TheAnalyticsTeam

Sprocket Central Pty Ltd

Data analytics approach

[Business Intelligence] - [George Maxwell], [Tony Smith], [Prashant Lal]

Agenda

1. Introduction
2. Data Exploration
3. Feature Engineering
4. Feature selection
5. Model Development
6. Interpretation

Introduction

Brief introduction about steps

The dataset provided are quite messy with lots of data are not in the right format to feed into machine learning model. Preprocessing the dataset is a necessary concern.

Exploratory data analysis helps us to get to know more about the data and the information contained in it.

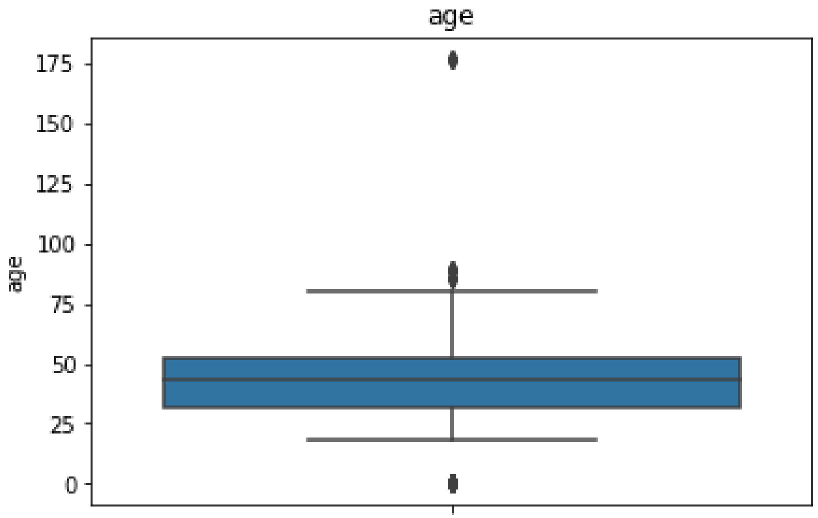
Featuring engineering helps to extract those information which are not readily available.

In feature selection, we select relevant attributes from irrelevant columns.

After cleaning and preprocessing the dataset, we can build the model, interpret the result based on the observed pattern and could address business problems.

