Ethical Approach for Handling Imbalanced Data

February 24, 2022

Registration number:	2111651
Project:	Imbalanced datasets
Link to GitHub:	`https://github.com/darshan16011989/Imbalanced_Datasets`

Executive summary (max. 250 words)	199
Introduction (max. 600 words)	382
Data (max. 500 words/dataset)	683
Methodology (max. 600 words)	434
Conclusions (max. 500 words)	218
Total word count	1916

Contents

Figure 1: Imbalanced data

Abstract

A dataset is said to be imbalanced if the distribution of the class or target variable is not equal. To study this, we will be loading the data which has two classes in target variable. The class proportions will be perfect (50:50) or slightly imbalanced (60:40). Now by techniques like under/over sampling, we will be introducing the imbalance in the data. Three surrogates are created from the data by sub-sampling one of the classes. It will be named as shown below

1. Low imbalance (65:35)
2. Medium imbalance (75:25)
3. High imbalance (90:10)

The machine learning classifier model performs well when the class proportions are balanced. If the dataset is imbalanced, the performance of the model reduces and tends to be biased towards the majority class. If the requirement is to predict the minority class, then standard model performance measures will not be effective and might need some modifications.

We can find an imbalance in the data where there are normal covid cases and omicron variant cases. Number of omicron variant cases are usually lower than normal covid cases. This results in producing imbalance in the data.

We will be using three datasets. We will be predicting the satisfaction of customer in flight services, predicting churn rate of customers for telecom industry, and predicting heart disease from patient data.

To train a model, we will be using techniques like clustering for segmenting the data and ensemble techniques like Random Forest. We will try to get the perfect blend of both supervised and unsupervised learning to solve the problem.

1 Introduction

Machine Learning algorithms play important role in predicting or forecasting from the given data. Basically, in supervised learning we can divide the predicting model in two types i.e. classification and regression. If there are no labels for target variables in the data, techniques like clustering and Principal Component Analysis (PCA) are used. Clustering is used to segment the data while PCA is used to reduce the dimensions or features from the data. Clustering technique can even act as a pre-processing step before fetching it to supervised learning model.

The dataset needs to be balanced before applying it to Machine Learning model. It means, the target class should have almost equal distribution for both the classes. But, in real world problems we need to often deal with imbalance in the data. The predicting power of algorithm decreases and tends to be biased over the majority class. Mostly we are interested in the predicting minority class like attrition rate for an organisation, classification of ham and spam and so on.

Techniques like under/over sampling can be applied to make data balanced. Depending on the size and nature of the data, under or oversampling method is decided. In under sampling, we randomly remove n number of data points from any class of the target variable. This results in imbalance in the data. Whereas, in over sampling we intentionally n number of data points to the required class of target variable. The best method for sampling is the SMOTE(Synthetic Minority Oversampling technique), which generates artificial data similar to the existing data and it insures that there are no duplicates created.

To know the problem, we will use the data which is already balanced and then we will make it imbalance with different ratios in target variable. We will be using combination of supervised and unsupervised learning to solve the problem. We will be using ensemble technique called Random Forest for predicting the target variable. Random forest consists of n number of decision trees with randomised features and then takes the best model by voting principle. To explore more and to make our model rigid, we will also perform clustering or segmentation for the data. This will overall add a good value to our data as well as some power for boosting the accuracy of the model.

.

