CHAPTER 3:

RESEARCH METHODOLOGY.

## 3.1 Introduction.

The research approach will serve as the foundation for this study, which will itemize the flow in multiple stages. This method is simply the sequential planning of each step to achieve the desired study objectives. The research methodology is presented and demonstrated in this section, along with logical explanations. The steps include; data exploration, data preprocessing, feature extraction and finally supervised and unsupervised modelling.

## 3.2 Research approach

The overall research approach flow is presented in figure 3.1. This methodology uses the CRISP-DM model as a foundation due to its broad applicability in solving data analytics problems. This flowchart provides a summary of all the phases with the steps involved in this study.

Graphical user interface, diagram

Description automatically generated

Figure 2 Research Framework

## 3.3 Data Extraction and Preparation.

**Overview of the data source.**

A secondary dataset is used for this study. The dataset is known as Cyclist-Divvy Trips 2021 which was collected to complete case study 1 of the google analytics certificate. The data was collected from the Chicago divvy Ride share service electronically through web scrapping and was then made available to Kaggle for other relevant case studies. The dataset is suitable for this study as the aim is to solve churn in enterprises with high churn rates and low switching costs.

**Raw data analysis**

The cyclists’ divvy data comprises 12 months of records of their client’s information. The datasets contain both categorical and interval variables such as gender, type of bicycles, the geographic location of the route and finally starting and ending stations of the trips.

## 3.4 Exploratory data analysis (EDA).

EDA is a process that conducts preliminary investigations on data to identify key characteristics and summarises the findings in tabular and graphical formats. Before conducting any analysis, EDA is vital as it gives more understanding of the dataset. The study will involve both statistical and graphical analytics for the univariate, bivariate and multivariate data.

The univariate analysis will focus on the statistical and graphical exploration of a single variable. R programming will be used as the primary tool for statistical analysis. R programming opts in this study as its one of the actively used tools for statistical analysis.

Furthermore, Tableau will be used as a primary visualization tool. Tableau was chosen because it offers a wide range of visualisation charts, stylisation mechanisms, and drill-down capabilities. The graphical outputs will be chosen based on the data characteristics and exploration direction. Options for charts include a bar chart, a pie chart, a stacked bar chart, and a tree map. Missing values will be excluded from the visualisation for each variable rather than the entire record being removed. This will aid in gaining a better understanding of the data while maintaining the authenticity of the variable performance.

The final output of the analysis will be analysed as the key goal of EDA is to look at distributions, outliers, and anomalies in the data to direct specific testing of your hypothesis and assist in proper data pre-processing (Komorowski et al., 2016).

## 3.5 Data pre-processing.

The data is pre-processed before the other phase to ensure result accuracy and meaningful interpretation. Firstly, the data is manipulated by transforming the variables with time data types to a measure which will represent the time taken for a person to complete one trip, the same approach is taken into the longitudes and latitudes which will be transformed into a relevant measure of the distance a person covered in one trip. Afterwards, the data will be summated to the clients with the same customer id. This is done to reduce the size of the dataset and to know the exact behaviour of each customer in terms of recency, frequency, distance covered, and time taken.

Pre-processing techniques such as treating missing values and treating noisy data such as outliers and skewed data are conducted. Graphical exploration such as box plots and histograms will be used to identify both the outliers and the data distribution to process each before further steps.

The imputation method of missing values will depend on the different missing value patterns identified (Phan, 2018). Truong et al. (2020) proposed that features with too many missing values and negligible impact on the target variable be excluded from the dataset. Imputing such variables is impractical because having too many imputed values may result in an unreliable result. However, variables with a small portion of missing values are imputed either using mean or median imputation. The imputation works by replacing the missing information for each variable with either the mean or the median of a specific variable depending on the distribution of the data.

## 3.6 Feature engineering and Sampling

This process involves the final preparation of the data before the actual machine-learning process. Feature engineering works can be divided into three categories: feature transformation, feature reduction, and feature construction.

Although some of the feature engineering processes are done on the data preparation and processing further engineering is more conducted. Specifically, the conversion of categorical variables to numeric variables so that the model can process accurate information. Also, the feature reduction process by reducing the variables through Principal Component Analysis (PCA). Finally, feature scaling which are techniques such as normalization and log transformation are used to reduce the wide scale of feature values that exists.

The dataset involves customer information from 12 different months, to get an ideal overview of consumer behaviour to extract the segments. The proportion of each sample of data from each month is then used to establish a training set to be used for the modelling phase. Therefore 60% of each month’s data is used as the training set and the remaining is used as the test set.

## 3.7 Clustering.

Customers will be divided into clusters based on their purchasing behaviour such as How often they rent a bicycle? How long do they ride? etc), and the characteristics of each cluster will be used to forecast subscription prices. The goal of this prediction is to identify all the segments according to consumer spending habits. Clustering is an effective technique for customer segmentation because it groups customers who share similar characteristics. For this study, three techniques will be used to segment customers. But before clustering the data is analysed the processed data is then fed into the RFM model, which determines the recency, frequency, and monetary values.

**RFM model.**

Recency, frequency, and monetary (RFM) analysis is a powerful and well-known database marketing technique. (“RFM ranking – An effective approach to customer segmentation”) It is widely used to rank customers based on previous purchases. RFM analysis is used in a variety of applications that involve many customers, such as online purchases, retailing, and so on. This method categorises customers based on three criteria: recency (R), frequency (F), and monetary value (M).

