**CCT College Dublin**

**Assessment Cover Page**

*To be provided separately as a word doc for students to include with every submission*

|  |  |
| --- | --- |
| **Module Title:** | Statistical Techniques for Data Analysis |
| **Assessment Title:** | CA1 |
| **Lecturer Name:** | Kayoum Khbuli |
| **Student Full Name:** | Oluwatimileyin Oladipo Ayomide |
| **Student Number:** | 2023383 |
| **Assessment Due Date:** | 17th Oct 2023 |
| **Date of Submission:** | 12th Nov 2023 |

**Oladipo Ayomide**

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

**Introduction**

**What is data?**

Data is plain facts gathered within a defined context. Statisticians would refer to it as a set of observations, which contained within variables or columns whin columns possessing varying/unique characteristics. Data can also be referred to as a piece of information after it has been summarised, and when subjected to analysis, data can be referred to as evidence of a hypothesis. Before data can become information and ultimately evidence it must go through a process. Data analysis is the process in which raw data is ordered and organized, to be used in methods that help to explain the past and predict the future (Hector, 2013).

Data analysis requires good use of statistical techniques and probability in other to understand, summarise and efficiently and accurately manipulate data hence facilitating making reliable predicts to stakeholders of a domain. In this project we would be using statistical and probability techniques to summarise Bank churn dataset and ultimately draw conclusion as informed by our analysis.

**Forms of data**

Data is a generalised term, but within this term are branches which can be referred to as forms/ types. Data can broadly be divided into 2 types:

1. **Categorical types:** Categorical data are also referred to as qualitative data. These are data’s that cannot be measured. They are in general any form of data that is not in numerical for. E.g., smoker/ non-smoker.

2. **Numerical types:** Numerical data on the other hand are data’s that are in number form E.g. [Age: 20, 18.]. They are also called quantitative data’s, this means they can be quantified by numbers, and increments is also proportional to quantity for example an 18year old person is more aged than a 2year old person.

It is worth saying that these 2 data types are just a broad overview of forms in which a data can be presented, however each data type has sub-division in them, and they mean different things as relating to characteristics of such data and how they would be treated in analysis.

**Understanding 'Bank churn' dataset**

The Bank churn dataset i have chosen to analyse is from ABC Multistate bank (Topre, 2022). Customer churn in simple term is when a customer of an organisation decides to pathways with the business. In this contest ABC multistate bank customer churn would be when a customer of the bank discontinues their patronage of the bank services. This data set has 12 variables in it, and the bank wants to be able to see how each if these columns influence a customer’s decision to churn the bank. The data is available in

**Data source**

This data is available in ‘comma separated format’ and has been downloaded from Kaggle open source:

[Kaggle](https://www.kaggle.com/datasets/gauravtopre/bank-customer-churn-dataset)

**Columns definition (Topre, 2022).**

1. customer id: Unique identification of customers.

2. credit score: Credit score of each customer

3. country: This is the region the customer is based.

4. Gender: This is the gender of each unique customer.

5. Age: This variable contains the age a customer.

6. Tenure: This is how long a customer has been with the bank.

7. Balance: This is the available balance of customer account at the time of data collection.

8. products number: This is the number of services a customer signed up for with the bank.

9. credit card: These states wither a customer has a credit card or not.

10. active member: This checks if a customer is still active customer of the bank.

11. estimated salary: The estimated gross salary of each customer.

12. churn: Checks if a customer has left ABC Multistate bank or not

**Objectives**

With this data, I would be implementing the following in python programming language:

1. Read data into pandas dataframe format.

2. Explore 5 number summary of data using method in pandas and functions.

3. Visualise distributions using appropriate visualisation technique.

4. Report on findings and conclusion based on analysis.

**Implementation**

**Data shape**

To work with data, I read the csv file into a pandas dataframe(df) and assigned the dataframe to a new variable ‘churn’. Loading the csv file into df facilitates manipulation of the data. Churn has 12 columns and 10000 rows and there are 2 categorical variable, 10 numerical variables out of which 8 are integers and 2 are floats.

**Converting categorical variables back to objects:**

Data dictionary defined ['active member', 'churn', 'credit card'] variables as categorical variable but when churn was loaded these variables had been encoded into numerical form to facilitate visualisation and modelling. To be specific, 'active member' 1 and 0 means yes and no respectively, stating if customer is still an active member of the bank. 'churn' 1 and 0 mean yes and no respectively, stating if customer have left the bank or haven't. 'Credit card' 1 and 0 respectively mean yes and no, stating if a customer has credit card or not.

