Report 3

**APPLYING DATA ANALYTICS IN CUSTOMER SEGMENTATION**

Toan Diec - 30032064

CSIS 4495 - 090

Table of Contents

**1. Statement of problem3**

**2. Significance of the Study4**

**3. Environment Setup5**

**4. Project Implementation7**

Data Processing and Wrangling9

Feature Engineering13

Univariate Analysis18

Exploratory Data Analysis 19

Customer Segment with Feature Selection, Feature Scaling and Transformation…………………………… 22

Feature Selection 22

Modeling ………………………………………………………………………………………………………………………………….23

Model Evaluation 27

Performance Indicator………………………….….……………………………………………………………………………….27

Interpret Cluster…………….…………………….….……………………………………………………………………………….29

**5. Proposed Revisions 20**

*Abstract*

Customer segmentation is the most common way to separate customers into groups that have same features such as behaviour, age, gender, etc. The expected result of this process is to identify high value customers for different marketing campaigns or general strategy and planning.

This project focuses on how to apply data analytics knowledge, skills, and tools that I have learnt during my time at Douglas college to discover customer segments from a raw dataset. To successfully achieve the outcome, the pipeline of data analytics from data processing, feature engineering, feature selection, modeling, model evaluation, and model deployment will be applied.

Besides the expected results of customer segmentation, the project also aims for the theories and insights behind the outcomes. Hence, the project is not only included a technical side of data analytics, but also includes important concepts and insights of business for deeper understanding of the reason why customer segmentation is necessary.

1. Statement of problem

Customer segmentation has a lot of benefit for businesses when applied correctly. From the customer segmentation, decision makers can develop more effective strategies for targeting the customers, which plays a major part in businesses succeed. For instance, company with high-end product, budget product, or a cheap product has different way to approach its own customer such as highly spend customer, frequent purchase customer, or low-value customer. By understanding and segmenting the customer, company could adjust the feature of its product to match with customer’s needs. Customer segmentation can also help a company to understand how its customers feel about the products, what is the important feature to them, and what is not. Finally, improving customer service, price testing, and upselling and cross-selling other product or services are common benefits when applying customer segmentation.

For a small size business, it is easier to learn and target a specific customer. However, when the business gets bigger in scale, it is very insufficient to target every individual in the customer list with human judgments. Therefore, to assist the human part of the equation, data-driven approach is crucial to large or even mid-size businesses. Using data-driven approach to segment the customer not only help business to understand and target new customers, but also to retain the current customers by discover any potential factor that increase customer churning rate.

In this project, a raw dataset will be utilized to apply the process of data analytics pipeline. Data processing, feature engineering, feature selection, modeling, model evaluation, and model deployment will be applied to generate an appropriate customer segment and gain some insights behind the results. To achieve the desired outcome, the project has the approach of batch learning and tries some different techniques in data analytics such as supervised learning or unsupervised learning to find out the most accurate model. By the end of the project, the project will not only give the reader the understanding of how to handle customer segmentation using data-driven approach but also provide an insightful information behind the results.

1. Significance of the Study

By applying data analytics skills and tools to analyze the dataset from a real-life business, organizations gain a great deal of strategic value from customer segmentation studies, but in order to exploit that value, researchers must go beyond standard reports and find novel ways to make segmentation more accessible and actionable.

This project provides a clear step by step of how to conduct customer segmentation by using standardize data analytics pipeline and many other data techniques. Moreover, insights and information behind the results will also be discussed throughout the project to give a reader deeper understanding. Finally, the project includes many concepts not only from data science viewpoint but also from the business viewpoint.

Achieving customer segmentation will give business a great advantage of information to target not only new but also current customer more efficient.

1. Environment Setup

To successfully run the code in the project, reader need to install Jupyter Notebook or Anaconda software which already included Jupyter Notebook. The documentation of how to install the software for various operation system can be found with the following link: [Installation — Anaconda documentation](https://docs.anaconda.com/anaconda/install/index.html) .