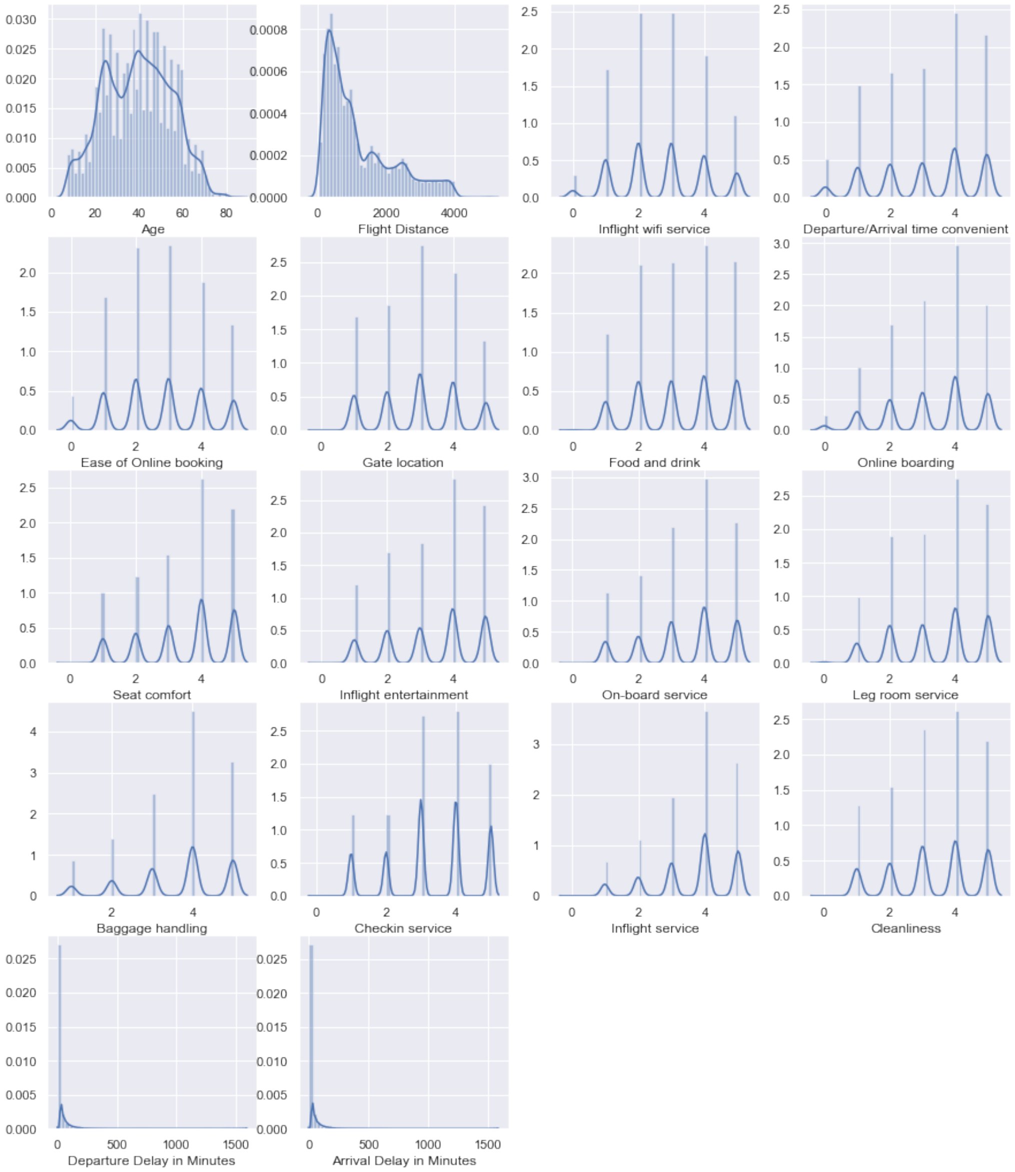

Figure 2: Histogram or Distribution plots

2 Data

2.1 Airline Passenger Satisfaction dataset

2.1.1 Basic Exploratory Data Analysis (EDA)

1. The data was downloaded from kaggle[1].

2. There are 129880 rows and 25 columns/features in the data.

3. There are 5 categorical and 20 numerical features in the data.

4. Data is almost balanced in target variable with ratio 56:44 for ’neutral or dissatisfied’ and ’satisfied’ respectively.

2.1.2 Distribution of numerical variables

1. Age feature is almost normally distributed. Features Flight distance, Departure Delay in Mins and Arrival delay in Mins are right skewed.

2. Apart from above features, rest features have natural clusters within them, hence we can identify them as categorical also. Since the value ranges from integers 0 to 5 only.

