***Data Mining and Machine Learning for a Classification Problem***

***Predicting Voters’ Inclination to Vote Democratic and Applying Uplift Modelling for Treatments.***

Abstract - The study seeks to predict voters’ persuasion towards a particular direction. The data set explored was voter persuasion which was downloaded from the Kaggle website. The data is made up of 10,000 rows and 79 columns which are the variable set. Multicollinearity was used to select the most accurate variables to carry out the classification. Two models were created using machine learning algorithms and these were the Light Gradient Boosted Machine Classifier and the Decision Tree Classifier. The findings obtained by both models were highly outstanding, and they illustrate their capacity to classify whether or not a voter is likely to switch their vote. The LGBM classifier gave a higher AUC score and thus, it was the best fit model for classification as it showed that NH\_WHITE was the best predictor for movement in the direction of voters.

***Keywords: LGBM, Voter Persuasion, Decision Tree Classifier, Prediction, Classification, ROC Curve, Uplift Modelling***

1. INTRODUCTION

This study seeks to predict voters’ persuasion towards an election such that they can come up with the possible outcomes of electoral events. The goal is to use various analytical methods towards making this prediction. Analyzing what propels the electoral decision of citizens is important as it helps to make valid inferences for stakeholders in cases where necessary. For predicting elections and voters’ persuasion, measurements via instruments are important and this is usually done via surveys or projections. This study finds applicability in the use of projections for analyzing voters’ behavior.

It has been well agreed and analyzed that campaigns have become less persuasive in changing the mind of citizens towards election as predictive analysis has gained more grounds. It was explained that recent campaigns now make use of data in various ways to the end that the campaign data will help to provide information of the likely events and occurrences as well as give information on those that would be contacted for the success of the election. [1]

Campaign data produces predictive models which produce predictive scores which produce three different types of scores from voters’ databases and this includes the behavior scores, predictive scores, and support scores [1].

The first two types of scores which are the behavior and support scores predict the behavior or attitude of voters. This is done to emphasize the focal traits of the citizen thereby avoiding overfitting of data. This finds application for this study.

Supervised machine learning which includes classification and regression trees is typically more appropriate for analyzing and modelling political data [2]. This method has an advantage as it is relatively scalable. Specifically, classical predictive modeling that classifies voters as moving towards the democratic side or not using LGBM and Decision Tree Classifiers will be used.

II. DESCRIPTION OF DATASET

The data, here for the study, are in the file Voter-Persuasion.csv. The data was downloaded from the Kaggle website[3]. The target variable is “MOVED\_AD”, where 1 = “opinion moved in favor of the Democratic candidate” and 0 = “opinion did not move in favor of the Democratic candidate.” This variable contained all of the information gathered from the pre-and post-surveys combined. Among the important predictor variables is Flyer, which is a binary variable that indicated whether or not a voter had received a flyer from the campaign.

III. PROCESSING AND EXPLORATION

1. *Data Preprocessing*

After the data was read in Python, it was seen that the data contains 10,000 rows and 79 columns. The variables in the column consist of both numerical and categorical variables. No missing values were found in the data set. Hence, there was no need to treat columns for missing values in the data presented. The target label was identified as the MOVED\_AD variable. From the value counts of the variable, it was observed that 6266 respondents were classified No while 3734 respondents were classified Yes. The summary statistics for the numerical variables in the data were then presented. (See Appendix II).

Further preprocessing tasks were performed on the data. The MESSAGE\_A column of the data set describes whether a voter got the flyer or not. If the voter gets the flyer, a score of 1 was assigned as its representation while 0 represented not getting the flyer. We further calculate how excellent the flyer did in moving the voters to the democratic side using the moved variable against those who got the flyers and those who did not. It was then discovered that 20% of voters who got the flyers were moved while 17% of voters who did not get the flyer moved. It shows that the flyer did well in moving the voters generally.

1. *Exploration*

To gain a better understanding of the relationships between the predictor variables and the MOVED\_AD, the data were analyzed. After showing supporting charts and tables, we select the predictors that appear to have the best predictive potential.

Exploratory Data Analysis was performed on some features using boxplots. They are useful in classification tasks because they allow you to evaluate the potential of numerical predictors before you use them. This is accomplished by using the x-axis to represent the categorical outcome and the y-axis to represent a numerical predictor on a graph. The first set of examples shown below helps us to see the effects of SET\_NO, OPP\_SEX, AGE, HH\_ND, HH\_NR, HH\_NI, MED\_AGE, NH\_WHITE, NH\_AA, NH\_ASIAN on MOVED\_AD. These pairs do not separate the outcome variable so we will use the correlation plot to select potentially useful variables. (See Appendix III).