Data Exploration

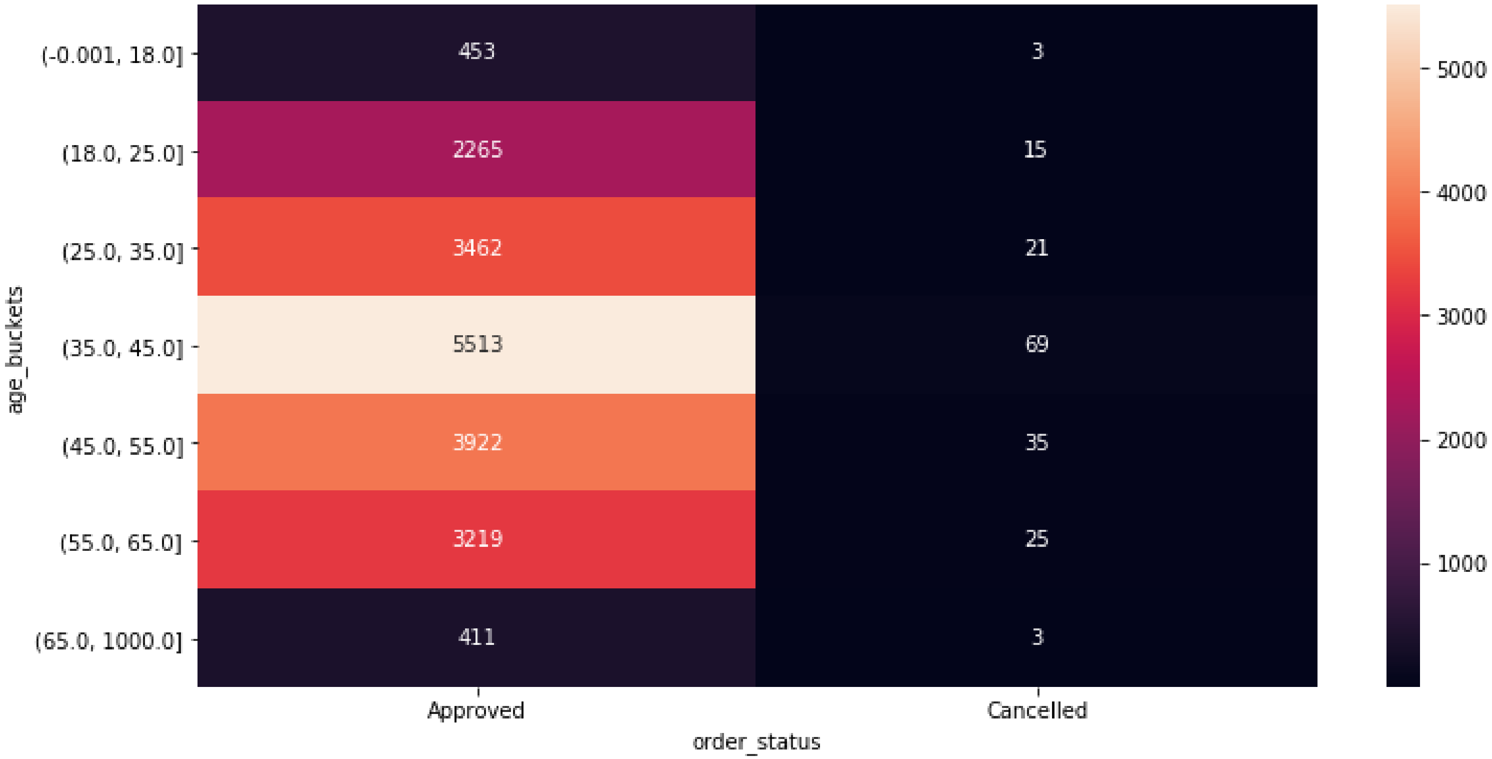
Checking of consistency of each columns , i.e. should be free from missing values. For example missing values imputation of independent variables based on their correlation with other variables or checking certain trend in the particular attributes and based on that imputing the missing space. Knowing the percentage of data missing and if less than 1 percent occurs then better to remove it.
Forming of correlation matrix among different numerical attributes to know the presence of duplicated columns or columns describing the same information or the interaction between different variables.
Understanding the distribution of continuous variables, checking the skewness and plotting the boxplot to obtain the mean, median and inter quartile range on which the attributes are distributed. Inferred from these analysis whether discretization of continuous variables is a better choice of not.