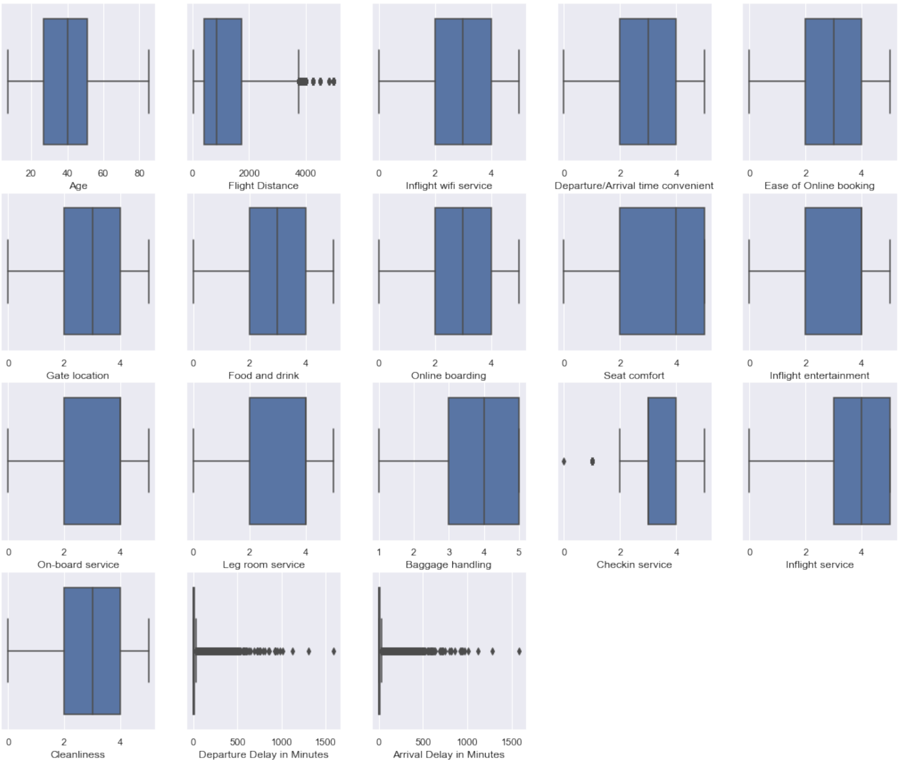

Figure 3: Box plots

2.1.3 Box plots for numerical variables

1. There are outliers in flight distance, arrival delay in Minutes, Departure delay in Minutes and check in service feature.