After successfully install Anaconda, open Jupyter Notebook

Graphical user interface, application

Description automatically generated

Then find the file *Customer\_Segment(ToanDiec\_300320364),* open it and click Run all. It might take 10 to 15 minutes, depend on the computer system, to completely run all the code.

A screenshot of a computer

Description automatically generated with medium confidence

1. Project Implementation

In this report, I will report my project’ progress for data preparation phase in the data analytics pipeline.

Diagram

Description automatically generated

After searching for a suitable dataset for implementing data analytics in the project, the Brazilian E-Commerce Public Dataset by Olist is utilized for further analyses. This is a real commercial data included 100 thousand of order information that made at difference marketplaces in Brazil from 2016 to 2018. The dataset has the following data schema as below and can be found using this link: <https://www.kaggle.com/olistbr/brazilian-ecommerce/home> .

Diagram

Description automatically generated

**Data Processing and Wrangling**

Graphical user interface, application, table, Excel

Description automatically generatedFor the purpose of the project, the dataset of olist\_order\_dataset, olist\_order\_item\_dataset, olist\_order\_customer\_dataset, olist\_product\_dataset are joined together and cleaned by using Tableau.

But before joining the dataset, I found out that the product category name of the olist\_product\_dataset is recorded in Brazilian language. So, I have to use a technique in previous data analytics course to translate the product category name into English using Tableau. Graphical user interface, application, table, Excel

Description automatically generated

After that, using Tableau, I joined all the dataset into one big table for the next steps. Even though I could join all the data using pandas and python, but I found it easier and more descriptive when using Tableau instead.

Graphical user interface

Description automatically generated

This is the raw dataset in form of Excel.

Graphical user interface, application, table, Excel

Description automatically generated

After using Tableau to join all the dataset, I use Jupyter Notebook to continue with the project.

Text

Description automatically generatedFirst, loading the required library and have a glance of the dataset.

Graphical user interface, text, application

Description automatically generated with medium confidence

Table

Description automatically generatedThe raw dataset included 17 columns and 117601 rows.

**Feature Engineering**

Graphical user interface, text

Description automatically generatedAfter that, I checked for duplicated entry in the dataset. I found out that there are more than 30 thousand duplicated rows in the dataset.

After went through the variable description, I understand that there are some customers order the same product multiple times with difference quantities, which is totally normal in this circumstance. Because of that, when I used Tableau to join several datasets with order dataset, Tableau automatically generate more rows to match with the data in order dataset. However, deleting duplicate rows would be a mistake.

So, I decided to create a new data frame included the sum of the order quantity and took the first value that appears in the Price column and Freight\_Value column. This gives me a new data frame without any duplicate Order\_Id.

Table

Description automatically generated with medium confidence

Moving forward, I merged the new data frame with the data frame at the beginning using Order\_Id as index.

Table

Description automatically generated

Even though it gives me duplicate rows just like the problem I have in the beginning, but now I can confidently remove duplicated rows and keep only the first version of it. Graphical user interface

Description automatically generated with low confidence

Next, I want to change columns that have category value such as Customer\_City, Payment\_Type, Order\_Status into numerical values and columns such as Customer\_State, Order\_Status, Payment\_Type into several columns with dummy values of 0 and 1 for further analysis.

Graphical user interface, text, application

Description automatically generated

**Graphical user interface

Description automatically generated with low confidence**

Lastly, transforming the Order\_Purchase\_Timestamp into column that contains only numeric value. For example, value of 29/12/2017 will become 29122017. Then drop all unneeded columns.Graphical user interface, text, application