We change the datatype of categorical columns and one-hot encode them before getting their correlation values.

1. Correlational Analysis

The use of correlation analysis to look for duplicates in the data is a powerful tool. It is not uncommon for the same variable to appear many times in a dataset (under various names) since the dataset was compiled from various sources, the same phenomenon is measured using different units, etc. A correlational analysis is carried out to examine the multicollinearity of variables. Multicollinearity occurs when two or more independent variables are highly correlated, meaning, they have the same linear relationship with the outcome variable. Multicollinearity effect should be avoided as highly correlated variables can override the model. [4]. Looking at the correlation matrix allows us to find redundancy. This creates an avenue to remove irrelevant predictors, helping to achieve data reduction. Avoiding multicollinearity issues in various models can be achieved by eliminating variables that are significantly connected. (See Appendix IV).

After this was done, using the correlation table, the analysis focused on the correlation figures of predictors with the target variable (MOVED\_AD). The predictors that reduce the effect of multicollinearity were selected and assumptions were made on the variables that have a good predictive potential. These variables include: HH\_ND 0.259323; HH\_NR -0.279504; NH\_WHITE -0.096613; PARTY\_R -0.415191; VPP\_08 0.115435; UPSCALEMAL -0.005611; MESSAGE\_A 0.059954; CAND1S\_S -0.657461; CAND2S\_S -0.187660; CAND1\_UND\_Y 0.366541.

1. Testing Measures of Central Tendency, Shape, and Spread

This was done using the getdistprops function in Python. It takes a series and generates measurements of central tendency, shape, and spread by using the getdistprops function. These values are returned in a dictionary by the function. When the Shapiro test for normality fails to yield a value, it is likewise handled by this method. When this occurs, it will not add keys for normstat and normpvalue.[4]. This explains the skewness, kurtosis, and the type of distribution data for each selected variable. See Appendix V-XIII. The summary of the result is presented below.

1. HH\_ND: positively skewed, leptokurtic, multimodal
2. NH\_WHITE: negatively skewed, platykurtic, multimodal
3. PARTY\_R: positively skewed, platykurtic, bimodal
4. VPP\_08: positively skewed, platykurtic, bimodal
5. UPSCALEMAL: Positively skewed, flattened tails, leptokurtic
6. MESSAGE\_A: perfect symmetry, bimodal, platykurtic
7. CANDIS\_S: negatively skewed, platykurtic, bimodal
8. CAND2S\_S: negatively skewed, leptokurtic, bimodal
9. CANDI\_UND\_Y: positively skewed, bimodal, platykurtic.

From the normpvalue obtained and as shown in the distribution plots for each of the predictor variables selected, it is obvious none of them are normally distributed. However, the rule-based algorithms that will be utilized for modeling don't need feature scaling (Standardization and Normalization).

IV.DATA MINING, MODELLING, AND METHODOLOGY

Data was then partitioned using the selected independent variables in the dataset. Only variables with good predictive models that fit two models accordingly were selected. The performance of each of the models is evaluated using the confusion matrix and ROC curve. One way to summarize the correct and faulty classifier outputs is to use the confusion matrix or classification matrix. The ROC curve is a commonly used graphical representation for evaluating a classification model's prediction accuracy across a wide range of classification thresholds. The False Positive Rate (FPR) is plotted on the X-axis and the True Positive Rate (TPR) is plotted on the Y-axis in the ROC curve.

TPR(Sensitivity) = TP/(TP+FN)

FPR(1-Specificity) = FP/(TN+FP)

The data was split into the train and test sets. 60% were used as the train set while 40% as the test set. Modelling was done with LGBMClassifier and the Decision Tree Algorithms using Python. We use a derived variable that is the opposite of Flyer(MESSAGE\_A). This is represented as MESSAGE\_A\_REV in the data. Instead of Flyer, we use the Flyer reversed variable as a predictor in our best-chosen model to re-score the validation data. The first three records in the validation set are included in this report.

LGBMClassifier: This is a gradient boosting framework that uses a tree-based learning algorithm. This has an advantage because it focuses on the accuracy of results. It finds application in its wide range of variable acceptability.