2. Other features do not have outliers.

2.1.4 Count plots for categorical variables

1. Most of the travel is done for business purpose compared to personal.

2. Customer satisfaction is comparatively poor

3. There are more loyal customers

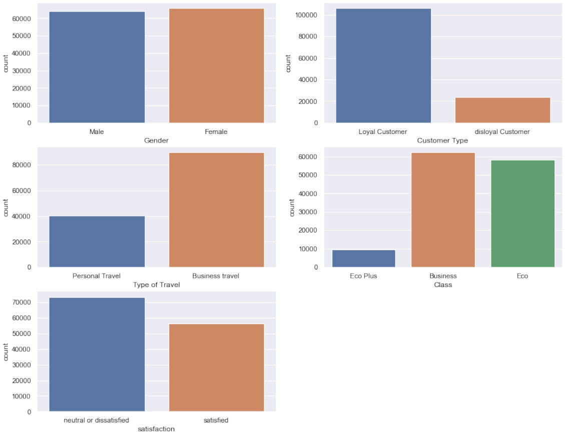

Figure 4: Count plots

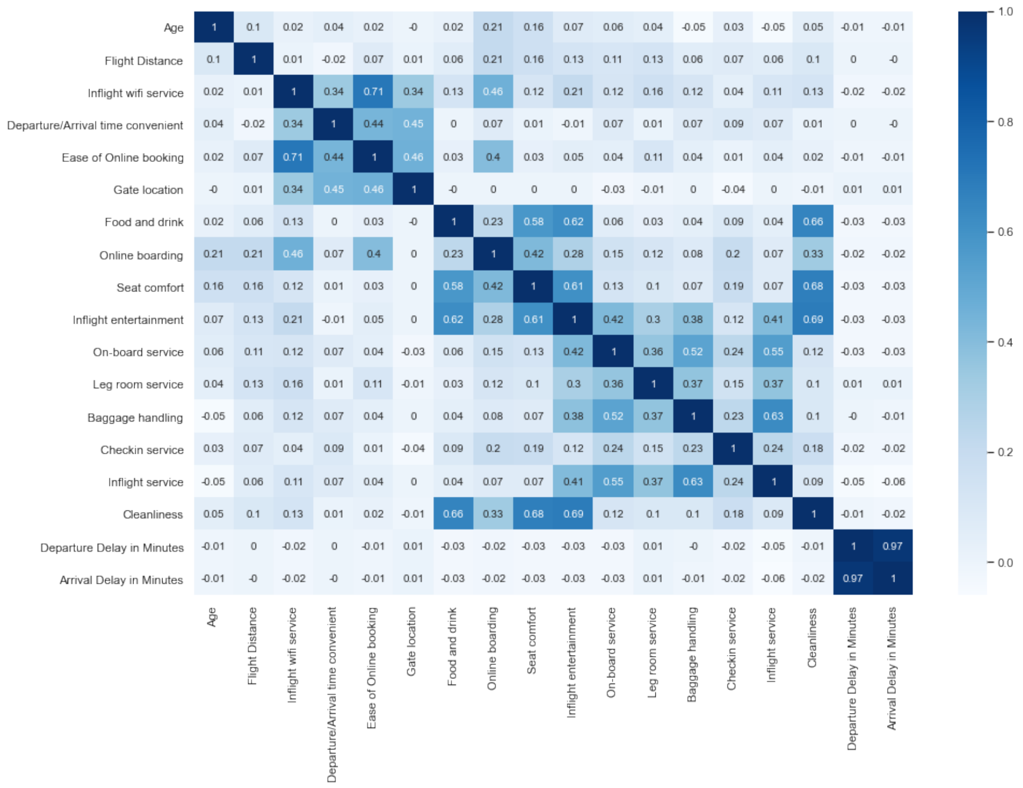

Figure 5: Heatmap

2.1.5 Correlation or Heat maps

1. We can see a strong correlation between Departure Delay in Minutes and Arrival Delay in Minutes.

2. There is a good correlation between ease of booking and in flight WiFi services.

3. There is also a good correlation between cleanliness, fooddrink, seat comfort and in flight entertainment.

2.1.6 Pre processing steps

1. There are 393 missing values in Arrival Delay in Minutes feature. Since the value is very small (0.3 %) compared to data points, we will drop missing values.

2. We will perform one hot encoding for categorical features.

3. Since Random Forest classifier is not sensitive to outliers and feature scaling, we will train the model using one hot encoded data directly.

4. When performing clustering we will perform feature scaling as well as treating the outliers as clustering is a distance based algorithm.

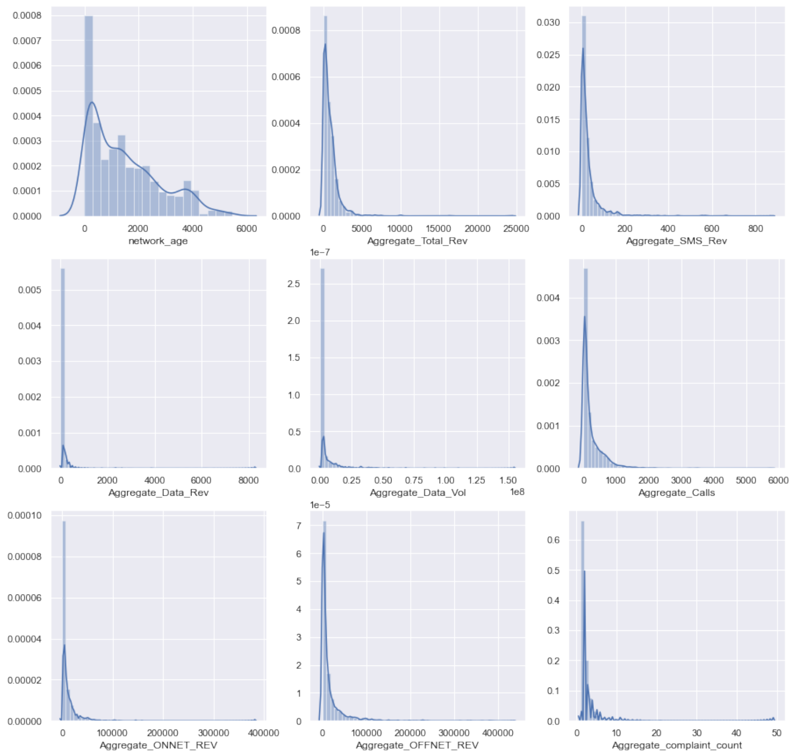

Figure 6: Histogram or Distribution plots

2.2 Churn Prediction for Telecom Industry

2.2.1 Basic Exploratory Data Analysis (EDA)

1. The data was downloaded from kaggle[2].
2. There are 2000 rows and 14 columns/features in the data.
3. There are 5 categorical and 9 numerical features in the data.
4. Data is perfectly balanced in target variable with ratio 50:50 for Churned and Active.
5. There are missing values in the data.
6. There are no duplicate records.