Boxplot also helps to detect the possible outliers. Further analysis of assumed outliers and cross checking on statistical evidence and preferred domain knowledge to deal with such values. If found technical error, can be relabeled or possible the removal of records, if found by doing so, we are not loosing much valuable information.
Forming the contingency tables across different attributes to get the insight about various factors among different categorical variables which shows strong interaction with other variables labels . For example frequency table between “Job industry category” and “Brand” showed that most in “manufacturing” sector prefer “Solex” brand .
Checking the cardinality of categorical variables, counting their percentage in the whole dataset and making correction of those categorical variable which are having different labels for the same category.
Forming word cloud to know which are the most assigned job title or which street area address most customers are from

Feature Engineering

Extraction of more features from different attributes.
Knowing the pattern hiding behind it. Getting extra information from outside of the dataset are some of the feature engineering steps
For example getting the age of the customers from their Date Of Birth, Discretizing the tenure, property valuation and age columns into small buckets to extracts valuable information by looking into their interaction with different variables and addressing the common business problems.
Removing some of the columns whose values are un explainable, unnamed columns or does not contain relevant data to address common business problems
Numerical encoding of the categorical variables and adding binary indicator columns for each numerical encoded columns .



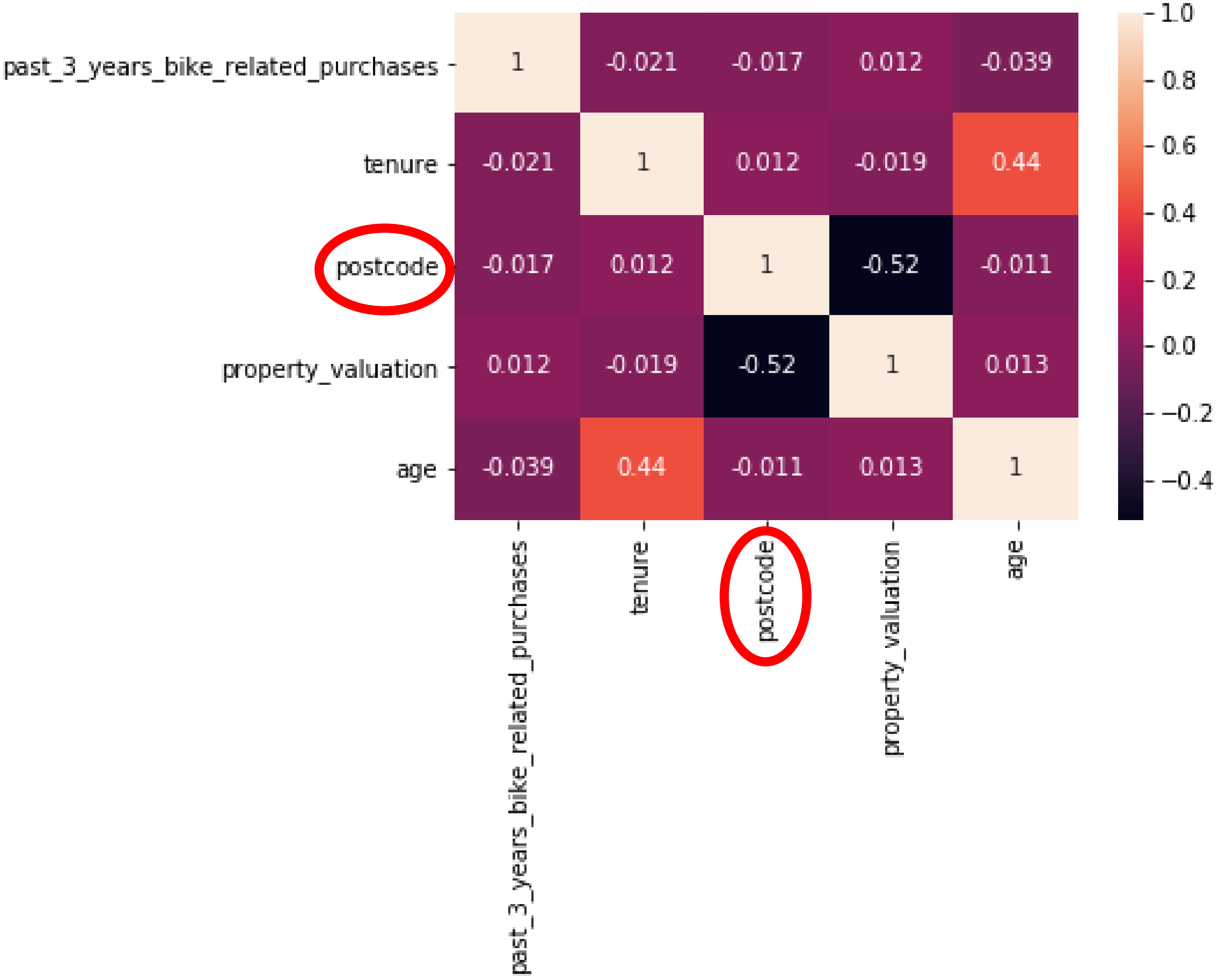
Feature Selection

We get to know the importance of each variables in Exploratory Data analysis and engineered some of the variables in feature engineering section. In feature selection we choose those columns which are most relevant for model building, training and deploying into production

For example we have age bucket columns and age columns, therefore picking the age bucket column and dropping the other.

Similarly for tenure and property valuation columns.

Selecting binary indicator columns and dropping the numerical encoded columns and it's original columns .



	past_3_years_bike_related_purchases	tenure	postcode	property_valuation	age
past_3_years_bike_related_purchases	1	-0.021	-0.017	0.012	-0.039
tenure	-0.021	1	0.012	-0.019	0.44
postcode	-0.017	0.012	1	-0.52	-0.011
property_valuation	0.012	-0.019	-0.52	1	0.013
age	-0.039	0.44	-0.011	0.013	1