This variable according to information output were interpreted has integers, statistical summary of this values was not interpreted as number. To explicitly define them as just classification values, I converted them to objects using ‘astype ()’ method.

**Descriptive summary:**

Statistics summary gives a high-level idea to identify whether the data has any outliers, data entry error, distribution of data such as the data is normally distributed or left/right skewed (Mahadevan, 2022).

Descriptive statistics has two broad aspects namely **summarization and visualisation**. It is a branch of statistical analysis, that involves summarizing, organizing, and presenting data in a concise and meaningful way (simplilearn, 2021).

Descriptive statistics however does not make any generalized conclusion on a general population, for instance the descriptive statistical result of 'churn' is not a representation of all the customers in ABC bank, it only focuses on describing the columns in 'chun'(simplilearn, 2021). Descriptive statistics summarizes the measure of central tendency (location), and variability in each column of a dataset (Torres-Reyna, 2008).

**Summarizing statistics:**

Using “. describe()” method, I obtained statistical description of churn Dataframe. This method returned result for: Mean, median, maximum, minimum, standard deviation, 25% and 50% of each numerical variable in churn.

**Credit-score:**

Statistical summary revealed that the average(mean) credit-scoreof the customer in churn dataframe is 650.528800 and the median is 652.000. The difference between the mean and median should that the distribution of the observation in the credit-score variable are uneven. Furthermore, boxplot shows the lower extreme value is 350 and upper extreme value is 85. Outliers were found below the minimum value on box plot, and because outliers affect the mean the credit- score distribution is slightly left skewed. The left skewness was displayed on an histoplot. The 1st Interquartile range (IQR1) and 3rd interquartile range (IQR3) are 584.000000 and 718.000000 respectively.

A graph of credit score

Description automatically generatedA diagram of a credit score

Description automatically generated

Figure 1: Credit score histoplot

Figure 1b: Credit score Box plot

The most frequent credit score in the variable is 850 and the range is 500. Functions is used to obtain the result for standard deviation and variance, and they are 96.653299 and 2.168621e-23 respectively. The variance reveals how close or dispersed are the observations from the mean and as shown in *fig:1* the variance in credit score makes the distribution close to the mean.

**Age:**

Mode function shows the most frequent age out of 10,000 customers is 37. The minimum and maximum customer age is 18 and 92 years old respectively and the range between the min and max is 74 years old. Statistical summary revealed that age distribution was right skewed and unsymmetrical because the mean was greater than the median. The average(mean) ageof the customer in churn dataframe is 38.921800 and the median is 37. Boxplot shows there are outliers, these are numbers that are significantly higher than 95% of the entire age distribution. It also shows the mid 50% line is not divided the plot equally. The variance in the age distribution is 8.529184e-28 and the standard deviation is 10.487806, this having an impact on the tightness of the observations to the mean as shown in *fig: 2..* A negative interquartile range (-34) was found in age variable, this suggest that the quarter of the age (IQR1) data is significantly lower than the (IQR3) as shown in the box plot.

A graph of a number of columns

Description automatically generated

Figure 2: Histo plot of the age distribution.

**Tenure:**

Statistical summary revealed that the average(mean) tenure of the customer in churn dataframe is 5.012800 and the median is 5. Histoplot show the observations in the tenure variable are mutually exclusive. They did not follow a pattern or variance. Furthermore, boxplot shows the lower extreme value is 0 and upper extreme value is 10. No outliers were found on box plot, and distribution did not appear to be skewed. The 1st Interquartile range (IQR1) and 3rd interquartile range (IQR3) are 3.0000 and 7.0000 respectively, and the range is 10. The most frequent tenure is 2.

**Products number:**

Most customers in the churn df have 1 products number, and the minimum and maximum number of products in sample customer are 1 and 4 respectively and their range is 3 products. All sample customers have an average of 1.530200, and median of 1, indicating a right skewed distribution. The minimum and the median 50% value are the same, this reflect on the box plot and one outlier of 4 was found above the 3. The standard deviation and variance are 0.581654 and 1.994105e-27 respectively. This reflected on the histoplot as the data distribution was largely spaced out from the mean.

**Balance:**

Mode function shows the most frequent balance out of 10,000 sample customers is 0.0. The minimum and maximum customer balance is 0.000000 and 250898.090000 respectively and the range between the min and max is 250898.09. Statistical summary revealed that balance distribution was right skewed and unsymmetrical because the mean was greater than the median. The average(mean) sample customer balance in churn dataframe is 76485.889288 and the median is 97198.540. Boxplot shows there are no outliers, however many balance values are high, therefore the median moved significantly to the right way from the mean hence mid 50% line did not divide the plot equally. The variance in the balance distribution is 1.270521e-15 and the standard deviation is 62397.405202, this having an impact on the closeness of the observations to the mean as shown in *fig: 3..* A negative interquartile range (-38166.76) was found in age variable, this suggest that the interquartile range 1 of the balance (IQR1) data is significantly lower than the (IQR3) as shown in the box plot.

