**CIND820-Project-Work-for-Austin-Okafor**

**My working Dataset can be found on my Github link**

**Github link:**[**https://github.com/AustinOkfor/CIND820-Project-Work-for-Austin-Okafor.git**](https://github.com/AustinOkfor/CIND820-Project-Work-for-Austin-Okafor.git)

**EDA Analysis**

For his phase of the project, we will utilize some R packages for Automated Exploratory Data Analysis. This is especially useful because of the size of the data we are working with, the complexity of the variables, and the number of columns in the dataframe.

We started by installing Tidyverse, DataExplorer, and SmartEDA. These were used for our EDA work.

Then, leveraging our previous knowledge of the data from the Descriptive Statistics analysis conducted in our last work, we will load the dataset and assign column types automatically. We will however not change any data value.

During the data upload, I converted the variables to factor chose not to alter the data by converting -99 and -88, which represent missing and unknows values respectively, to NAs because most of the column types are factors. I felt that changing the original values may distort the true understanding of how the data is distributed, because certain levels will be lost.

Also, due to the number of columns, it was very tedious trying to obtain the column types and ensuring no error was made during assignment. I had to redo this step several times to ensure accuracy.

After reading the file, I removed the UID column as this is not useful for any analysis, and then ran "str" and "summary" to understand the distribution of the data better and confirm the variable types were properly captured during upload. Irregular data was then rectified.

From running summary, we can observe that quite several columns have missing data represented by -88 and -99. This information is useful in performing further analysis as it would better help understand outliers and data distributions in our charts.

We can observe that quite several columns have missing data represented by -88 and -99. This information is useful in performing further analysis as it would better help understand outliers and data distributions in our charts.

Now let us move on to more advanced EDA tools. I then create an EDA report using DataExplorer and SmartEDA. This produced as extensive diagrammatic analysis of the data, lotting bar chart, correlation, scatter plots, ad many more.

The HTML output displays among other: Basic Statistics Raw Counts Percentages Data Structure Missing Data Profile Univariate Distribution Histogram Bar Chart (with frequency) Bar Chart (by SERVICES) QQ Plot QQ Plot (by SERVICES) Correlation Analysis Principal Component Analysis Bivariate Distribution Boxplot (by SERVICES) Scatterplot (by SERVICES)

It also shows there are 33,615 rows, 75 Columns, 73 Discrete columns, and 2 Continuous columns.

The correlation analysis also shows that there are some correlations between variables but due to the number of variables, this is difficult to read due to the volume of data I had to work with, and this further highlights the extent of work required just for data cleanup and preparation.

The principal component analysis sections display feature importance and this is useful for feature selection when creating out samples.

A snippet of the data description is provided below:

Sample size (nrow) 33615 No. of variables (ncol) 75 No. of numeric/integer variables 2 No. of factor variables 44 No. of text variables 29 No. of logical variables 0 No. of identifier variables 0 No. of date variables 0 No. of zero variance variables (uniform) 0 %. of variables having complete cases 100% (75)

From the SmartEDA and DataExplorer report, I reviewed the variable importance based on information value, and this was particularly useful for feature selection, as it displayed the various variables based on their predictive strength and relevance. Referencing the outputs of SmartEDA and DataExplorer for feature selection, I then created a subset of the original data by reducing the number of columns from 75 to 19, using only the columns of high and medium importance. This was then be used for the modeling stage of our analysis in this project.

After Creating a Subset dataframe using columns with High and Median importance features, and adding labels for variables in services, I prepared the data for modeling by splitting the dataset into training and test sets, checking dimension of train and test set to confirm accuracy of split.

Since our dependent variable, Services, has more than 2 possible outcomes, a multi-class classification algorithm was required for the first modeling operation and Using Classification Model was used.

We used rpart.plot for the classification procedure to build the model using type = class. As this is a preliminary analysis, no further investigation was performed. I however tested the model by using the output of the classification to build a model.

As part of the next phase, I will be investigating Accuracy and other measures to determine the effectiveness of this model.

I then proceeded to attempt creating a KNN Algorithm and Logistics regression but ran into issues. The error displayed was mostly because the data were in factor form and the fact that the dependent variable had more that 3 possible outcomes, meant my options were very limited.

After many failed attempts to rectify the errors I was getting, I decided to group two of the outcomes of the dependent into one to allow me use a binomial classification method for classification. This allowed me to use the Logistics Regression algorithm to predict the Services. but first, since Linear regression is useful for binary classification, I needed to prepare the data afresh to be able to use regression analysis.

To achieve this, I created a new column that groups the data into people who use both residential and non-residential services, versus those who only use either. Regression analysis will then be done on this new column as class.

After preparing the data afresh and leaving them as numerical variables to allow the algorithm to run successfully, I did have some concerns with the number of data considered as outliers, mainly -99 and -88 which represent missing values. For the next phase of my analysis, I hope to have enough time to attempt to remove them and compare the results of running the regression analysis without missing values. Only 5 rows remained when I attempted to remove all the rows with NAs, so I will still be forced to work with at least some NAs if I do have time to investigate this in the next phase of my work.

After running the model, from the output of the analysis, we notice that not all variables are significant for prediction, C\_QRB2, C\_QS1, C\_QRB1\_3, PR\_QM1C, and PR\_QO1\_C\_COMBINE are highly significant, while C\_QS2\_3 and C\_QS3 and less significant in that order.

In the next phase of my analysis, I will be creating another model using only the significant variables and will also be running the Anova function to analyze the table of deviance.

Due to time constraints resulting from the amount of work needed to prepare the data for analysis, I will not be able to provide a report on a third classification algorithm at this time. This will be investigated in the next phase of this project.

Thank you.