The customers are assigned three different scores for the variables of recency, frequency, and monetary value. Scoring is done on a 5-point scale. The top quintile receives a 5, while the others receive 4, 3, 2, and 1. The scores can be assumed to have distinct characteristics. The three attributes are then fed into three clustering algorithms: K-Means and Hierarchical clustering.

**Hierarchical clustering.**

This approach does not require an initial number of clusters before modelling in which each data point is considered a separate cluster. A specific distance metric is then used to calculate the proximity of two points, and the closest pairs are combined into a single cluster. This process is repeated until all data points have been combined into a single cluster.

**K means**

This approach requires an initial number of clusters before analysis. The main aim is to re-divide data objects and update cluster centres, an iterative heuristic process is used. The algorithm's basic idea is to generate a set of element objects and clusters at random. Initially, a random sample element is chosen as the initial cluster centre, and the distance between them is calculated. The optimal number of clusters is found using an elbow approach.

**Cluster interpretation.**

After successful customer segmentation, the effectiveness of the clustering approaches is examined. The time required to execute each technique is examined, and the performance of both models is compared, the model that consumes less time while also reducing the number of iterations is chosen. The resultant segments are then interpreted and according to a specific segment the subscription prices are assigned. Therefore, the resultant clusters will be assigned a competitive subscription price, this will complete the target variable of the training set. The subscription price from the segments will be the target variable in the dataset.

## 3.8 Predictive models

According to Azimlu et al. (2021), Artificial Neural Network (ANN) is the best-performing clustering-aided model for in-house price prediction. As a result, the Artificial neural network will be used as the benchmark model in this study.

**Neural networks.**

This algorithm works by constructing a computational structure made up of neurons, which are also known as interconnected elements. A neural network is made up of input layers, hidden layers, and output layers. (“Neural Network Basics and Concepts – Learn by Marketing”) Input layers introduce the data, hidden layers extract required intermediate data to determine the final solution, and output layers produce results. During the training phase, the layers adjust the relationship between predictors and the expected outcome to obtain accurate results. The weights obtained from the phase are then prefixed to the validation phase to observe the data prediction outcome. Following validation, the model can be used to predict new cases with unknown outcomes.

A neural network built with multi-layer perceptron and back-propagation will be used as it is the most effective algorithm. Preliminary testing has revealed that feed-forward back-propagation neural networks outperform other types of neural networks (Ghose & Tran, 2010).

**Performance evaluation.**

The training and test results of models will be compared to determine the best fit model for prediction. The best fit model will be chosen based on the model's high accuracy, which is contributed by the lowest error rate. Three statistical error measures will be evaluated primarily for testing and then training: "Maximum Absolute Error (MAX).", "Root Mean Squared Error (RMSE)," and "Average Squared Error (ASE),"   When the error value of all statistical criteria is lower, the accuracy of a model is higher.

## 3.9 Summary.

This chapter lays the groundwork for a series of sequential steps that will be taken to carry out the research. As a result, in this chapter, a research approach method has been plotted to structure the flow of this study. The survey data set will be examined from a variety of perspectives, including descriptive analytics, data pre-processing, and supervised and unsupervised modelling via hierarchical, k means, and artificial neural networks. Methods for validating achieved results for predictive modelling are also iterated. As a result, this chapter merely depicts the entire process of analysing and transforming the dataset into attaining the key objectives of the study.

CHAPTER 4

IMPLEMENTATION

## 4.1 Data description.

The dataset contains cyclists’ information collected through web scrapping from the company’s website. The dataset contains 12 months’ record of customers’ information described through unique features such as gender, type of bicycles, etc. The table below is the summary of the metadata of the dataset.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable Name** | **Description** | **Data type** | **Sample data** |
| Ride ID | Unique ride id | Character | CFA86D4455AA1030 |
| Rideable type | Type of bicycle | Nominal | classic\_bike |
| Started at | The date and time on which the ride started | Date | 3/11/2021 21:17 |
| Ended at | The date and time on which the ride ended | Date | 3/16/2021 8:36 |
| Start station name | The station at which the ride started | Nominal | Humboldt Blvd & Armitage Ave |
| Start station id | The id of the starting station | Character | 15651 |
| End station name | The station at which the ride ended | Nominal | Stave St & Armitage Ave |
| End station id | The id of the starting station | Nominal | TA1308000043 |
| Start lat | The Latitude in which the ride started | Numeric | 41.91751339 |
| Start\_lng | The longitude in which the ride started | Numeric | -87.70180896 |
| End\_Lat | The longitude in which the latitude ended | Numeric | 41.917741 |
| End\_lng | The longitude in which the longitude Ended | Numeric | -87.691392 |
| Member casual | The type of membership of the customer | Nominal | Casual, member |
| Gender | The gender of the customer | Nominal | Male, female |

Table 4.1: Metadata of the dataset

## 4.2 Data Exploration

Firstly, the data is extensively explored. This stage is the first phase of predictive modelling. It involves the use of data visualizations and statistical techniques to uncover initial characteristics and patterns of the data. Data exploration helps to understand more the important trends and the point of study.

Each variable is explored independently to identify missing values, outliers, and any hidden characteristics of the variables. This analysis serves as a foundation for a thorough understanding of the Divvy Cyclist dataset.

### 4.2.1 Univariate analysis

1. **Rideable type.**



Figure 4.1: Rideable type statistical summary

The statistical analysis of the rideable type explains the mode distribution of the variable. The figure above shows most cyclists prefer classic bikes to docked and electric bikes.

