness was statistically significant. Above is an ANOVA test across the three states, from the table, we notice that the p-value is 0.01285 which is < 0.05 (95% confidence level), hence, we reject the null hypothesis and conclude that there is a statistical difference in the number of bikes bought across the three states in the past 3 years. The omega_sq is known as the “eta” value, that is, the effect of the effect size, this is to tell us if the statistical significant is meaningful. From the value above, 0.0000522 seems small, hence we can say that, although it is significant, but the effect is small.

Exploratory Data Analysis (**Graphs**)

Figure 3: Customers property valuation by state

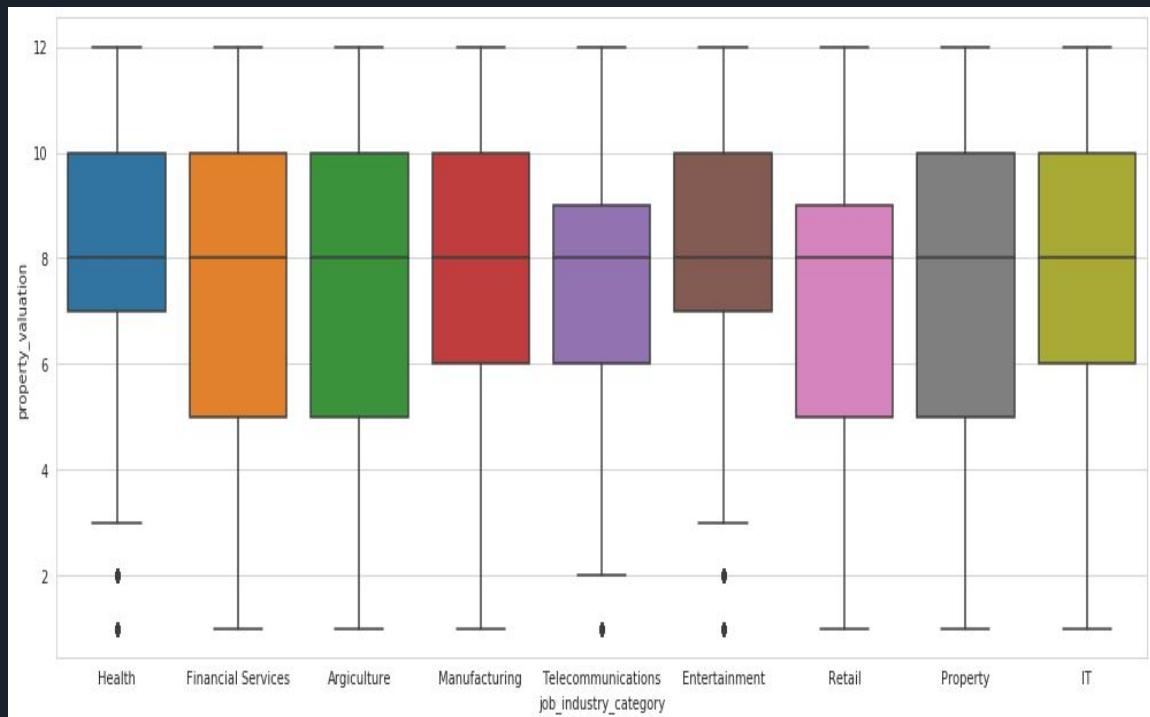
This **bar chart** displays the property valuation by state. As seen from the plot, we can deduce that New South Wales has the highest property valuation with the least variance as well. However, they have four outliers with low property valuation. QLD has the least property valuation as well as the highest variance across the property valuation. Aside New South Wales, Victoria also have a high property valuation.