Decision Tree Classifier: They are visual representations of decisions and the potential outcomes of those decisions. Nodes and branches are the building blocks of a network. The nodes represent the entirety of the sample and are then divided into multiple sets, with branches illustrating the various options available to the user. For each feature and branch of a decision rule, there is a node that represents the result. The Decision Node and the Leaf Node are two of the nodes in a Decision Tree.

V. ANALYSIS AND EVALUATION OF RESULTS

Using the LGBM Classifier, it was seen that NH\_WHITE is the most important feature for predicting MOVED\_AD, that is, the movement towards the democratic direction. This model was confirmed with an 87.6% accuracy for the confusion matrix and a 0.95 AUC Score. (See Appendix XV)

For the Decision Tree Classifier, it was revealed that CAND1S\_S is the most important feature for predicting movement in the democratic direction. The confusion matrix of this classifier also has 87.2% accuracy with the AUC Score having 0.94.

In terms of predictive power, the best model is the light GBM classifier model. It was chosen because it has the highest AUC score.

Furthermore, using a cut-off of 0.5, we report the propensities for the first three records in the validation set for the lightgbm model.

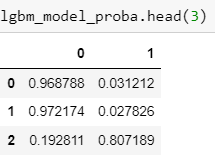


Figure 2: LGBM Model

The first record has a 3.12% propensity of moving in favour of a democratic candidate. The second record has a 2.78% propensity of moving in favour of a democratic candidate. The third record has an 80.72% propensity of moving in favour of a democratic candidate.

For further analysis, a derived variable that is the opposite of Flyer(MESSAGE\_A) was used. This is represented as MESSAGE\_A\_REV in the data. Using the best-chosen model, we re-score the validation data using the Flyer-reversed variable as a predictor, instead of Flyer. We report the propensities for the first three records in the validation set. The LGBM classifier which was the fittest model was used; the confusion matrix and AUC Curve are 87.85% and 0.95 respectively (Appendix XV). The first record has a 66% propensity of moving in favour of a democratic candidate. The second record has a 1.66% propensity of moving in favour of a democratic candidate. The third record has an 80.38% propensity of moving in favour of a democratic candidate.

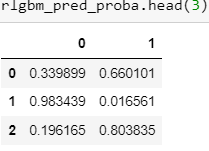


Figure 2: Reverse LGBM Model

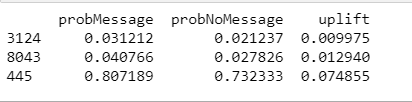
The uplift score is calculated using this formula: P(Success | Flyer = 1) – P(Success | Flyer = 0). Our validation set includes the first three records, therefore we compute the uplift for each voter.

Figure 3: Uplift of first Three Voters in the Validation Set

There were three values here: 0.0099, 0.0129, and 0.0749. This uplift model will be used to determine whether or not to send a persuasive message or possibilities from sent messages to all voters with positive uplift.

When it comes to marketing and politics, uplift modeling is most commonly used. It serves two primary functions:

(i) To determine whether or not you should try to influence someone, or if you should just leave them alone.

(ii) Choosing a message to send from a list of possibilities when a message is going to be sent.

VI. CONCLUSION/FUTURE WORK

It is possible to experiment with no message, A, and B, as well as alternative treatment options. The distinction between the two is important to practitioners, who prioritize the first. Consumers who would buy or renew subscriptions anyway are not interested in receiving discount offers. The same is true for political campaigns. A message or offer with a negative effect is especially undesirable for both parties.

The study concluded that accurate predictions of voters’ persuasion and voters’ behaviour can be made via classification models and this can be used to predict if voters will respond to a message or move in a particular voting direction or not.