Simply, this can be explained by the affordability of the classic bikes compared to the docked and electric bikes. However, classic bikes have another advantage rather than only being affordable, they stand out as a last-mile option as the riders can leave them anywhere. Whereas, with docked bikes, the riders must find a station to leave the bikes.

Chart, bar chart

Description automatically generated

Figure 4.2: Rideable type bar plot

The figure above shows the graphical representation of the rideable type of variable. As presented, the classic bikes have thrice the number of users compared to the electric bikes. And 11 times the number of users compared to docked types. Finally, the graphs show the absence of any missing values.

1. **Started at and ended at**

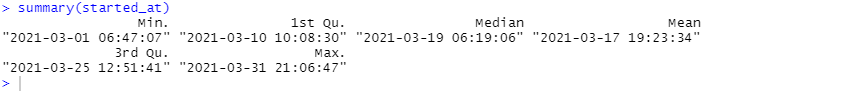


Figure 4.3: Statistical summary

The statistical analysis shows the distribution of the time in which cyclists started their rides. The analysis shows that the observations have taken place from 1st March to 31st March. The analysis shows the data is normally distributed as there is a slight variation between the median and mean values.

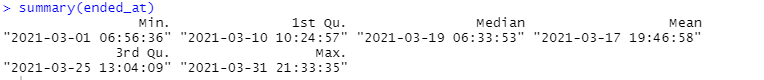


Figure 4.4: Statistical summary

The statistical analysis shows the distribution of the time in which cyclists ended their rides. The analysis shows that the observations have taken place from 1st March to 31st March. The analysis also shows the data is normally distributed as there is a slight difference between the median and mean values.

1. **Start station name**

Graphical user interface, application

Description automatically generated

Figure 4.5: Statistical summary

The analysis shows the starting station names, from the analysis May St and Taylor St have the highest mode, this means that many rides were started at that station compared to other stations.

May St & Taylor St had 45 instances followed by Calumet Ave & 18th St. The analysis shows that the starting station is somehow distributed, and there are no significant outliers. This means although May St & Taylor St has the highest number of starting rides it does not mean that it is the only hot spot for the rides.

1. **End station name**

A picture containing table

Description automatically generated

Figure 4.6: Statistical summary

The analysis shows the end stations for different rider routes. The analysis helps to identify which places in which most people ride to. The stations with the highest frequency show those are the points to which most people go and are the busiest places in the city.

From the figure above, 3 stations have significantly high frequencies compared to other stations. Clark St & Lake St have 245 instances, followed by Wood ST & Taylor St and Halsted St & Willow St which have 228 and 199 instances, respectively.

The identification of stations with abnormally higher frequencies compared to others shows that these stations are at the hotspots of town. Further analysis is done to identify why those points have higher frequencies compared to others.

1. **Geographical points**

By utilizing the longitude and latitude, geographical points were plotted on a map. Moreover, the routes which the riders used were plotted on the map. This analysis assists in identifying why some stations have higher frequencies compared to others.

Diagram

Description automatically generated with medium confidence

Figure 4.7: Bike routes.

The map above plots the routes which the riders have used. The analysis shows that there are 4 main hotspots. The hotspot means the places with a significantly considerable number of activities. These hot spots are identified as they have a high number of routes.

**Point A**

Chart

Description automatically generated

Figure 4.7: Bike routes in point A.

From the analysis, several distinct reasons led to Spaulding Ave & Armitage Ave station being a hot spot. Firstly, the map shows the presence of residential buildings around the station. This means the station is used by the people who are living around the area as a means of transport from other areas to the station. Secondly, the presence of Rico’s fresh market, Armitage produce, and a school these services attract the attention of different people and that is why the station is a hotspot.

**Point B**

Diagram

Description automatically generated

Figure 4.8: Bike routes in point B.

The analysis shows that the presence of people’s residences around the Halsted St & Willow St station is the reason behind the high usage of that station. Residents around the area use the station to navigate through various places around town. Also, the presence of a church might be a reason for the rides to the station.

**Point C**

Diagram

Description automatically generated

Figure 4.9: Bike routes in point C.

The analysis shows that the Clark St & Lake St station is in the city centre. The location that these stations are located has led to the high demand for bikes to navigate around town or to transport to town. Both the two stations around the areas are associated with a high number of routes to and from the city centre to access different services that are around such as restaurants, parks, and malls.

**Point D**

Diagram

Description automatically generated

Figure 4.10: Bike routes in point D.

The analysis shows that the station is around important services such as healthcare and education this explains the high number of activities conducted around the station. The station facilitates the users to access the University of Illinois, the Jesse Brown Medical centre and the UIC College of Medicine.

This hotspot is associated with thehigh student population around the area.

1. **Member casual**



Figure 4.11: Statistical summary

From the analysis, there is a higher proportion of riders who have a membership card with Divvy Cyclists compared to casual customers. The casual riders are the ones who purchase a full-day pass or a single ride while the riders with membership have purchased an annual plan.

The proportion does not mean that there are 716 annual members and 283 casual customers. But it means out of 999 rides that took place in March, 71.6% are annual membership rides and 28.4% are casual rides.

Chart, bar chart

Description automatically generated

Figure 4.12: Membership type bar plot

The figure above shows the graphical representation of the membership type variable. As presented, the annual members have thrice the number of rides compared to the casual members. Finally, the graphs show the absence of any missing values.

