**Module 2 (Part II) – Exploratory Data Analysis**

**Name: Maviya Shaikh**

**Student number: 100766785**

**Date: March 13th,2020**

**Table of Contents**

[1. Problem Statement and Analytical Statement 3](#_Toc34988314)

[2. Exploratory Data Analysis (EDA) Action Plan 3](#_Toc34988315)

[3. Exploratory Data Analysis (EDA) 4](#_Toc34988316)

[i. Summary of EDA 4](#_Toc34988317)

[ii. Insights 5](#_Toc34988318)

[iii. Analytical Score Card 5](#_Toc34988319)

[iv. Target 6](#_Toc34988320)

[4. References 6](#_Toc34988321)

1. Problem Statement and Analytical Statement

**Problem Statement**: The analysis for the test of TNM stage 1 pre-screening is conducted to identify whether the patient has cancer or not.

**Analytical Statement:** The model is developed to identify the patient has stage 1 cancer or not.

1. Exploratory Data Analysis (EDA) Action Plan

* Convert the dataset into **binomial** as it is multi-class dataset.
  + The dependent variable class has three values 0 represent no cancer, 1 represent stage 1 cancer with low risk and 2 represent stage 1 cancer with high risk. However, problem statement is to identify patient has cancer or not, therefore we will **combine 1 and 2 as stage 1 cancer using single naming convention 1**
* Basic Statistics is calculated using pandas profiling and describe() method:
  + **Pandas Profiling**: It gives the high-level analysis of the dataset that includes number of variables, observations, missing values, duplicate rows and memory occupied. Apart from this, it will also show the correlation between variables
  + **Describe method:** The describe() method is used to get the brief overview of the dataset with the total number of counts, mean, median, standard deviation, minimum and maximum
* The missing values are fixed by removing the values from the dataset
* The outliers are detected using Tukey method:
  + **Tukey Method**: This method will help in detecting and removing the outliers that lies beyond **1.5** times the Interquartile Range
* The dataset is balanced using SMOTE technique:
  + **Synthetic Minority Over-Sampling Technique (SMOTE)**: This technique will synthetically create the data for minority class to balance the dataset. So, that model will not be biased towards specific class
* The data is scaled using Standard Scalar or MinMax Scalar:
  + **Standard Scalar**: This scalar method rescales the value of the data points in the range of mean 0 and standard deviation 1
  + **MinMax Scalar**: This scalar method rescales the value of the data points within the range of 0 to 1

The following graphs will be used for the analysis of the dataset:

* **Boxplot**: The boxplot shows the significance of each variable and the distribution of the data around the mean. It also helps in identifying the outliers in the dataset
* **Histogram**: The histogram will help identify whether the data is normal, right skewed or left skewed. It also represents the occurrences of the specific value in the dataset and also detect the outliers in the dataset
* **Heatmap**: The heatmap will determine the intercorrelation between the variables that will help in selecting the model

1. Exploratory Data Analysis (EDA)

## **Summary of EDA**

* After converting into binomial, the dataset has 9 independent variables V1, V2, V3, V4, V5, V6, V7, V8 and V9 that represent the cell structure of the body and one dependent variable class that represent **no cancer if it is 0 and cancer if it is 1**
* **Analysis of Basic Statistics:**
* With the help of **pandas profiling**, it was found that dataset has total 1690 observation with 100 missing cells and 433 duplicate rows. In addition, all the independent variables are numerical continuous values whereas dependent variable is Boolean i.e. 0 or 1. Also, it shows the warning that most of the variables are highly correlated and the class has unbalanced data for no cancer and cancer. **Note**: The duplicate rows will not be removed because it seems to be duplicate records, but it is original data of different patients
* With the help of **describe() method**, the number of observations is 1680 that indicates this method does not count missing values. The median value of all the independent variables is less than the mean value except V9 which represent that data for V1 to V8 is right skewed, whereas for V9 the median value is more than the mean value that represent V9 data is left skewed
* There are **10 rows** that has missing values which is equal to 100 cells. These values are removed in the dataset as there is not a single value in the row.
* Tukey method is used to detect the outliers and it detects that there are **120 rows** represented as outliers. The decision to remove or keep outliers will be taken based on the model selection
* The **dataset is not balanced**, so to balance the dataset SMOTE technique will be used that will remove the biased and synthetically create the datapoints for minority class. This method is dependent on the train size used for the model as it creates the synthetic values for the trained data
* The **dataset is not normalized** and to normalize the data, the standard scalar or minmax scalar will be used depending on the performance of the model
* **Analysis of the graphs:**
  + With the help of **boxplot**, it is determined that dataset contain outliers and the distribution of the dataset is not normal
  + With the help of **histogram**, the distribution of the independent continuous variables is drawn i.e. for V1 to V8 variables, distribution is right skewed or positively skewed whereas for V9 variable, distribution is left skewed or negatively skewed. Also, the outliers can be easily identified by looking at the bars of the histogram which is quite far away from most of the distribution
  + With the help of **heatmap**, the correlation between variables are determined that except V3 and V6 all other independent variables are highly correlated with **more than 90% correlation**