2.2.2 Distribution of numerical variables

1. The distribution of data appears to be right skewed for all features.

2.2.3 Box plots for numerical variables

1. All features have outliers in the data.

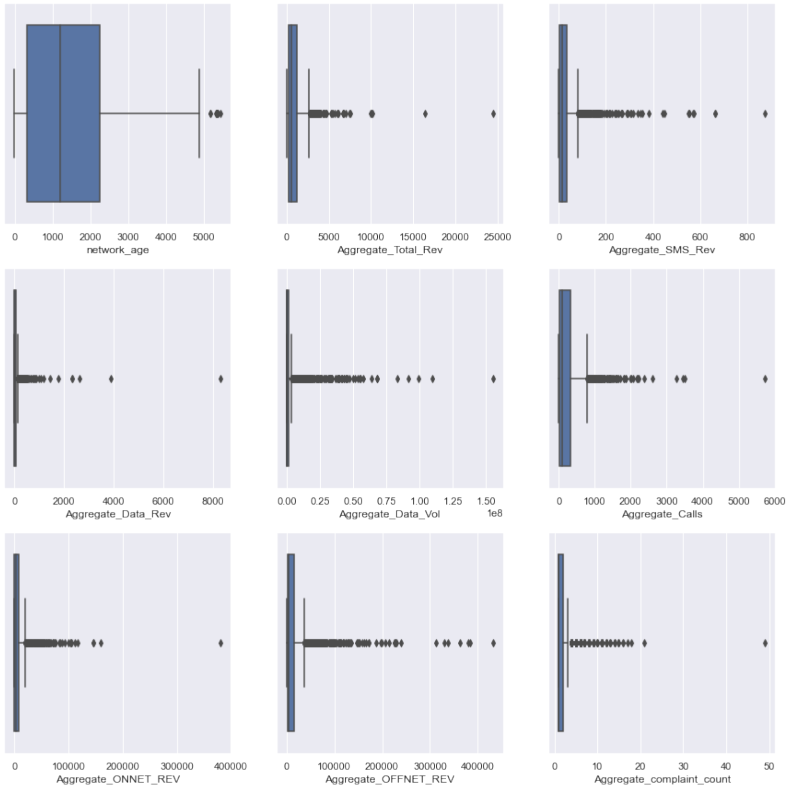

Figure 7: Box plots

2.2.4 Count plots for categorical variables

1. The target variable 'class' is perfectly distributed.
2. PTCL service provider customers dropped heavily in the month of September.
3. Ufone service provider saw a great increase in the month of September.

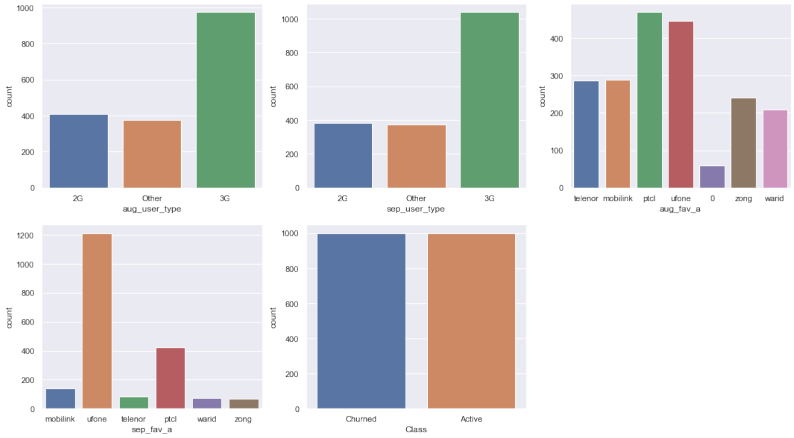

Figure 8: Count plots

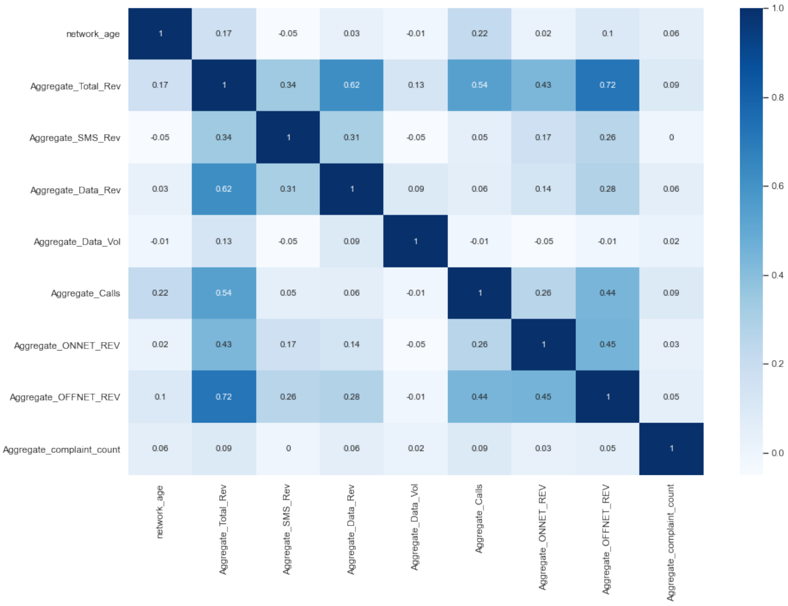

Figure 9: Heatmaps

2.2.5 Correlation or Heat maps

1. Feature Aggregate offnet Rev. and Aggregate Total rev has a good relation.
2. Other features have a minimal correlation.
3. Hence we can conclude that there will be a minimal effect of multi co-linearity.