1. **Gender**



Figure 4.13: Statistical summary

From the analysis, there is a higher proportion of male riders compared to female riders. 62% of the total rides had male riders and 38% of rides had female riders.

### 4.2.2 Missing values

Missing data affect statistical analysis because of discrepancies in data patterns and knowledge loss. When dealing with missing data, it is crucial to comprehend the percentage of the distribution of missing values for each variable.

The analysis shows there are no missing values available in the data.



A picture containing table

Description automatically generated

Figure 4.14: Missing value plot

## 4.3 Data pre-processing

### 4.3.1 Feature engineering.

The process involves selecting and transforming the raw data into useful information that can be used in machine learning or to improve the performance of the machine model.

From the data, there are variables which cannot be implemented in the machine model. Firstly, the dates and times, are reliable sources for feature engineering they cannot be directly used in the model. Secondly, the geographical points (Latitude & Longitude) cannot also be used in a predictive model.

Therefore, two new variables are created in the dataset through feature engineering.

1. **Distance**

The variable is created through a geosphere function which calculates the geospatial distance between two geographical points using a Haversine method.

Text

Description automatically generated

Text

Description automatically generated with medium confidence

Chart, histogram

Description automatically generated

Figure 4.15: Distance statistical and graphical analysis

From the analysis, there are no missing values or rides with unknown distances in the dataset. The histogram, mean, and median values all show that the data for this attribute is positively skewed, with the mean and median at 2199.0 and 1806.4, respectively. The maximum value is greater than the (mean + 3x standard deviation) value, indicating that this variable contains extreme outliers.

Chart, box and whisker chart

Description automatically generated

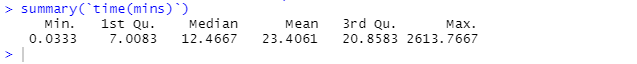
Figure 4.15: Distance Boxplot

The box plot above shows the presence of outliers, the rides which cover more than 6000m are considered outliers as they exceed the (mean + 3x standard deviation) value. The outliers in this instance will not be treated as they explain some riders ride more compared to others. Riding 12km is usual for an experienced rider.

1. **Time**

The time variable is created by taking the difference between the time the ride started and the time the ride ended.





Chart, box and whisker chart

Description automatically generated

Figure 4.15: Time statistical and graphical analysis

From the analysis, there are no missing values or rides with unknown times in the dataset. The mean time of the rides is approximately 23 minutes. The box plot shows the presence of outliers. The maximum value is greater than the (mean + 3x standard deviation) value, indicating that this variable contains extreme outliers.

The maximum time is 2613 minutes (about 2 days) which is equal to 43.55 hrs. These observations are outliers and do not make meaningful sense in the analysis.

The observations in which the time taken for the rides are greater than 8 hours are removed from the analysis. This is to remove the outliers in our analysis for better model performance.

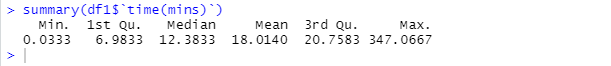


Figure 4.16: Statistical summary

Finally, the figure above shows the cleaned time variable in which the observations which have time greater than 500 mins are removed from the dataset. From this statistical analysis, the average time for each distance is 18 mins while the median time is 12.38 minutes.

1. **Level**

To further explain the characteristics of the rides, a new variable speed is created that measure the amount of distance covered per time in each ride. The speed calculated will be able to identify the cycling level of the riders.



Figure 4.17: Speed statistical summary

The analysis above shows the average speed used in the rides is 2.7 meters/ second. The average speed is slow due to distinct factors such as traffic or having many stops in between thus the time taken to cover a short distance is exceptionally long. The maximum speed is 6.778 meters/second equal to 15 mph, the average speed for an experienced cyclist in a short-medium distance.

Therefore, the speed variable is used to create three main levels which explain the degree of urgency between the rides. The first level is low which means the degree of urgency is extremely low and the speed used in these rides is between 0 to 2.051 metres/s (1st quartile).

The second level is medium, meaning the degree of urgency is normal and the speed used in these rides is between 2.051 to 3.648 metres/s. Finally, the high degree of urgency in which the speed is more than 3.648 metres/s.

A picture containing background pattern

Description automatically generated

Figure 4.18: Level statistical summary

Finally, three levels are created with frequencies 249.497, and 249, respectively. The medium urgency has the highest frequency compared to others.

### 4.3.2 Bivariate analysis

This involves the analysis of two variables to identify the relationship that exists between two variables. To determine how to improve customer retention it is vital to understand the relationship between the variables

1. **Level vs Bicycle type**

Chart, bar chart

Description automatically generated

Figure 4.19: Level vs Bicycle type bar plot

This helps to understand the relationship between the degree of urgency and the type of bicycle that is used. From the analysis, it can be observed that classic bikes are preferred by customers with a medium or low degree of urgency. Electric bikes are preferred by customers with a high degree of urgency and finally docked bikes are preferred by customers with a low degree of urgency.

Therefore, due to the speed of electric bikes, it is more useful to customers which is more urgent. Divvy Cyclists should ensure that more electric bikes are available at stations near busy areas for instance the Wood St & Taylor St station which is near the University of Illinois.

Moreover, the docked bikes should mostly be available at stations which are near parks or fun areas. Because most riders who use docked bikes are less urgent, riders who want to relax and ride bicycles for fun.