Exploratory Data Analysis (Graphs)

Figure 4: Customers Job title by state

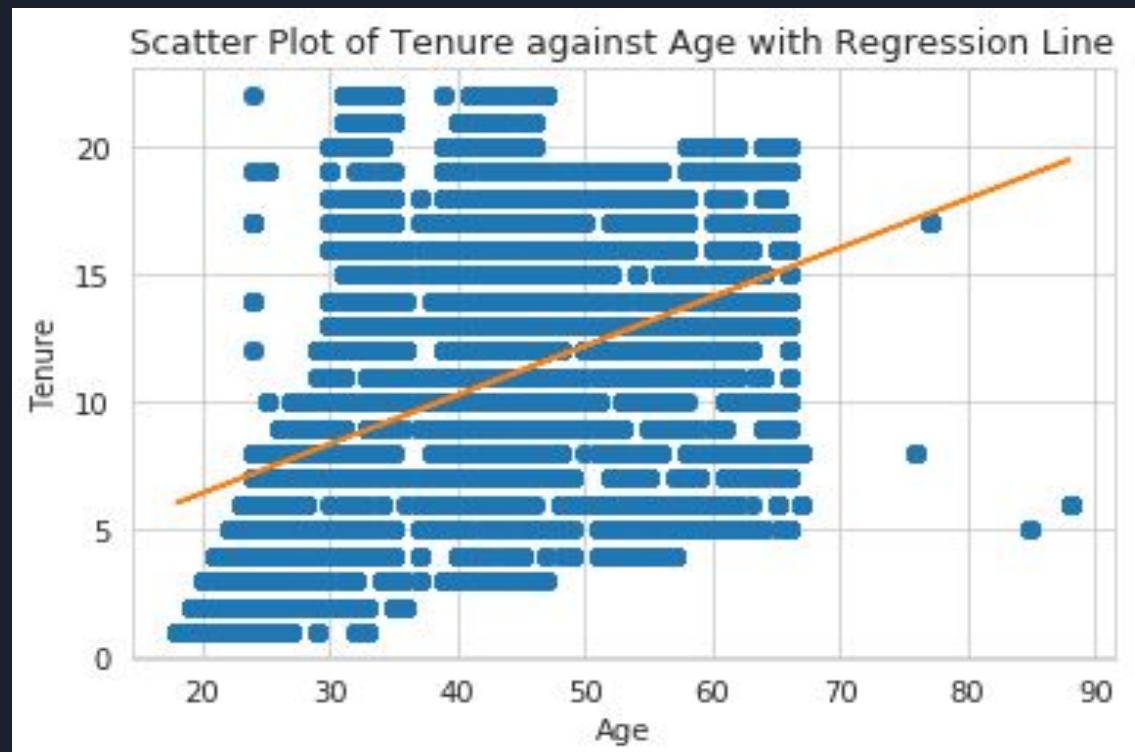
This **bar chart** displays the job title by state. As seen from the plot, we notice that the mean looks somewhat the same across the job titles, however, there are variations across the job title. However, Health, Telecommunication and the Entertainment sector has the least variations, but Health and Entertainment has more customers with higher property valuation. Financial Service, Agriculture and Property sector has the same variations across them as well, with few customers above property valuation 8.



Exploratory Data Analysis (Graphs)

Figure 5: Relationship between customers age and tenure

The **scatter plot** displays the relationship between customers age and tenure. From the plot, we see that, there is a moderate positive linear relationship between customers age and tenure. It is also noticeable that most of the customers are within the ages of 25 and 68. We also have few customers above 70 years of age.