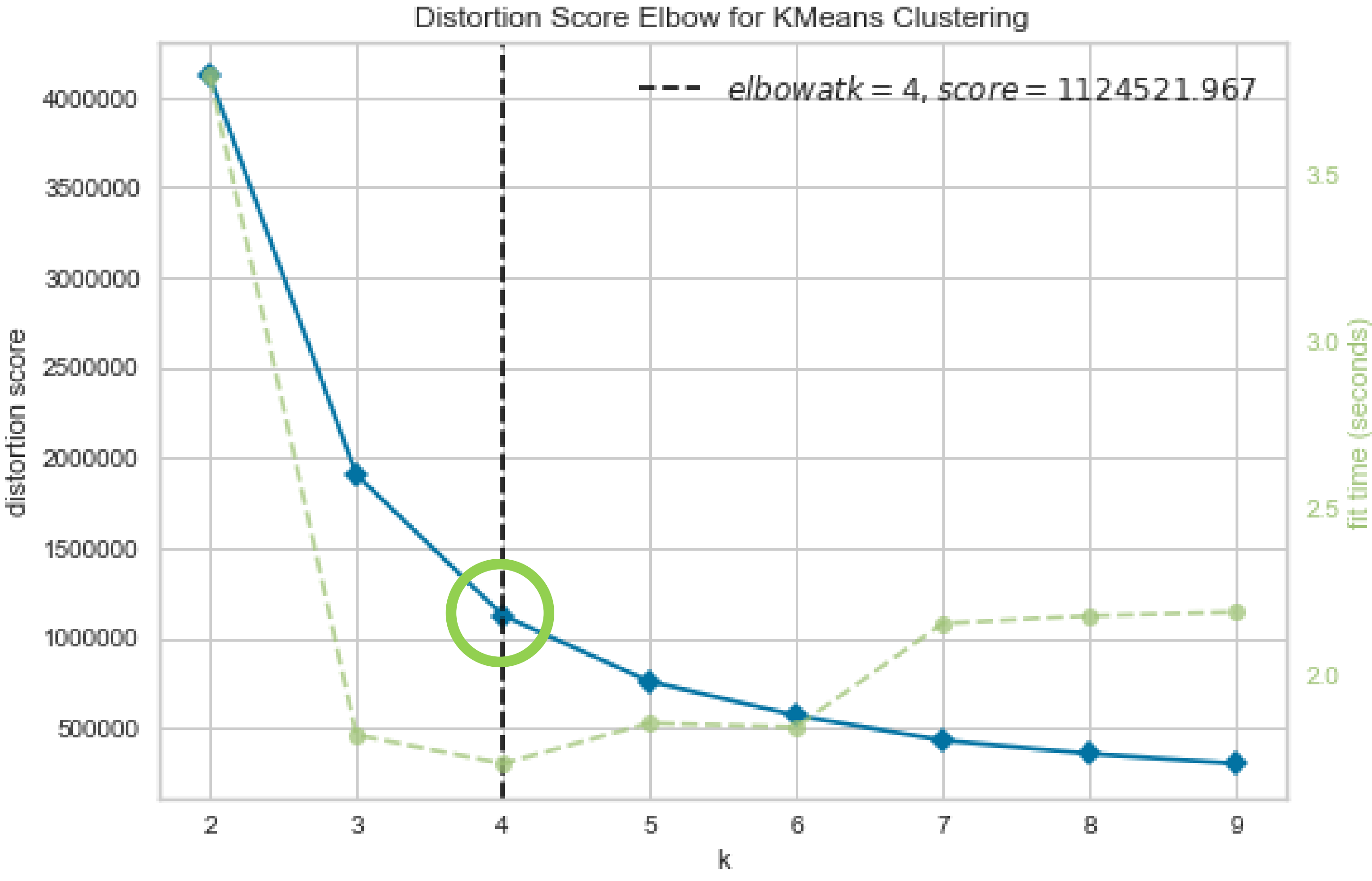
Model Development

Machine learning model building and implementation

Since the problem is of customer segmentation, so we are using k mean clustering algorithm to build our machine learning model.

This algorithm clusters those values of attributes which shows similarity with each other and labels each cluster into separate groups.

We will use Elbow method to get the best number of clusters.