Description automatically generated

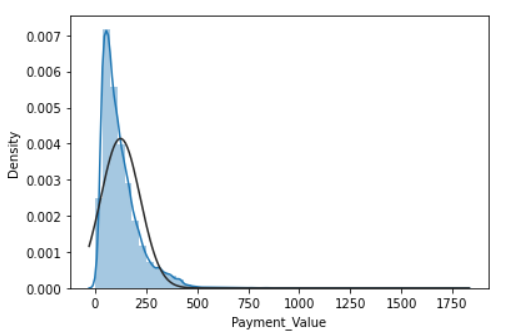
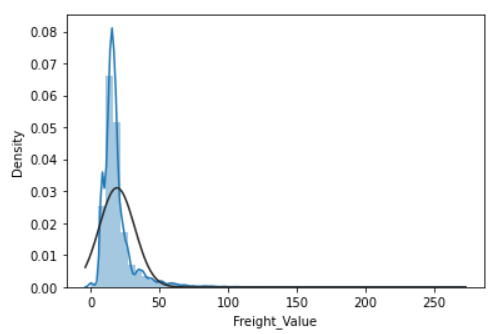
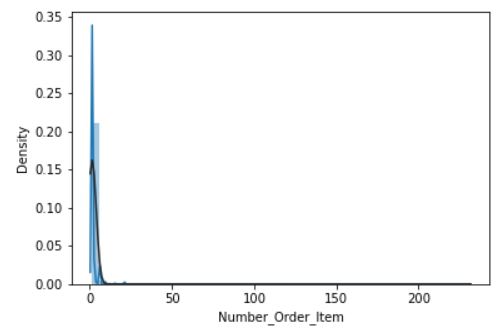
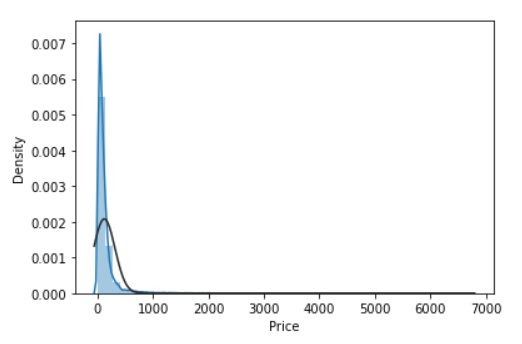
After preprocessing the dataset, the dataset now has more than 86 thousand rows and 34 columns

Text, letter

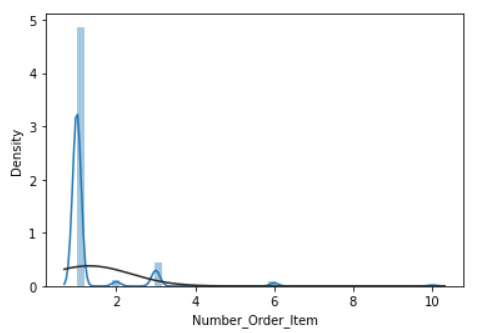
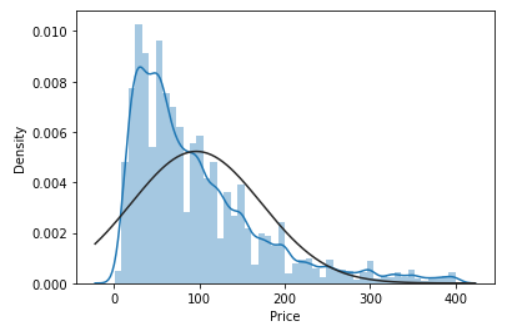
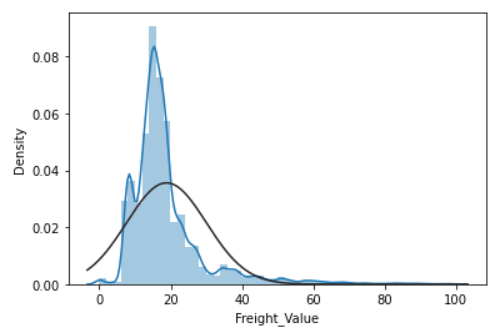
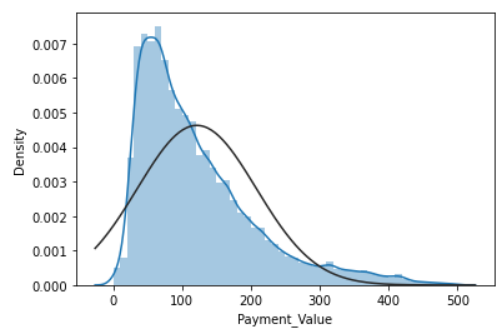
Description automatically generated

**Univariate Analysis**

To remove any outliners, I used distribution plot to watch for any sign of outliner.