1. **Bicycle type vs member**

Chart, bar chart

Description automatically generated

Figure 4.20: Member vs bicycle-type bar plot

The plot explains the relationship between the type of bicycles used and the type of membership. As explained earlier, there is a higher proportion of rides with annual membership compared to casual membership. A higher proportion of annual membership rides prefer classic bikes followed by electric bikes but nonannual membership rides use docked type

1. **Level vs member**

Chart, bar chart

Description automatically generated

Figure 4.21: Member vs Level bar plot

From the analysis, it is observed that casual members' rides are highly associated with the degree of urgency. As the degree of urgency increases the number of casual rides decreases

1. **Rideable type vs distance**

Chart, box and whisker chart

Description automatically generated

Figure 4.22: Distance vs Bicycle type bar plot

From the analysis, the distance covered on the rides using electric bikes is slightly higher compared to the other bikes. This is proved by the mean distance of different rideable types, with the electric bike’s average distance being 2718.733. But also, there is high dispersion of distance in the electric bikes compared to the other bikes. Due to variability of the minimum and maximum value of the income.

1. **Rideable type vs time**

Chart, box and whisker chart

Description automatically generated

Figure 4.23: Time vs bicycle-type box plot

From the analysis, the time taken on the rides using docked bikes is higher compared to the other bikes. This is proved by the average time of different ride types, with the docked bike average time being 40.100. But also, there is high dispersion of time in the docked bikes compared to the other bikes. Due to variability of the minimum and maximum value of the income.

1. **Distance vs Time**

Chart, scatter chart

Description automatically generated

Figure 4.24: Time vs distance scatterplot

The analysis above explains the relationship between distance and time. The scatterplot shows there is a positive relationship between distance and time although the distribution of the points along the plot shows the relationship that exists is weak, thus a small correlation coefficient.

Moreover, the analysis explains the relationship between distance and time according to rideable type. It can be noted the classic and electric type rides have a positive relationship between distance and time, as the distance increases time also increases. Although the slope of the classic type shows a stronger relationship compared to the electric types.

Nevertheless, the scatterplot explains the docked bikes’ relationship between distance and time. There is a negative relationship between distance and time, which means as the distance increases time decreases. This is an unusual outcome from the analysis.

From the analysis, it seems like the riders using the docked bikes have a low degree of urgency as they use a lot of time covering short distances. They can either be riding at the parks, enjoying the scenery, or learning how to ride a bike.

Chart, scatter chart

Description automatically generated

Figure 4.25: Time vs distance scatterplot

Furthermore, the relationship between distance and time is analysed in terms of membership status. It can be noted that the strength of the relationship between the two variables is stronger with the annual members compared to the casual members.

The stronger relationship between distance and time within the annual members corresponds to what is expected. The correlation between distance and time is high, which means they are more experienced riders, with a higher degree of urgency compared to casual members.

## 4.4 Clustering

Customers will be divided into clusters based on their purchasing behaviour such as How often they rent a bicycle? How long do they ride? etc), and the characteristics of each cluster will be used to forecast subscription prices. The goal of this prediction is to identify all the segments according to consumer spending habits. Clustering is an effective technique for customer segmentation because it groups customers who share similar characteristics.

### 4.4.1 Data preparation

The preparation of data involves the selection of segmentation variables (clustering variables). Secondly, it involves scaling the data into one scale to perform effective clustering.

Selecting the appropriate variables for clustering depends on the output expected from the analysis. From this study, it is expected to segment customers according to their riding behaviour. The data exploration highlighted how the relationship between distance and time explains distinctive characteristics between the riders.

The segmentation variables that will be used are distance and time. Moreover, the rider id can be used to determine the frequency at which the riders rent the bikes. This variable is effective in determining the character of each rider and the frequency with which riders rent the bikes.

Therefore, the other variables will be dropped remaining only with distance, time, and rider id. The distance and time variable will be aggregated according to rider id to determine the total distance and time according to each rider. Finally, the frequency of each rider id will be determined.

Table

Description automatically generated with medium confidence

Figure 4.26: Aggregated data sample

Secondly, the prepared data is then scaled into a specific range. This is important as scaling prevents the variability between the datasets. Moreover, the scaled data is essential in generating good-quality clusters and it improves the accuracy of the clustering method.

A picture containing table

Description automatically generated

Figure 4.27: Normalized data sample

### 4.4.2 K means clustering

The fundamental idea of the method is to randomly produce a collection of element objects and clusters. A random sample element is first selected as the initial cluster centre, and the separation between them is computed.

To determine the number of optimal clusters to be used an elbow method is used. The knee plot is then used to determine the optimal number of clusters.

Chart, line chart

Description automatically generated

Figure 4.28: Knee plot

The plot above shows the total within sum of squares against the number of clusters. The within sum of squares is a measure that demonstrates the degree of dissimilarity among group members; the higher the sum, the greater the degree of dissimilarity within a group.

From the graph, it can be observed after the 4th cluster the slope of the plot decreased the total sum of squares started to reduce slowly from the 4th cluster. Therefore, the optimal number of clusters to be used is 4.

Moreover, there was a sharp decrease in the total sum of squares from 3 to 4 clusters. Hence again 3 clusters should be considered for the analysis.