REFERENCES

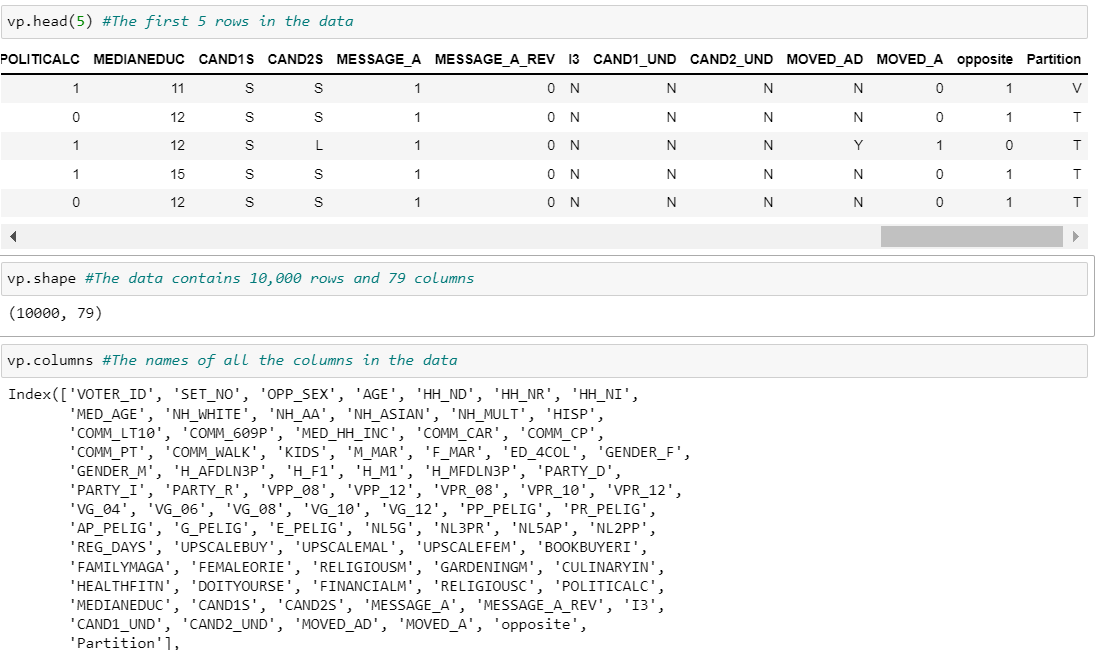
[1] Nickerson, D. W., & Rogers, T., “Political Campaigns and Big Data”, Journal of Economic Perspectives, 28(2), 2014. 74, —Pages 51–

[2] Rusch, T., Lee, I., Hornik, K., Jank, W. &Zeilei, A., “Influencing elections with statistics: targeting voters with logistic regression trees”, The Annals of Applied Statistics 2013, Vol. 7, No. 3, 1612–1639 DOI: 10.1214/13-AOAS648 c Institute of Mathematical Statistics, 2013

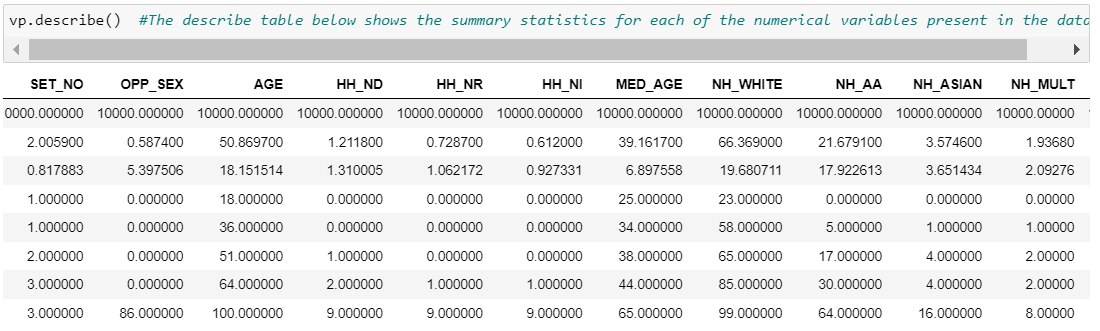
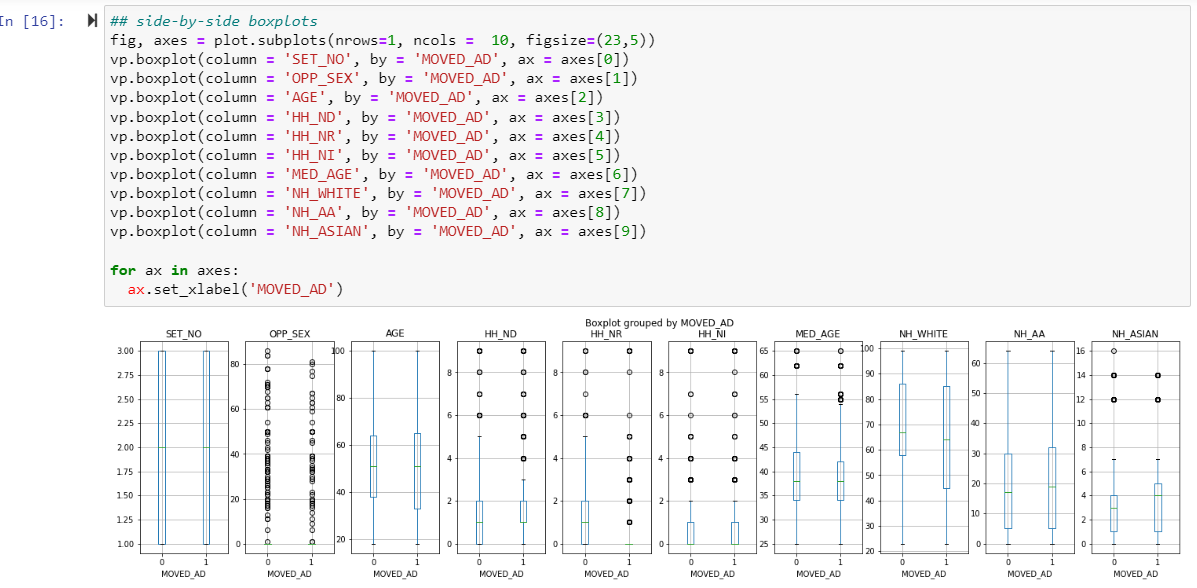
[3]<https://www.kaggle.com/aakarkale/voterpersuasiondataset/code>