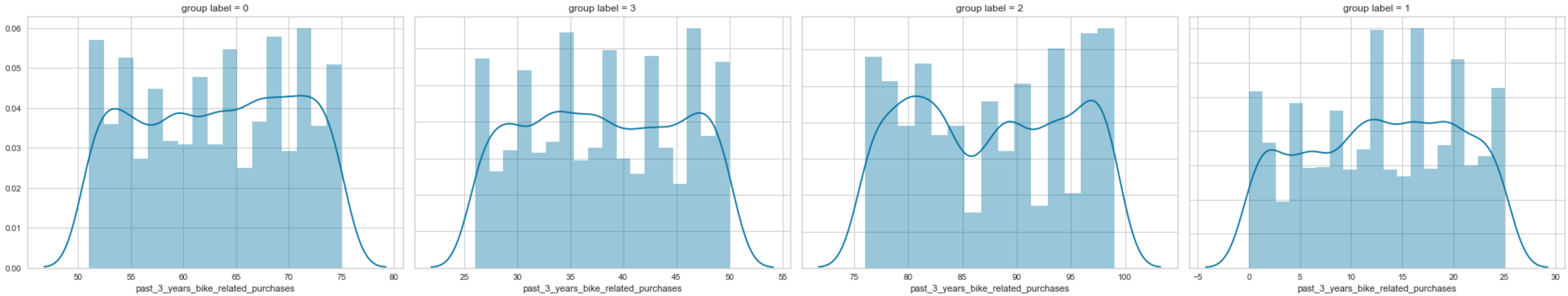


Interpretation

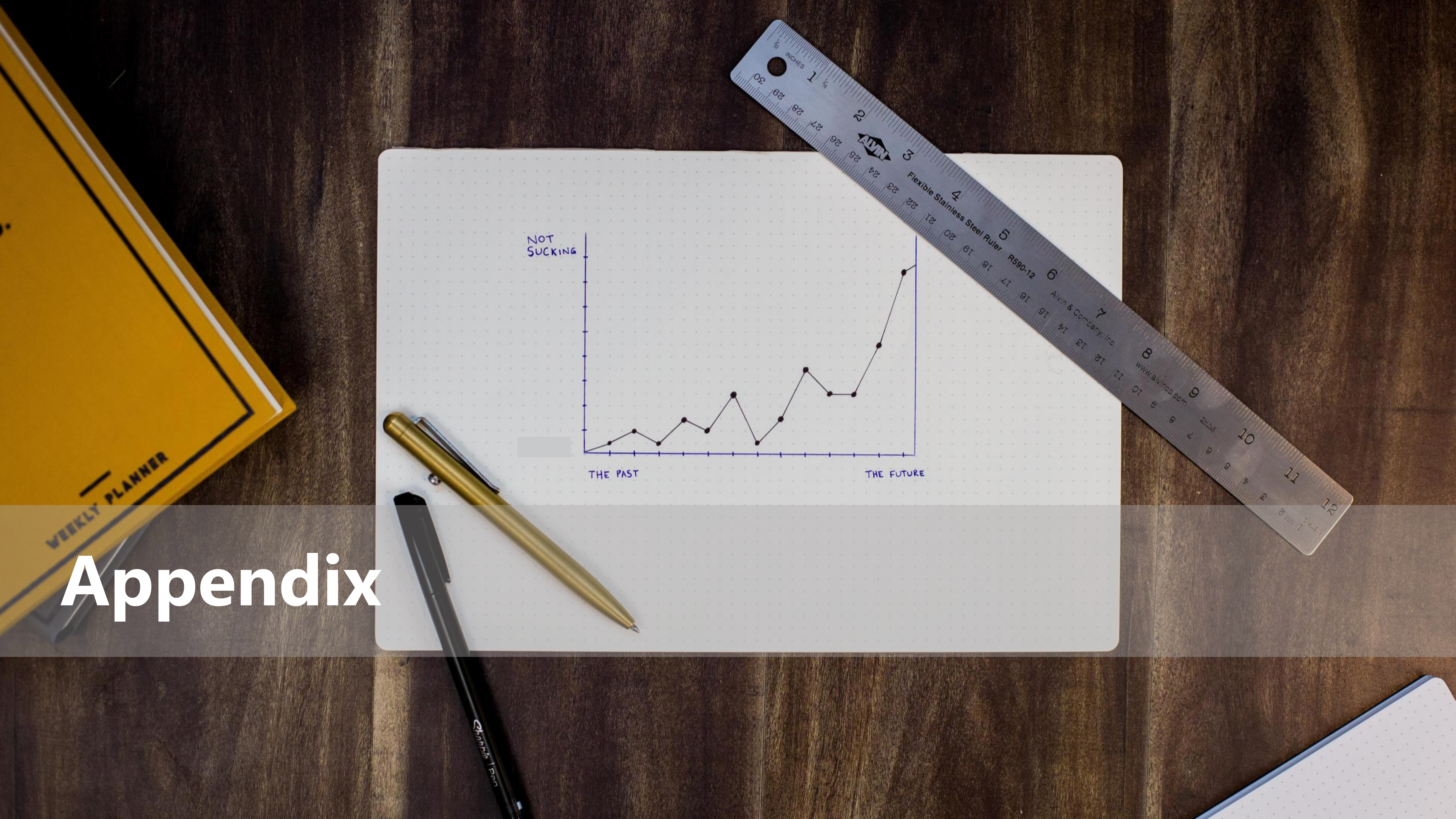
Post model building

After the machine learning model is build, we have to interpret the results based no clustered we formed. Going through each group, their attributes and how they distinguish from other groups we can answer some of the most complex problems related to business development.

For example group one’s past 5 years bike related purchase attribute ranges form 50 to 75 and group two ranges from 0 to 25 for different age buckets.



Appendix


NOT SUCKING
THE PAST
THE FUTURE
WEEKLY PLANNER
Flexible Stainless Steel Ruler
R590-12
Alvin & Company, Inc
www.alvinco.com

Appendix

All the analysis have been done on the dataset provided by **Sprocket Central Pty Ltd** on Python language (version = 3.8, anaconda environment) in Jupyter notebook as IDE

Packages used for analysis : NumPy, Pandas, Matplotlib, Seaborn
Package used for Model building : Scikit Learn