Customer Retention Project Report

Abstract: Customer satisfaction has emerged as one of the most important factors that guarantee the success of online store. Five major factors that contributed to the success of an e-commerce store have been identified as: service quality, system quality, information quality, trust and net benefit. The combination of both utilitarian value and hedonistic values are needed to affect the repeat purchase intention (loyalty) positively. A comprehensive review of the dataset is done using data exploration, EDA, preprocessing, data manipulation and feature selection.

I. Data Analysis

Customer satisfaction is the most important objective of any sales business. Customer satisfaction retains the old customers, brings in new customers, increases sales, decreases the spend on advertising, etc., The research furthermore investigated the factors that influence the online customers repeat purchase intention. The combination of both utilitarian value and hedonistic values are needed to affect the repeat purchase intention (loyalty) positively. The data is collected from the Indian online shoppers. This dataset contains 269 rows and 71 columns to determine whether the customer is satisfied or not.

II. Data Analysis

The purpose of Data Analysis is to extract useful information from data and taking the decision based upon the data analysis. Linear model assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x). Implementation includes importing necessary libraries, cleaning and analysing the dataset, building various models and using the best model for prediction.

Exploratory Data Analysis is the process of investigating the dataset to discover patterns, and anomalies (outliers), and form hypotheses. Getting maximum insights from a data set, Uncover underlying structure, Extracting important variables from the dataset, Testing underlying assumptions and determining the optimal factor settings.

1. Explore the data, check shape and explain the data, replace the column names inot numbers.

2. Check for the unique values of each column. 39 is the maximum unique value present in the dataset.

3. Check for the information of the data - object and integer type of data.

4. Distribution of categorical and continuous variables are determined using visualisation techniques.

5. Remove if there is any duplicate rows.

6. Check for missing values and null values. Treat it using central limit theorem or remove it if filling gives biased result.

7. Check for outliers and skewness. Treat it using z-score or Inter-quartile range.

8. Check for correlation among the features. If presence of correlated variables observed, reconfirm using graph like scatter plot and remove it.

9. Check for linear association between the features(multicollinearity) using variance inflation factor. It is absorved when two or more features are highly linearly realted. Remove it if it is present in high amount.

10. Feature reduction is already done through multicollinearity and correlation. Where there are more features which does not give information to predict the target remove it using Principal Component Analysis to get better performance.

11. Transform the data using any encoding techniques. Here, ordinal encoder is used to encode the data.

12. Normalisation and standardisation of data is necessary to keep all values in same range to get enhanced result.

III. Concluding Report

The column names are renamed to numbers. Checked the unique values of all the columns which has a maximum value of 39. There are 70 object type of data and 1 of integer type. On visualising the categorical variable, it is imbalanced. There were 166 duplicates. Removing these duplicated affects the percentage loss of data but removing this is necessary to not get biased result. Enoded the data using ordinal encoder for all 70 columns. There is no null values present in the dataset. On checking the statistics of the data, it is concluded that the data has to be standardised. 10 outliers are detected and removed. If it is not removed, it affects the result. No correlated variables are found. Standardise the data using Standard Scaler. There was multi-collinearity found in the data. Principal Component Analysis is used to find the best 30 features. Thus the 30 features are selected using data analysis.