Project: Grover Data Science Assessment
Author: Adegboyega Adesanya
Date: Feb 28, 2022

Project Overview: This is a binary classification project that aims to classify a certain target variable "y" optimally.

General Observations and Findings:

1. The dataset is of dimension (2671,21) meaning there are 2671 samples or observations in the dataset and 21 features or variables including the target variable (" y "). Making it 20 independent variables and 1 dependent variable y.
2. At first sight, the dataset has a general data type of "object". A closer look and subsequent steps to data cleaning sorted that out.
3. It is observed and assumed that " ? " represents missing values usually "NaN " in normal dataset and treated as such. As such variables with high " ? " was dropped during the task.
4. Some variables have mixed data in them. i.e., having strings and supposed numerical data.
5. Numerical data are comma " , " separated to indicate decimals instead of being separated with periods " . "
6. Also, some element appears in double forms e.g. "bb", "ff"
7. Another major observation noticed is the imbalance in the dataset. The target variable (y) is not evenly distributed or at least close in proportion. There are 2398 Good samples and 273 Bad samples.

All that was observed was treated accordingly during data preprocessing/data cleaning steps

General Approach

Data Understanding:

To understand the data, I checked the shape, info of the dataset and did a value count which returns the unique element and their count in each variable.

Data Cleaning Approach:

1. **Handling Missing Values (" ? "):** variables with missing values greater than 10% of the variable population was dropped. "x.16" and "x.20" fell in this category. Using 10% is almost like a rule of thumb as variables with missing values up to 10% can still be handled without bias by either replacing missing values with mean or median if it is a numerical variable or replacing with the modal class in case of categorical variables. This was the approach taken to handle missing values or " ? " in the dataset.

2. **Mixed Data Variables**: Most of the variables with mixed dataset are the numerical variables, the foreign elements ("f", "b", "t", etc.) in these variables are replaced with the median of the numerical variables. I opted for median instead of mean because of the effect outliers have on mean as a statistic.

3. **Numerical Values with commas**: This is handled by replacing "," with "." And converting the variable to float.

Data Preparation Approach:

1. **Feature Selection**: I used chi2_contingency approach for selecting relevant categorical variables and ANOVA F-statistic for selecting numerical features. These methods have been proven to work with those data types. They both estimate some statistics with p-value being one of them. I basically select variables with p-value < 0.05 which represents their significance in predicting the target variable.

2. **Check for Correlation**: correlating variables tend to predict themselves rather than the target variable. Since they hold information about the target variable, it is best to use just one of these features in training our model. I used a correlation > 0.90. as variables with this much correlation would affect model performance and will cause unnecessary redundancies in the dataset. Variable "x.17" has a correlation of 1 with "x.1" meaning they are giving exactly the same information about the target and variable "x.18" has a 0.98 correlation with "x.2". Variables "x.17" and "x.18" were dropped.

3. **Categorical variable Encoding:** Machine learning algorithms don't work with strings as such we encode all categorical variables in the dataset. Categorical variables can be nominal or ordinal. Ordinal meaning, they have a rank to them and as such their encoding should represent their ranks. Nominal variables however don't have ranks. As such they are one-hot encoded by creating dummy variables. These dummy variables just indicate the presence or absence of each element that make up the variable using 1s and 0s. 1 – variable present, 0 – variable absent. Dummy variable trap was avoided by dropping the first dummy variable and also dropping the original variable creating the dummy to avoid unnecessary correlations.

4. **Dataset Balancing:** Synthetic Minority Oversampling Technique (SMOTE) was used to balance the very imbalanced dataset we have. This technique helped balance the original 2398 Good samples and 273 Bad samples. As such during training, we have 2398 Good samples and 2398 Bad samples. SMOTE used k nearest neighbor technique in creating data similar to the minority class therefore balancing the dataset as a result. The advantage of this is, we don't get to drop samples from the majority class by undersampling to balance our dataset.

Modelling Approach:

1. **Machine Learning Algorithm Selection:** Cross validation was used in algorithm selection. While cross validation is majorly used for preventing overfitting, it has been proven to help in selecting the best algorithm that fits the data best. As such this approach was used and Random Forest classifier was selected as the best to fit the data.

2. **Hyperparameter Tuning:** This approach helps in getting the best out any machine learning algorithm by tweaking and selecting parameters of the algorithm to optimize the model performance. Random forest was tuned and the best parameters were used in training.

Result:

The resulting model has a 99% accuracy on training data and 93% on validation data. As shown in the confusion matrix in the notebook. It was able to predict 320 of 336 Bad samples correctly and 138 of 154 Good samples correctly.

In conclusion, the model is performing well but can be improved on by getting more Bad sample dataset to be used in training of the model instead of generating data using SMOTE.