[4] Brandusoiu I, Toderean G, Ha B. Methods for churn prediction in the prepaid mobile telecommunications industry. In: International conference on communications. 2016. p. 97–100.

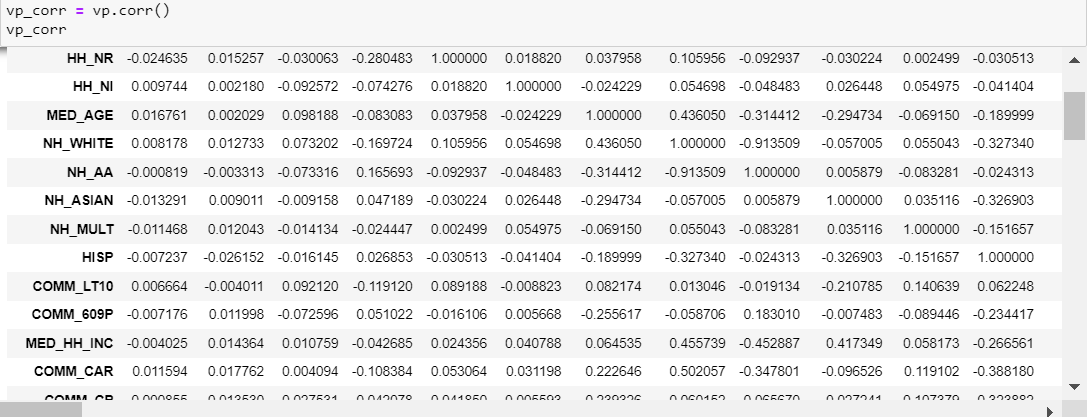
Appendix I: The first 5 rows of the dataset



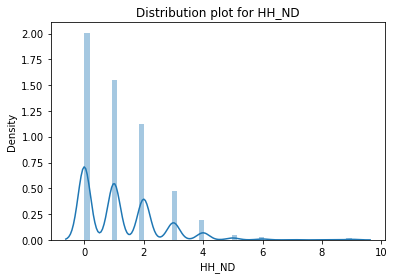
Appendix II: Summary Statistics for Numerical Variables

Appendix III: Box plot for Some Independent Variables

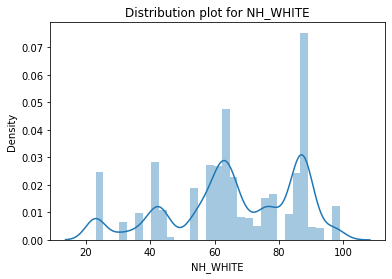
Appendix IV: Correlation Matrix for all Variables



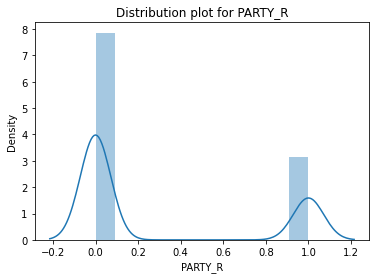
Appendix V: Distribution Plot for HH\_ND Variable