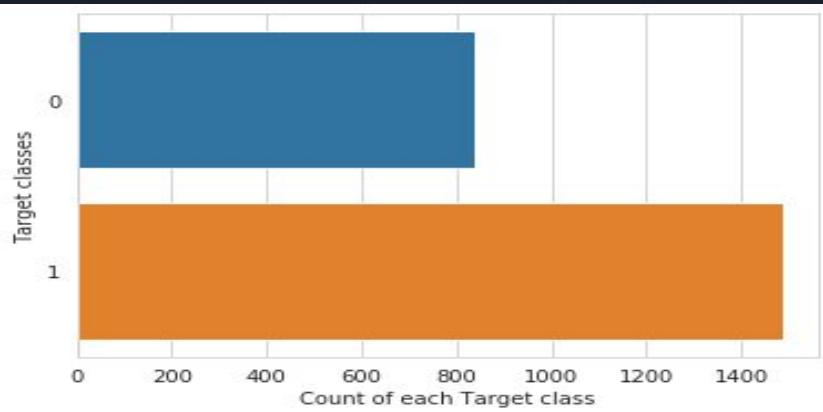
Model Development (**Pareto Principle**)



Pareto Principle:- According to legend economist, Pareto, he noticed 20% of the pea pods in his garden provided 80% of the peas. He then determined 20% of the population in Italy owned 80% of the land. The use of the 80-20 rule has since expanded beyond the alleged humble beginnings in Pareto's garden. Hence, in business, the pareto 80/20 rule, states that for many phenomena 80% of the result comes from 20% of the effort, meaning that, Sprocket Central Pty Ltd gets 80% of their income from 20% of their customers.

Now, the big question is, who are the customers that contributes to the 80% profit made by Sprocket Central Pty Ltd. To answer this, we need to create a new column called “profit made” (as discussed in slide 03). From this column we can then sort and have customers with the highest profit on top. We then calculate those whose profit made up 80% of Sprocket Central Pty Ltd income. These ones are categorised as group 1, while those who contributed to 20% of Sprocket Central Pty Ltd profit are categorised as group 0. This forms the basis for the classification tasked (A supervised learning problem). This column is called **focus**.

Model Development (Pareto Principle)



Here is the categories of people Sprocket Central Pty Ltd should focus on. From the screenshot on the right, we can see the Sprocket Central Pty Ltd made a profit of \$8446.55 from the customer with id 1913, we can also see that the group of customers focus 1 contributed to 80% of Sprocket Central Pty Ltd profit. The count plot above shows that more of the customers (1490) contributes to 80% of their profit.

	customer_id	profit_made	focus
0	1913	8446.55	1
1	2476	8370.19	1
2	2659	8131.46	1
3	2183	8115.96	1
4	1227	7636.98	1
...
2322	3168	312.26	0
2323	2115	311.95	0
2324	2047	258.82	0
2325	130	258.80	0
2326	2423	237.94	0

Model Development (Modelling Techniques)



Before, building a predictive model, we needed to work on the data structure as we had both categorical and numerical variables. The categorical variables were converted into numerical via the label encoder from Sklearn.

The gender variable which had inconsistent level i.e. m, male, femal and female were all replaced by Male and Female for consistency. Once we had all variables in numerical, there was a need to normalize the data, cause we had the data in different scale. I.e gender as 0 and 1 while list price was in thousands. Hence, we need to make the data have a consistency, i.e, a common data format of mean 0 and standard deviation of 1. This is achieved with the function StandardScaler() in from Sklearn.

Since, it's a classification problem, different machine learning techniques were explored as well as a deep learning approach.

Model Development (Modelling Techniques)

The different machine learning models tried as well as their accuracy on test score are as follows:

★ Linear Regression

★ Support Vector Machine

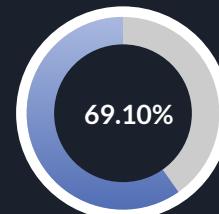
★ Decision Tree

★ Random Forest

★ Gaussian Naive Bayes

★ Gradient Boosting

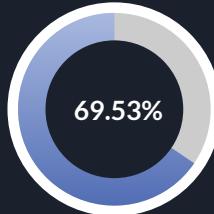
★ Deep Learning approach



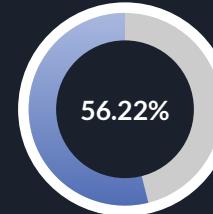
Gradient Boosting



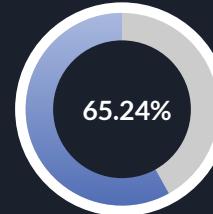
Linear Regression



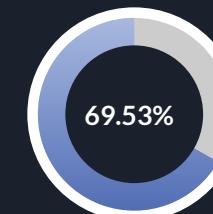
Support Vector Machine



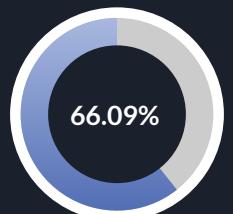
Decision Tree



Random Forest



Gaussian Naive Bayes



Keras deep learning approach



Prediction

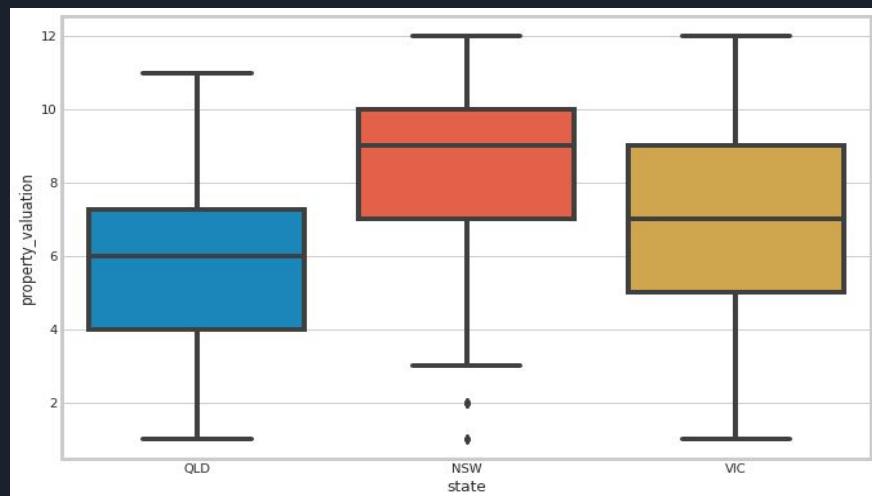
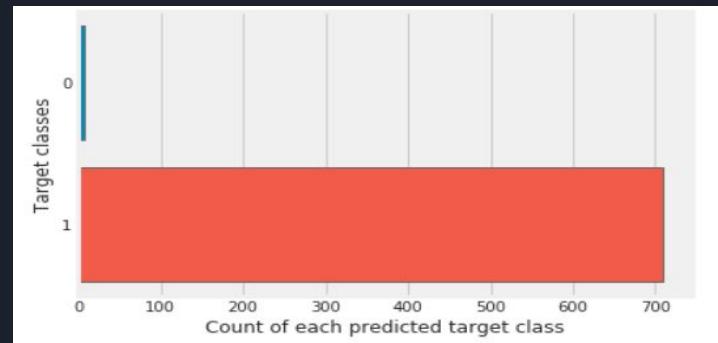
Finally, the results of the different classifiers (with accuracy score of 69.53%) presented in the previous sections were combined to improve the classification model. This can be achieved by selecting the customer category as the one indicated by the majority of classifiers. To do this, I use the `VotingClassifier` method of the `sklearn` package. As a first step, I adjust the parameters of the various classifiers using the best parameters previously found. The final score which was 69.53% was used to predict the 1000 new customers in Sprocket Central Pty Ltd.

Before any prediction could be done, I preprocessed the 1000 new customers datas as I did for the existing customers, after that, I normalized it as well, and then made prediction.

Conclusion

From the prediction on the right, we can see that majority of the new customers will contribute to the profit of Sprocket Central Pty Ltd by 80%. Only few will fall into the categories of 20%

As confirmed by the first transaction data set on slide 08, we also see that NSW has a great influence in predicting the profit made by Sprocket Central Pty Ltd, hence, Sprocket Central Pty Ltd should focus more on customers from this state. This is because they have the highest property valuation, even it's 25 percentile is the median property valuation of customers from Vic. This is not the only factor, all others also contributes in predicting the customer focus at Sprocket Central Pty Ltd.