2.2.6 Pre processing steps

1. There are missing values in all categorical features except target variable.
2. We will treat missing values with most repeated i.e. by mode value.
3. We will perform one hot encoding for categorical features.
4. Since Random Forest classifier is not sensitive to outliers and feature scaling, we will train the model using one hot encoded data directly.
5. When performing clustering we will perform feature scaling as well as treating the outliers as clustering is a distance based algorithm.

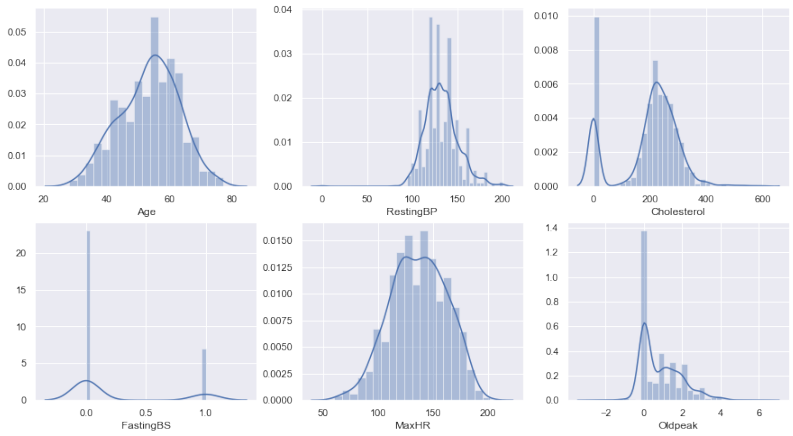

Figure 10: Histogram or Distribution plots

2.3 Heart Disease Prediction data

2.3.1 Basic Exploratory Data Analysis (EDA)

1. The data was downloaded from kaggle[3].

2. There are 918 rows and 12 columns/features in the data.

3. There are 5 categorical and 7 numerical features in the data.

4. Data is almost balanced in target variable (Heart Disease) with ratio 55:45 for Yes and No respectively.

5. There are no missing values in the data.

6. There are no duplicate records.

2.3.2 Distribution of numerical variables

1. Feature Age and MaxHr feature appears to be normally distributed.

2. Cholesterol feature appears to be having a anomaly, because cholesterol can never be so low like zero. Rest all readings in this feature finds normal.