After plotting some important features, we can easily see those features are skewed and have very long tail. This signifies there are many outliners that we can cut off to prevent bad influence on the outcome. Then I decided to cut off the long tail with my objective observation.



**Exploratory Data Analysis**

I used histogram plot to make sure that the data does not have the following warning sign:

* Possible outliers that cannot be explained or might be measurement errors
* Numeric features that should be categorical.
* A picture containing diagram

  Description automatically generatedBoundaries that do not make sense such as percentage values greater than 100

I try to investigate the correlation among variables, but there is no significant correlation is founded.

Graphical user interface

Description automatically generatedTo have the overall picture of my dataset. I create a small data frame of summary numbers of numeric columns. The dataset has about 28 thousand products with total of 71 product categories. There are more than 2 thousand sellers working in platform and they were serving over 80 thousand customers. We can see that the company provides their services and products in a very big scale, which covers more than 3 thousand cities and 12 states. Chart

Description automatically generated Chart, histogram

Description automatically generated

The charts above answer the question which product category is the most popular among the others. The gray one represents the quantity of the product category that was bought by the customer. The green chart is the total revenue of each product category. This can help company adjusting suitable strategies for different products.

Table

Description automatically generated

Finally, the table of 10 most valuable customers can give the company a glance of the behaviour and some details of their top customer. For example, we can easily see that most of the big spending customers come from state 11 and they likely to use credit for their purchase.

**Customer Segment with Feature Selection, Feature Scaling and Transformation**

**Feature Selection**

Feature selection can be done:

* manually by observing dataset
* manually by observing the correlation between features
* statistically using variance threshold
* statistically using generalized linear models
* recursive feature elimination, best-k, classification, ensemble method
* principal component analysis

Our goal is to do the customer segmentation, but the problem is the raw dataset has so many features, and we only interested in feature that makes sense in our case. However, there are no way we could know which feature to be selected. So, at the beginning of the project, we are going to use our common sense to first choose a set of features that could describe the customer features.

Then we will use Variance Threshold method to select significant variables among the others. About Variance Threshold feature selection, it removes all features with variance does not meet the threshold, in this project the threshold is set to be 0.5. The method assumes that features with a higher variance may contain more useful information than others. However, the drawback of this method is ignoring the relationship between feature variables.

**Modeling**

In this project, three popular clustering method will be used, which are k-means clustering, RFM segmentation and mini batch k-means clustering.

* Kmeans

Kmeans algorithm is an iterative algorithm that attempts to partition the dataset into distinct non-overlapping clusters where each data point belongs to only one cluster. The algorithm tries to group the data points of each cluster as similar as possible while also making sure that the clusters are different as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster’s centroid is at the minimum. However, one of the challenges of Kmeans is the user need to predefined number of clusters in advance. One of the approaches is to iterate through a list of predefined clusters, which will take a lot of time and resources, the other method is to use Elbow method to find the best number of clusters. In this project, Elbow method will be used to find the suitable number of clustering. Below is the function used to generate kmean clustering and to print out calinski harabaz score and silhouette score for performance measurement.

Text

Description automatically generated

* Mini batch Kmeans

The other clustering method is mini batch kmean clustering. The main idea is to use small random batches of data of a fixed size, so they can be stored in memory. Each iteration a new random sample from the dataset is obtained and used to update the clusters and this is repeated until the cluster is formed. The advantage of this method in comparing with kmeans is to reduce the computation cost and time required per iteration. Below is the function used to generate mini batch kmean clustering and to print out calinski harabaz score and silhouette score for performance measurement.