**Clusters**

**K = 4**

Table

Description automatically generated

Figure 4.29: K = 4 statistical summaries

The figure above outputs the result of clustering when K = 4. It can observe that the cluster means, the clusters to which each observation was assigned (clustering vector), and a percentage (75.7%) that denotes the compactness of the clustering, or how similar the individuals within a group are to one another. The within sum of squares is used to measure the percentage of similarity within the members of the same group, if every observation inside a group was located at the same location in the n-dimensional space 100% compactness can be attained.

Therefore, in selecting an optimal k value the highest percentage of the within-cluster sum of squares will be used

Chart, line chart

Description automatically generated

Figure 4.30: K = 4 Cluster plot

The plot shows the distribution of the clusters. It can be observed that there are some observations which are far away from other observations. These observations have unusual characteristics compared to others. Therefore, these observations can be outliers.

**K= 3**

Table

Description automatically generated

Figure 4.31: K = 3 statistical summaries

The figure above outputs the result of clustering when K = 3. It can be observed the clustering sizes are 3, 73,75 respectively. The within-cluster sum of squares by cluster is 64.5% which means the dissimilarity between clusters has reduced compared to the analysis. These clusters are less compact compared to the previous clusters hence less accurate

Chart, line chart

Description automatically generated

Figure 4.32: K = 3 Cluster plot

The cluster plot confirms the presence of unusual observations this is what is also observed with the K = 4. Furthermore, this plot shows that the previous cluster 3 clusters were combined to form two clusters.

### 4.4.2 Hierarchical Clustering.

This method entails figuring out the Euclidean distance between the data points. The closest pairs are then grouped into a single cluster after being utilised to determine how near two locations are to one another. Until every data point has been merged into a single cluster, this procedure is repeated.

Hence, in this analysis agglomerative analysis is applied in which each observation is firstly considered as a separate cluster. At each step of the analysis, the two observations with the highest similarity between them are combined to form a new cluster. The process is iterated until all the observations are combined to form one big cluster (all observations).

It should be noted as the number of clusters formed decreases the dissimilarity between them increases. Therefore, the clusters which are combined at the later stages have a high percentage of dissimilarity.

Diagram

Description automatically generated

Figure 4.33: H = 4 Dendrogram

Each leaf represents a single observation in the dendrogram shown above. Similar observations are joined into branches as we climb the tree, and these branches are then fused at a higher level.

The height shown on the vertical axis shows how similar or dissimilar the two observations are. Therefore, the height of the cut in the plot determines the total number of clusters obtained.

From our analysis, the dendrogram is cut when the height is 4.

Chart, scatter chart

Description automatically generated

Figure 4.34: Cluster plot

From the cluster plot, it can be observed a total of 4 clusters were obtained with 87,60 1 and 3 observations, respectively. It can be noted again that 4 observations have an unusual characteristic compared to other observations there is a higher dissimilarity between them. Therefore, these observations are treated as outliers.

### 4.4.3 Clusters Improvement

The clustering accuracy in the analysis can be improved by removing the outliers from the analysis. As observed from the K means clustering and HC clustering approximately 4 observations had an unusual characteristic compared to others. In the K means when k = 4, three observations formed their clusters, when k = 3 they also formed their clusters. Moreover, in the HC clustering, 3 observations formed their cluster, and 1 observation formed its cluster.

This shows there is high dissimilarity within the observations, as HC clustering uses similarity to merge clusters thus the clusters with the least observations are very dissimilar compared to others. Hence, observations 71, 144,139 and 138 are outliers.

Text

Description automatically generated with medium confidence

Figure 4.35: Outliers statistical summary

The unusual characteristic that is present within this observation is that the time taken on the rides is unusual compared to the other, the mean time taken for the rides is 154. 85 minutes but these clients’ times are high than expected.

Therefore, the 4 observations are removed.

**K means Clustering.**

Chart, line chart

Description automatically generated

Figure 4.36: Knee plot

The knee plot determines 3 as the optimal number of clusters. Because more than three clusters the total within sum of squares decreases slowly. Therefore, there is a trade-off between the model complexity and the proportion of dissimilarity.

Table

Description automatically generated

Figure 4.36: K = 3 statistical summaries

The figure above outputs the result of clustering when K = 3. It can be observed the clustering sizes are 61, 26 and 60, respectively. The within-cluster sum of squares by cluster is 73.6% which means the dissimilarity between clusters is 73.6%.

**Cluster Validation.**

The silhouette coefficient is used to evaluate the performance of the clustering. When the Si coefficient is greater than 0 it means the observations are well clustered, closest to 1 means the observations are best clustered.

Chart, histogram

Description automatically generated

Figure 4.37: K = 3 Cluster silhouette plot

The silhouette plot shown above shows that our three-group clustering is effective because there are no negative silhouette widths, and the average silhouette width of the values is 0.42.

Chart, scatter chart

Description automatically generated

Figure 4.38: K = 3 Cluster plot

Finally, the cluster plot shows the distribution of the observations in the three clusters. There is a clear classification of observation between the clusters. Therefore, the k means has well classified the observations into three distinct clusters

**Hierarchical Clustering**

After the removal of outliers, a new dendrogram is plotted and the tree is cut when the height is equal to three.

Diagram

Description automatically generated

Figure 4.39: H = 3 Dendrogram plot

The resultant cluster from the dendrogram is plotted in the cluster plot. Each cluster contains 71, 60 and 16 observations, respectively. It can be observed there is a slight intersection between the observations of the 3rd and 1st clusters. Therefore, it is observed there is a little misclassification between the two clusters.