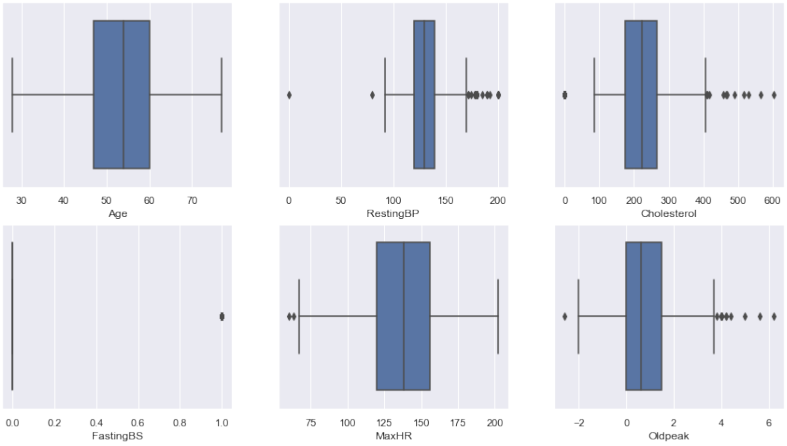

Figure 11: Box plots

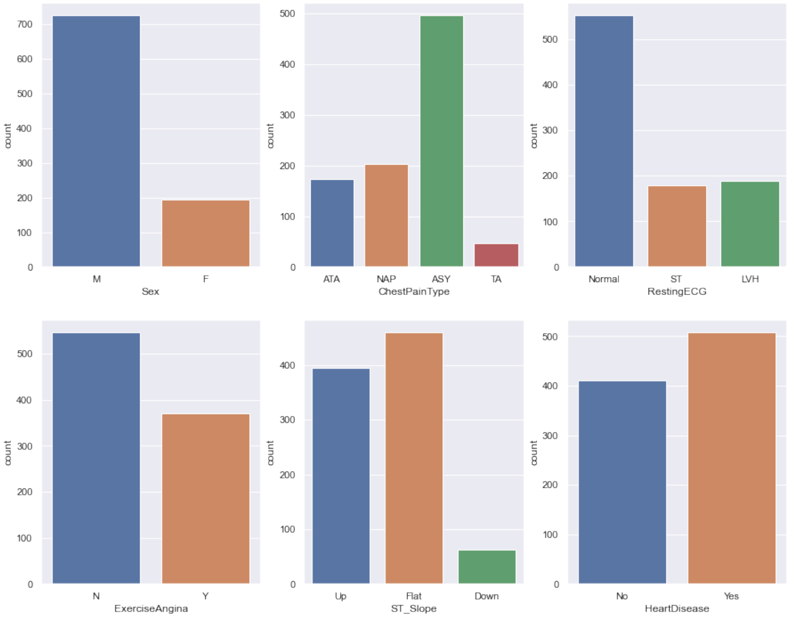

Figure 12: Count plots

2.3.3 Box plots for numerical variables

1. Except age feature, all other features has outliers.

2.3.4 Count plots for categorical variables

1. There are more male candidates than female in the data.
2. Candidates with ASY chest pain type are more.

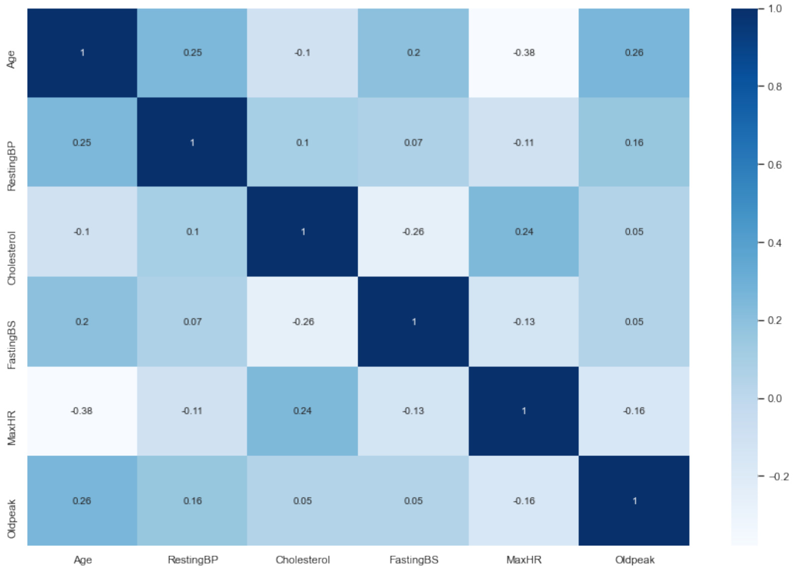

Figure 13: Heatmaps

2.3.5 Correlation or Heat maps

1. There is a negative correlation between Age and MaxHR feature.

2. Age and oldpeak has a positive correlation.

3. Rest all features have a minimal correlation.

2.3.6 Pre processing steps

1. There are no missing values or duplicate records in the data.

2. We will perform one hot encoding for categorical features.

3. Since Random Forest classifier is not sensitive to outliers and feature scaling, we will train the model using one hot encoded data directly.

4. When performing clustering we will perform feature scaling as well as treating the outliers as clustering is a distance based algorithm.

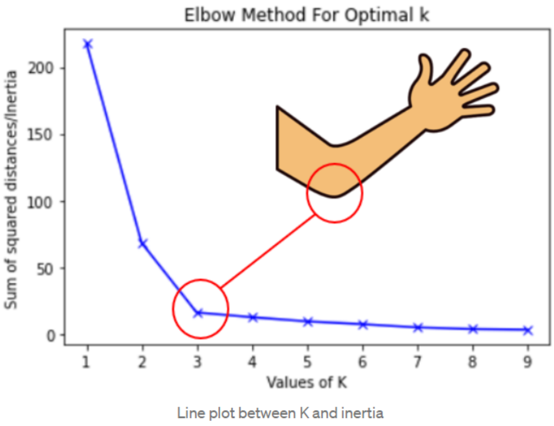

Figure 14: Elbow method

3 Methodology

Now as our data is cleaned and encoded, it is ready to be fetched for Machine Learning Algorithm. We will be doing following steps to train the model

3.1 Separation and splitting of data

Initially we will separate features and target variable. And then we will perform train and test split with the ratio 70:30 respectively. The split is done with stratification so as the the balance or target class remains the same.

3.2 Baseline Model

This model will be our reference or baseline model, any further models developed will be compared with this model. Baseline model will be trained using 10 fold cross validation with Random Forest Classifier. So out of 10 folds, 9 will be for training and 1 for testing. This will stabilize our results since it gives the overall average performance. The model will be trained for all the 3 datasets and their surrogates.