Text

Description automatically generated

* RFM

The last method is RFM segmentation (Recency, Frequency, Monetary). RFM analysis is a commonly used technique to give score to each customer based on how recent their last transaction was (Recency), how many transactions they have made (Frequency), and what the monetary value of their transaction was (Monetary). This technique gives the user information about who was company most recent customer? How often that customer bought something from the company? How much value in term of money did the customer bring to the company? This information can give company a hint of their own customer. The method will assign score for every customer based on recency, frequency and monetary. For example, we divide the metric into 4 cuts. For recency, most recent customer will have the highest score of 4. For frequency and monetary, the highest value will be assigned to the top 25% customer that frequently buys product from the company and has highest monetary value respectively. After that, we can combine all the score into one column, for example 414. This column can help us break the customer into segment. Then sum up all the score to have the total RFM score column. With this column, we can easily use the range of score to segment the customer. The figure below is the result after the process of RFM segment.

Table

Description automatically generated with medium confidence

Then for every model we perform the same process of finding the best number of clusters by using Elbow method. After plotting the Elbow, four clusters are chosen for all the modelling approaches.

Chart, line chart

Description automatically generated

Then with feature selection methods, manually and variance threshold, combining with feature scaling using standard scaler from sklearn, all the models are run one by one to generate the results. The below picture is one of the results after running multiple models with variance threshold feature selection.

Text

Description automatically generated

**Model Evaluation**

**Performance Indicator**

* Calinski Harabasz score

The Calinski-Harabasz score also known as the Variance Ratio Criterion. This method is one of the internal clustering criteria. The technique will sum all of the distance between-clusters dispersion and of inter-cluster dispersion for all clusters, the higher the score, the better the performances. The best way to use Calinski Harabasz score is to compare different clustering solution on the same data.

* Silhouette Score

Silhouette score can tell us how separation distance between the resulting clusters is. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters. This measure has a range of from -1 to 1. If the score is near +1, the number indicates that the sample data is far away from the neighboring clusters. Opposite with that, -1 means that the sample might be assigned to the wrong cluster. With the value near 0, it tells the user that the boundary between clusters is very blur. A higher Silhouette value relates to a model with better defined clusters.

With the combination of feature selection, finding the best cluster with Elbow technique and modeling using kmeans, mini batch kmeans and RFM, Calinski Harabasz scores and Silhouette Scores is given in the table below.

Graphical user interface

Description automatically generated

The best model after running all the combination of feature selection, feature scaling and modelling is the model using kmean with k equals 4, with variance threshold feature selection and no feature transformation. The Calinski score and Silhoute score of the model is the highest among the others. With 0.68 of Silhoute score, it is not the best clustering model, but it is acceptable for the purpose of the research. Further improvement can also be made using combination of more machine learning techniques and insider business knowledge. More features could be added to the dataset to make the model more accurate. Or another combination of machine learning method could generate a better outcome, such as Agglomerative Clustering or Birch clustering.

**Interpret Cluster**

To interpret the clustering result, first I need to take the cluster label from the best model and assign it to the original dataset.

Graphical user interface, text, application

Description automatically generated

Graphical user interface

Description automatically generated

So now, the original dataset has a new column, cluster. This is the result of clustering every transaction using our best model. Based on the Elbow method, four clusters is the best outcome the model can achieve, so we set the model to generate clusters in the range from 0 to 3.

Chart, bar chart

Description automatically generated

The count plot above indicates that most of the transaction is clustered as cluster 1. Let’s dig deeper into the detail.

Table

Description automatically generated with medium confidence

However, the average difference between the cluster in term of number of order item, price and payment value are not very significant.

Chart, bar chart

Description automatically generated

The plot above clearly shows that for all the clusters state number 11 is a very important state for the company to focus on. Cluster 1 is clearly taking the highest proportion in every states. Another reason to question what type of customer belong to cluster 1.

Lastly, a look at what kind of product each cluster has will be useful to understand the cluster.

Chart, bar chart, histogram

Description automatically generated