A diagram of a bar chart

Description automatically generated

Figure 3: Box plot for balance

**Estimated salary:**

Mode function shows the most frequent estimated salary out of 10,000 sample customers is 24924.92. The minimum and maximum customer balance is 11.580000 and 199992.480000 respectively and the range between the min and max is 199980.90. Statistical summary revealed that estimated salary distribution was left skewed and unsymmetrical because the mean was lesser than the median. The average(mean) sample customer estimated salary in churn dataframe is 100090.239881 and the median is 100193.915000. Boxplot shows there are no outliers *fig:4*. The variance in the estimated salary distribution is 4.082964e-16 and the standard deviation is 57510.492818.

A blue rectangular object with black lines

Description automatically generated

Figure 4: Box plot for estimated salary

**Categorical data:**

Descriptive analysis of credit card shows they are more customers (7055) that own credit card than those that don’t one (2945). Furthermore, there are 5457 male customers in the gender variable than females which are just (4543). In the country variable, France is the most frequent followed by Germany and Spain with 5014, 2509 and 2477 customers from 10000 sample size coming from each respectively. Finally, there are 5151 active members and 4849 in active members. This values where visually represented using barplot. Barplot was chosen because it good for comparing frequencies of categorical values.

**Relationship plots:**

In *fig:5* I compared the relationship between the active customer in the sample space and those that have churn ABC bank in this sample. Result shows there are more active customers still with the bank compared to those that have churn.

**A graph of a number of blue and orange bars

Description automatically generated**

Figure 5: Proportion of active customers compared to customer churn.

In fig 6 I used scatter plot to compare the age of customers to their available balance and used churn as hue to see the relationship. The plot shows that most customer have balance between 5000 and 20000. And customers from age 18- 94 have some that has available balance of 0.00. Customers between age 40 and 70 years old showed significant churning in the sample population compared to other age group.

**Conclusion:**

In conclusion, statistical analysis shows that data is overall good and usable for advance analysis and predictive analysis. However, all numerical variables appeared skewed, and some variables have significant outliers. In other to use this data to produce meaningful and reliable insights, it is imperative that churn is subjected to feature engineering and scaling methods to be able to adjust the outliers whiles preserving the trueness of the data. MinMax scaling technique is recommended for normalization because this data is skewed. The range of numerical variables are also high, as there are many high values in variables such as balance, estimated salary and credit score showing significant ranges. Smooth function machine learning models are sensitive to outliers, therefore using this value in their current form, would introduce bias to machine learning predictions as they are more likely to learn such patterns. Scaling these values will ensure the pattern are preserved and the values are reduced to be between 0 and 1, hence eliminating bias potential. Furthermore, churn would benefit from principal component analysis to reduce its dimensions and whiles preserving the pattern as well, to eliminate the noise in that data, before introducing to machine learning for predictive analysis.

**References:**

Hector Cuesta (2013) Practical Data Analysis. Birmingham, UK: Packt Publishing. Available at: <https://search.ebscohost.com/login.aspx?direct=true&db=e250xww&AN=654543&site=eds-live&scope=site> (Accessed: 18 October 2023).

Topre, G. (2022). Bank Customer Churn Dataset. [online] www.kaggle.com. Available at: <https://www.kaggle.com/datasets/gauravtopre/bank-customer-churn-dataset.>

Torres-Reyna, O. (2008). Data Preparation & Descriptive Statistics. [online] Available at: [https://www.princeton.edu/~otorres/DataPrep101.pdf](https://www.princeton.edu/~otorres/DataPrep101.pdf%20) [Accessed 14 May 2022].

Mahadevan, M. (2022). Step-by-Step Exploratory Data Analysis (EDA) using Python -. [online] Analytics Vidhya. Available at: [https://www.analyticsvidhya.com/blog/2022/07/step-by-step-exploratory-data-analysis-eda-using-python/.](https://www.analyticsvidhya.com/blog/2022/07/step-by-step-exploratory-data-analysis-eda-using-python/)

Casella, G., Fienberg, S. and Olkin, I. (2005). *A Modern Introduction to Probability and Statistics*. Springer Science+Business Media.