3.3 Clustering or Segmentation

Clustering is a unsupervised learning technique in which data points are segmented as per the groups or clusters we provide. We will be using k-means clustering to segment our data. Since k-means is a distance based algorithm, we will standardize the data before fetching to clustering algorithm. To know the optimal value of k, we will be plotting WSS (Within Sum of Squared) or inertia for each value of k from 1-10. We will decide k-value by looking at the elbow[4] point as shown in Figure 14.We will also note the centroid value and number of samples in minority class in the cluster.

To validate the clustering process, we will be using silhouette method. It ranges from -1 to +1. Positive sign indicates that data is correctly classified, negative sign indicates that it is wrongly classified and zero indicates that it's in the boundary line.

3.4 Blend of supervised and unsupervised learning

After providing cluster labels to all the datasets, we will again train a Random Forest classifier. Here, unsupervised learning technique i.e. k-means clustering becomes the pre-processing step before applying it to Random Forest classifier model(Supervised Learning). Hence the name blend or mixture of both techniques. We will only train when there are samples of both classes.

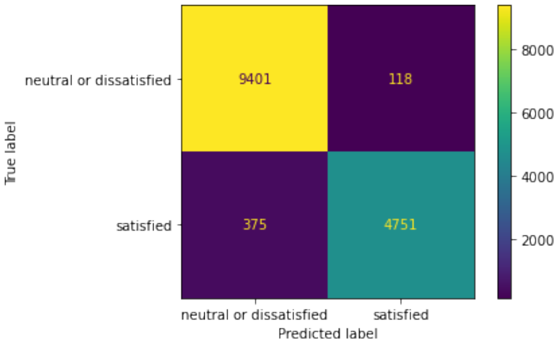

Figure 15: Confusion Matrix

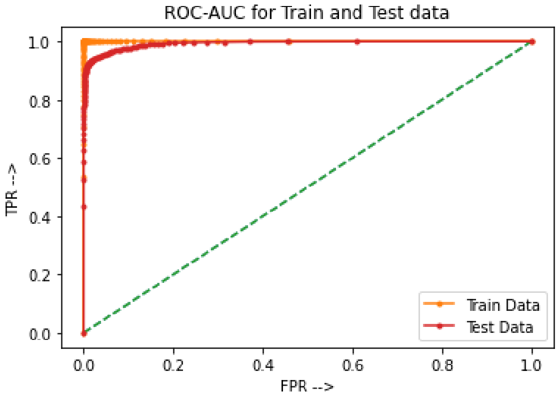

Figure 16: ROC-AUC Curve

3.5 Model Evaluation

3.5.1 Confusion Matrix

We will be using this type of matrix for evaluating the model(Refer Figure 15). We can calculate accuracy, precision, F1-Score and other model evaluation parameters. We can also find all above features in classification report also.

3.5.2 ROC-AUC Curve

1. AUC - Area Under the Curve

2. ROC - Receiver Operating Characteristics

3. This graph can visually give a good idea about the model performance. It is plotted with True Positive Rate(TPR) and False Positive Rate(FPR) on y and x-axis respectively. Refer Figure 16.

4. If the gap between the train and test data is minimum, the model performance is better.

4 Conclusions

Blend of supervised and unsupervised learning might surely pave the way to solve the problem of imbalances in the data. It will take many iterations to come up with optimal solution. Performance can be tuned by pruning the Random Forest Classifier. We can also explore other ensemble techniques like bagging or boosting to increase the predicting power of the model.

For any machine learning algorithm, they have its own strengths and weaknesses. For Random Forest, we will be sure of stabilizing the model by training n number of decision trees with different set of samples and features. We can get the feature importance value for each feature; it will help in the process of feature engineering. The most important advantage is that RF model can handle unscaled data and is not sensitive to outliers. Since RF model uses n number of decision tress, it will take more processing time to train a model. Sometimes model tends to overfit.

Similarly for k-means clustering, its advantage is that it is simple distance based clustering approach and is more efficient. For clustering algorithm, data needs to be standardised. Hence feature scaling is very important. And as it is a distance based approach, it is very sensitive to outliers. We need to treat outliers and scale the data before performing k-means clustering.

References

1. https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction
2. https://www.kaggle.com/mahreen/sato2015
3. https://www.kaggle.com/fedesoriano/heart-failure-prediction
4. https://www.analyticsvidhya.com/blog/2021/05/k-mean-getting-the-optimal-number-of-clusters/