Recommendation

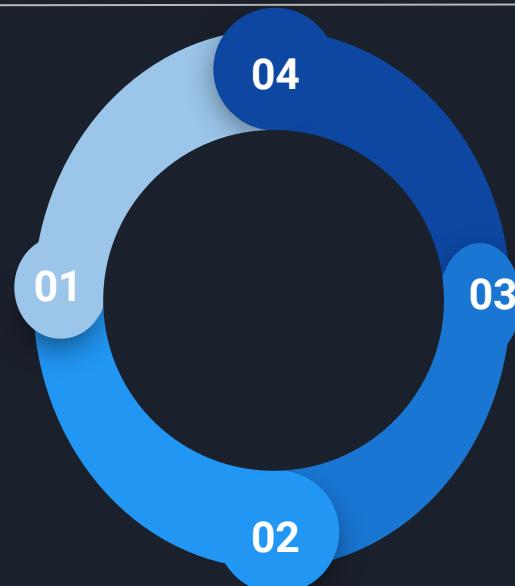
All the variables played their individual roll towards predicting the customer focus at Sprocket Central Pty Ltd , however, more attention should be place on variables listed below

State

As established before, customers from New South Wales should be targeted, as they have the highest property valuation.

Wealth Segment

The affluent customers as well as the mass customers contributed immensely to the profit made in Sprocket Central Pty Ltd, they should be targeted. See slide 04, table 2.



Job Title

We deduce that customers in the health and entertainment industry has the highest property valuation. These customers should be targeted.

Focus

Following the pareto principle, customers whose transactions contributes to 80% profit of the organization should be targeted.

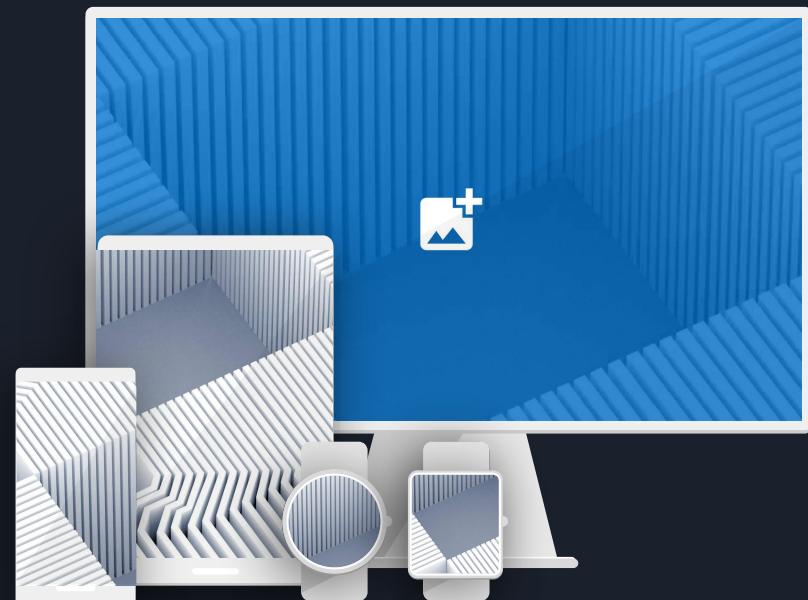
Future work

Below are some observations and proposed future work that could be done.

For further analysis, we would like to explore the task as an unsupervised learning problem and cluster customers based on the feature in the data set.

Also, for future analysis, it would be better if we have more features like the the customers income, number of item purchased, e.t.c

The full detail of all the variables were not provided, it was a little challenging knowing what each of the variables represent.



Thank you!

Essential Cycling Accessories



CYCLE
RACING



©2020



©2020



©2020