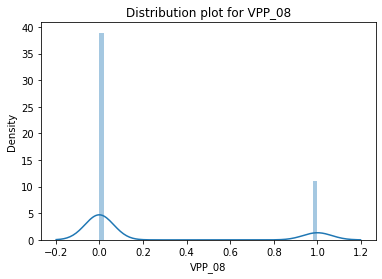
Appendix VI: Distribution Plot for NH\_WHITE Variable



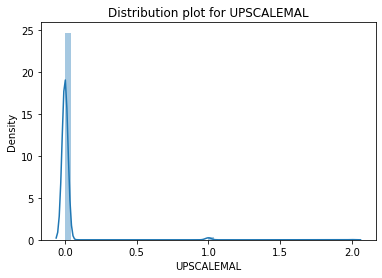
Appendix VII: Distribution Plot for PARTY\_R variable



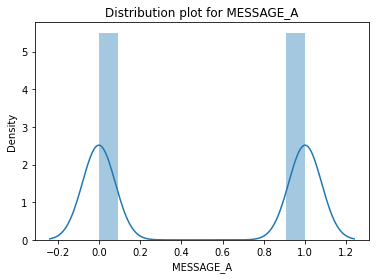
Appendix VIII: Distribution Plot for VPP\_08 Variable



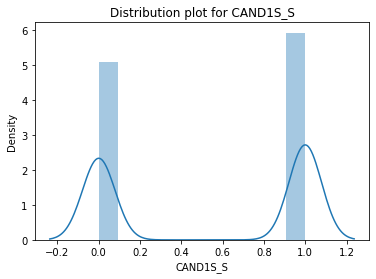
Appendix IX: Distribution Plot for UPSCALEMAL Variable



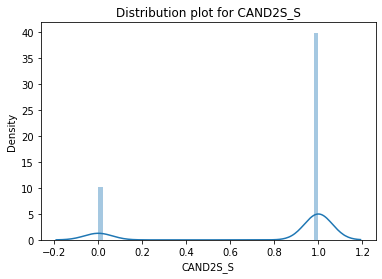
Appendix X: Distribution Plot for MESSAGE\_A Variable



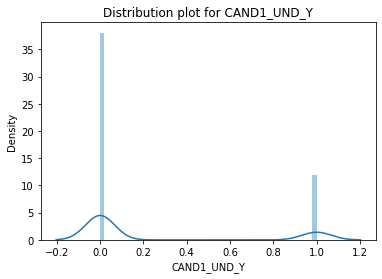
Appendix XI: Distribution Plot for CAND1S\_S Variable



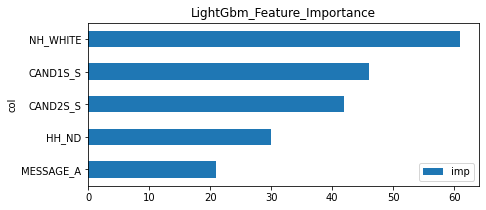
Appendix XII: Distribution Plot for CAND2S\_S Variable



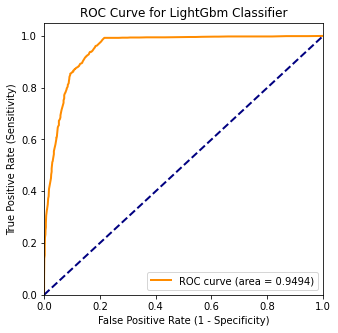
Appendix XIII: Distribution Plot for CAND1\_UND\_Y Variable



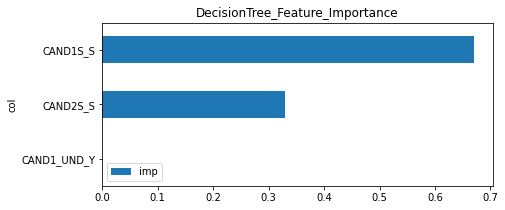
Appendix XIV: LightGBM Feature Importance



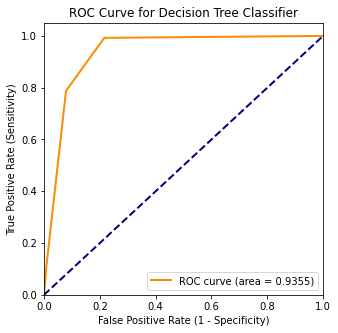
Appendix XV: ROC Curve for LightGBM Classifier



Appendix XVI: Decision Tree Feature Importance



Appendix XVII : ROC Curve for Decision Tree Classifier



Appendix XVIII: ROC Score for LightGBM when MESSAGE\_A\_REV is used