## **Insights**

From the Exploratory Data Analysis, the key insights developed that will help in model selection process are as follows:

1. **The dataset is not balanced:** This nature of the dataset will restrict the use of certain metrics in the machine learning models. So, while selecting the evaluation metrics, this insight should be considered. Moreover, there is a constraint that it is not possible to collect or add more observations into the dataset to balance it. However, we can use the SMOTE technique to balance the training data to eliminate the bias in prediction.
2. **The dataset is not normally distributed:** The provided dataset is not normally distributed as there is skewness in the data points. Also, there are outliers in the dataset that adds more skewness. However, we can remove the skewness of outliers with Tukey method, but if the original data still has the skewness then we cannot select the model that assumes ‘data is normally distributed’.
3. **The features of the dataset are highly correlated with each other:** According to our assumption that all independent variables in the dataset are useful to predict the outcome variable, we cannot remove any features from dataset to eliminate high correlation. Thus, it needs to be considered while selecting the model because some of the models are sensitive to high correlation of the features.

## **Analytical Score Card**

The number ofevaluation metrics that will be used to analyse the model are described in the below table:

|  |  |  |  |
| --- | --- | --- | --- |
| **Sr. No.** | **Metrics** | **Explanation** | **Its use in the project** |
| 1 | Accuracy | It is the ratio of number of correct predictions to the total number of predictions made. Thus, accuracy represent the percentage of predictions that model got right. This metric will give the correct result only if there is equal distribution of the class. | This metric will be used to compare the result of different models. Thus, it will help in selecting the best model. |
| 2 | Precision | It tells how often the model is correct when it makes the prediction. | This metric will use to analyse the specific class in the dataset. |
| 3 | Recall | It is the ratio of correctly classified the positive instances or negative instances from the total number of positive instances or negative instances respectively. | This metric will identify all the relevant instances from retrieved instances. Thus, it will use to analyse the specific class. |
| 4 | F1-score | It is the harmonic mean of precision and recall. If there is an uneven distribution of the class then this metric is considered for evaluation of the algorithm. | This is the **main metric that will be used to evaluate the model** as our dataset is unbalanced. |
| 5 | AUC-ROC curve | The full form of the AUC – ROC is Area Under the Curve – Receiver Operating Characteristics. It is a performance measurement tool for classification problem to know the performance of the model in differentiating classes of the dataset. | This is also an **important metric that will help to know the capacity of the model on how well it distinguishes the classes** **No cancer (0) and cancer (1)**. |

Table 1: Analytical Score Card

## **Target**

The problem description is quite sensitive in which the model is predicting whether the patient has cancer or not. Therefore, targets are set for the metrics after achieving that only the model should be in the useable form.

|  |  |  |
| --- | --- | --- |
| **Sr. No.** | **Metrics** | **Targets** |
| 1 | Accuracy | Should be greater than 95% (to select the model) |
| 2 | Precision | Should be greater than 95% (to analyse the specific class of the dataset) |
| 3 | Recall | Should be greater than 95% (to analyse the specific class of the dataset) |
| 4 | F1-score | Should be greater than 95% (to analyse the overall model) |
| 5 | AUC-ROC curve | AUC should be at least 0.95 (For instance: if AUC=0.95 there is 95% chance that model will be able to distinguish between no cancer and cancer.) |

Table 2: Targets for evaluation metrics

1. References

* Narkhede, S. (2019, May 26). Understanding AUC - ROC Curve. Retrieved from <https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5>