Chart

Description automatically generated

Figure 4.40: H = 3 Cluster plot

### 4.4.4 Summary.

From the analysis done above both K means, and Hierarchical cluster were used to group the observation with the same characteristics. Firstly, according to (Kaushik, & Mathur, 2014) the K means performance is better than Hierarchical clustering. This can also be observed in the cluster plot of the K means and Hierarchical algorithms. There is misclassification of the observation in the hierarchical cluster plots whereas in the k means there is no misclassification.

Moreover, there is a trade-off in the choice of the optimal number of clusters in the k means. Although the analysis with k = 4 has a higher dissimilarity percentage of 79.7 compared to that of k = 3 with 73.7 % of dissimilarity between clusters. The silhouette coefficient shows the 4-group clustering is not effective as it has negative silhouette values, and the average silhouette value is less compared to the 3-group clustering.

The k means with k = 3 is the best clustering technique in grouping customers together.

## 4.5 Cluster Interpretation.

This involves further interpretation of the specific clusters. To link the generated clusters and the business objectives it is important to identify the relationship that exists between each cluster this information will be important in identifying the characteristics that each rider has, and an effective marketing strategy can be used to prolong the customer lifetime value and thus reduce churn.

Therefore, the main segmentation variable that is used in this analysis is the distance covered by each rider and the time used by each rider. We can find the relationship that exists between distance and time according to the clusters.

Chart, scatter chart

Description automatically generated

Figure 4.41: Time vs Distance scatterplot

The plot above shows the relationship between distance and time according to the respective cluster. It can be observed that cluster 1 is the furthest in the plot and is characterized by higher distance values of more than 21000m (21km), cluster 2 is in the middle of the plot while cluster 3 is nearest the origin.

### 4.5.1 Cluster 1 (Gold)

Cluster 1 contains 26 customers. To interpret the behaviour of these customers a descriptive statistic of the variables is conducted to identify the mean, sum and standard deviation of each variable in the 1st cluster.

|  |  |  |  |
| --- | --- | --- | --- |
| **Statistic** | **Sum** | **Mean** | **Standard Deviation** |
| **Distance** | 677,264.0 | 26,048.616 | 4,104.731 |
| **Times** | 4617.133 | 177.58205 | 55.10123 |
| **Frequency** | 266 | 10.230769 | 1.394495 |

Table 4.2: Cluster 1 statistical summary

From the table above, it can be observed that the average distance covered by riders in the first cluster is about 26,048 m. This means that the 26 riders covered an average of 26 km in March. This average is higher than the total average which is 14 km. Therefore, this cluster is characterized by higher distance covering individuals.

Secondly, the average time used by the riders is 177.5 mins which are higher than the total average time used by the riders. Therefore, this cluster is also characterized by the longer time taken on the rides.

Finally, the mean frequency of the riders is 10.2 which is higher than the average. This means the riders in this cluster have rented the bikes approximately 10 times a month.

From the analysis of the statistics, this cluster can be termed **Gold**. This is because the characteristics of customers that are found in this group are the loyal and important customers in the Divvy Chicago dataset.

### 4.5.2 Cluster 2 (Silver)

This cluster contains 61 customers.

|  |  |  |  |
| --- | --- | --- | --- |
| **Statistic** | **Sum** | **Mean** | **Standard Deviation** |
| **Distance** | 978,287.8 | 16,037.505 | 3259.105 |
| **Times** | 8,549.833 | 140.16120 | 71.08070 |
| **Frequency** | 461 | 7.557377 | 1.323185 |

Table 4.3: Cluster 2 statistical summary

From the table above, it can be observed that the average distance covered by riders in the first cluster is about 16,037.05 m. This means that the 61 riders covered an average of 16 km in March. This average distance covered by these customers corresponds to the average distance covered by all the customers.

Secondly, the average time used by the riders is 140.16 mins which are slightly higher than the total average time used by the riders. Therefore, this cluster is also characterized by the longer time taken on the rides.

Finally, the mean frequency of the riders is 7 which is the same as the average. This means the riders in this cluster have rented the bikes approximately 7 times a month.

From the analysis of the statistics, this cluster can be termed, **Silver**. This is because the characteristics of customers that are found in this cluster are the same as the total average characteristics of the total customers, these customers cannot be termed as loyal and long-lasting customers neither can be termed as disloyal and temporary customers. Effective marketing strategies for this group can lead to the long-term success of the business because they can easily shift to loyal long-term customers.

### 4.5.3 Cluster 3 (Bronze)

This cluster contains 60 customers.

|  |  |  |  |
| --- | --- | --- | --- |
| **Statistic** | **Sum** | **Mean** | **Standard Deviation** |
| **Distance** | 511450.8 | 8524.179 | 3410.762 |
| **Times** | 4423.633 | 73.72722 | 48.07334 |
| **Frequency** | 250 | 4.166 | 1.379 |

Table 4.3: Cluster 3 statistical summary

From the table above, it can be observed that the average distance covered by riders in the first cluster is about 8524.179 m. This means that the 60 riders covered an average of 8 km in March. This average distance covered by these customers is ridiculously insignificant compared to the distance covered by customers in different clusters.

Secondly, the average time used by the riders is 73.7272 mins which are lower than the total average time used by the riders. The average distance of 73.7 corresponds to the 1st quartile of the time taken by total riders Therefore, this cluster is also characterized by the shorter time taken on the rides.

Finally, the mean frequency of the riders is 4 which is less than the average. This means the riders in this cluster have rented the bikes approximately 4 times a month.

From the analysis of the statistics, this cluster can be termed **Bronze**. This is because the characteristics of customers that are found in this cluster correspond to one-time customers or new customers. This is because the average distance and time are minimum, they spend less time on the bikes and they cover noticeably short distances. Effective marketing strategies for this group may turn these new and one-time customers into loyal and long-lasting customers.

Therefore, the efforts of the company should be giving incentives to these types of customers, this will ensure the reduction of customer churn.

## 4.6 Predictive analysis.

**Neural Network.**

A neural network is composed of three main layers the input layers, hidden layers, and output layers. The data is introduced in the input layer, hidden layers extract the necessary intermediate data to compute the definitive answer, and the results are generated in the output layer. The layers modify the link between the predictors and the anticipated result throughout the training phase to provide correct results.

The most efficient algorithm is a neural network created with multi-layer perceptrons with back-propagation (Ghose & Tran, 2010).

**Scaling the data.**

Firstly, the data is normalized using a min-max normalization. The normalization process enables speeding up the process of learning the data which leads to faster gradient convergence.

Text

Description automatically generated

Figure 4.42: Normalized data.

**Splitting the data.**

After scaling, the dataset is divided into training and validation data in proportions of 70% and 30%, respectively.

The data is randomly divided each time the loop runs, and the Neural network is constructed from the training data and assessed for accuracy using the validation data.

Text

Description automatically generated with medium confidence

Figure 4.43: Training and validation data sample.

**Model Implementation.**

The training data is trained in the neural network.

Graphical user interface, text

Description automatically generated

Figure 4.44: A neural network model

Notably, the activation function used is logistic because it performs better than the tanh function. Moreover, the suitable set of hidden layers is 13,10 and 3.

Chart

Description automatically generated with low confidence

Figure 4.45: The neural network plot

**Model Evaluation.**

A screenshot of a computer

Description automatically generated with low confidence

Figure 4.46: Model summary

From the analysis, it can be observed the model misclassified only one observation. The accuracy of the model is 97.78% and a kappa value of 96.7%. This shows a higher performance of the model in predicting the clusters.

Moreover, it can be observed that the model successfully predicted the third class with having a sensitivity and specificity of 100% respectively.

As the predictive accuracy of the model is remarkably high. This model will be used to predict the clusters in a test set with no target variable. From the statistics obtained, the model will successfully predict the third cluster although the 1st and 2nd clusters will be misclassified.

**Predicting the test set.**

The resultant model is then used to predict the cluster in a test set.

A picture containing table

Description automatically generated

Figure 4.47: Test set sample

The observed test set does not have a specific cluster. The neural network trained in the training data is used to predict the resultant cluster for each observation. The neural network will analyse each observation in terms of the distance covered, the time taken by the riders and the frequency with which each rider rents a bike. These variables will be used to predict the category which each rider contains.

A screenshot of a computer

Description automatically generated with medium confidence

Figure 4.48: Predicted test set sample

The figure above shows the predicted clusters in each observation. From the 143 observations gold, silver and bronze clients were 13,42 and 88, respectively.

CHAPTER 5

CONCLUSION.

This project aimed to increase customer retention through personalized subscription prices using machine learning. In the context of the dataset used, the project aimed to offer a dynamic subscription price that changes per customer’s behaviour. With this, it was vital to identify the characteristics of each variable and how they affect prices.

The dataset had 15 variables, and through data exploration, it was observed that the variables related to time and geographical points were worthy. These variables were engineered into new variables, the time taken in each ride and the distance covered in each ride. These variables were useful in identifying the characteristics of each ride.

From these data, it was easy to differentiate the types of riders. Some rides covered more distance whereas some covered short distances. Moreover, some rides covered an abbreviated time whereas others covered a long time.

The riderid variable was useful because it determined the frequency with which the riders used the bikes. This variable was used to differentiate between new customers and loyal customers as it determined the number of times, they rented the bikes.

Therefore, these three variables were useful in studying consumer behaviour. The variables answered three main questions:

* How often does a unique rider rent a bike?
* How long do they ride?
* How much distance does each rider cover?

The two variables distance and time were aggregated according to the rider id. This is to determine how much distance and time each rider covered. With this, it is possible to study the characteristics of the riders according to the total distance and time covered in a month and the frequency in which the bike was rented and thus suggesting a unique monthly price according to each rider.

Clustering aided to extract the optimal number of customer segments that exist between the riders’ characteristics rather than the fine customer segments that existed. Therefore, through K means clustering the 3 clusters were identified as Gold, Silver, and Bronze. Gold segments are considered the top and loyal customers while bronze segments are considered the bottom and new customers.

Therefore, through the characteristics obtained from each cluster in terms of distance, time and frequency, an optimal subscription price may be calculated.

For example, the gold segments contain customers with an average renting frequency of 10 times a month. They cover an average of 26 km and have an average of 3 hours in total on a bike.

Rather than charging a $10 (the current divvy monthly rate) for all members. A cheaper membership rate of $8 for the gold segment, $10 for the silver segment and $12 for the bronze segment will be a good marketing strategy.

Moreover, the price for the riders will change each month according to their behaviour through the neural network model generated. The riders from the bronze segments will want to have a cheaper plan therefore they will increase their engagement with the company.

Therefore, the prices will not only be personalized but also dynamically change according to consumer behaviour. This will not only increase the switching barriers but also the need for the client to increase engagement with the company.