JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 7, ISSUE 08, 2020

A PREDICTION MODEL FOR IMBALANCED DATASETS USING MACHINE LEARNING

Owk Mrudula[1], A. Mary Sowjanya[2]

[1]Research Scholar, Department of Computer Science and Systems Engineering, Andhra University College of Engineering(A), Visakhapatnam, AP, India

[2]Assistant Professor, Department of Computer Science and Systems Engineering, Andhra University College of Engineering(A), Visakhapatnam, AP, India

Email id: [1]mrudulaowk1990@gmail.com, [2]sowmaa@yahoo.com

Received: 11.03.2020 Revised: 12.04.2020 Accepted: 28.05.2020

ABSTRACT: The analysis of health datasets, becomes more challenging when the distribution of classes among the data is imbalanced. Hence medical data needs to be first balanced before proceeding to prediction analysis. A wide range of machine learning algorithms are available for classification. As such after performing data preprocessing, we focused on checking whether the dataset is balanced or not. If the dataset is balanced, we continued with prediction analysis. For imbalanced datasets we have used sampling techniques like over sampling, under sampling and hybrid sampling. After the dataset is balanced machine learning algorithms like Logistic Regression and Decision Tree, and ensemble classification algorithms like Boosting and Random Forest have been applied. Our results indicate that the performance of the prediction model increases after balancing the dataset. It can be seen that the ensemble classifier Random Forest gives better accuracy than the remaining classifiers for the Framingham dataset.

KEYWORDS: Class imbalance, Sampling Techniques, Machine Learning, Ensemble Classification, Random Forest, Prediction.

I. INTRODUCTION

Machine learning has become widely used term encompasses different types of programs in data analytics. Generally, machine learning algorithms can be categorized into main categories into Supervised, Unsupervised and Reinforcement. In supervised Learning the model maps the inputs to the corresponding outputs whereas Unsupervised learning algorithms cluster the dataset into different groups. Reinforcement learning algorithms are trained on taking decisions and by experience give predictions.

Traditionally algorithms assume that the classes are more or less equally distributed but the problem arises when the majority class(negative) is more than the minority class(positive). In the context of medical datasets negative class implies absence of disease and positive class implies presence of disease. Hence imbalanced medical datasets pose a severe problem for disease prediction. As such sampling techniques can be used to balance the dataset before prediction analysis.

II. RELATED WORKS

Haixiang, Guo etal [1] have collected and reviewed technical, practical papers to learn about class imbalanced data. Their prediction task was to detect rare events and develop a taxonomy of existing applications. A cost sensitive extension of least mean square algorithm was proposed by Belarouci, Sara etal [2]. They compared the obtained results before and after balancing. Fernández, A etal have presented imbalanced classification in big data and analyzed the behavior of standard preprocessing techniques. They have also discussed the challenges and future directions [3]. Different oversampling techniques were used by Huda etal [4] to build an ensemble classifier to identify faulty components in safety critical software. A complete survey on high class imbalance in big data was given by Leevy, J.L., etal [5]. M. A. U. H. Tahir etal developed a classification model for class imbalanced dataset using genetic programming [6]. Hasanin, T etal [7] discussed the challenges posed by imbalanced data while investigating different data sampling approaches like random over sampling, random under sampling, SMOTE etc. An improved oversampling algorithm for classifying imbalanced data was implemented by Xie etal [8].

ISSN- 2394-5125 VOL 7, ISSUE 08, 2020

III. METHODOLOGY

Algorithm: Class Imbalance _Prediction

Input: Imbalanced Dataset (DS_i) with $r_1, r_2, r_3, \ldots, r_n$ are records and $X_1, X_2, X_3, \ldots, X_m$ attributes
Output: Balanced Dataset (BD) with disease =Yes or disease =No

Step-1: Perform Data preprocessing
Step-2: Check for class imbalance
Step-2.1 If the ratio of distribution of classes is more or less equal, then dataset is balanced. Goto Step-4
Step-2.2 If the ratio of distribution of classes is unequal, then dataset is imbalanced. Goto Step-3
Step-3: Use sampling Techniques for balancing DS_i
Step-4: Perform classification on new_DS_i
Step-5: Generate n random numbers $k_0, k_1, k_2, \ldots, k_n$, between 0 and 1, where $n = |D_j| - |D_o|$ such that symmetry is created.
Step-5.1 For each random number k_j
Step-5.1.1 Generate a new record s
Step-5.1.2 Append new record to D_o
Step-7: For each record r_i in D_o, where $D_o \in$ new_DS_i, predict disease =Yes or disease =No with the algorithm selected in Step-5

Fig.3.1 Prediction model for Imbalanced Dataset

3.1 Data preprocessing

Since medical datasets are collected in different environments proper preprocessing of the data has to be done before processing to analysis. Missing values, errors, outliner analysis, dimensionality reduction and feature extraction etc. are handled here.

3.2 Class Imbalanced data

Here the distribution of positive and negative classes from the dataset is identified. If the ratio of distribution of majority and minority classes is more or less equal then it is a balanced dataset. We can proceed directly to classification. But if the ratio of distribution of majority and minority classes is more or less is unequal (60:40, 70:30 and so on) then it is an unbalanced dataset. More the difference in the ratio more the unbalance of the classes.

3.3 Sampling Techniques

To re-balance the above imbalanced dataset sampling techniques have been used. Sampling techniques include under-sampling, over-sampling, hybrid sampling and advance sampling. In this work we have used random under sampling (RUS), random over sampling (ROS) and a combination of over sampling and under sampling.

3.4 Supervised Learning Classifiers

Many machine learning algorithms like naive Bayes, SVM, KNN, RF, Regression are available for classification purposes. We have already used Naive Bayes classification previously for handling class imbalance [9]. Hence, we have chosen regression techniques – Linear Regression and Logistic Regression, Decision Tree and SVM, Ensembled techniques- Boosting and Random Forest.

ISSN- 2394-5125 VOL 7, ISSUE 08, 2020

IV. RESULTS AND DISCUSSION

4.1 Dataset

The Framingham dataset has been taken from Kaggle [10]. It comprises of 16 variables and 4240 observations. Out of which 3596 belong to majority class(negative) and 644 belong to minority class(positive).

4.2 Evaluation metrics

Assessing the classifier performance is important for any prediction task. They are many evaluation metrics for this purpose. Evaluation metrics like Accuracy, Precision, Recall, F-measure etc. The accuracy metric is not a proper choice for imbalanced datasets as it does not take into consideration the minority class. In case of medical datasets this may lead to improper conclusion and poor performance. As such class independent measures are preferred.

4.3 Results

The figures 1, 2 show the imbalance ratio (85:15) and the importance of variables in the Framingham dataset.

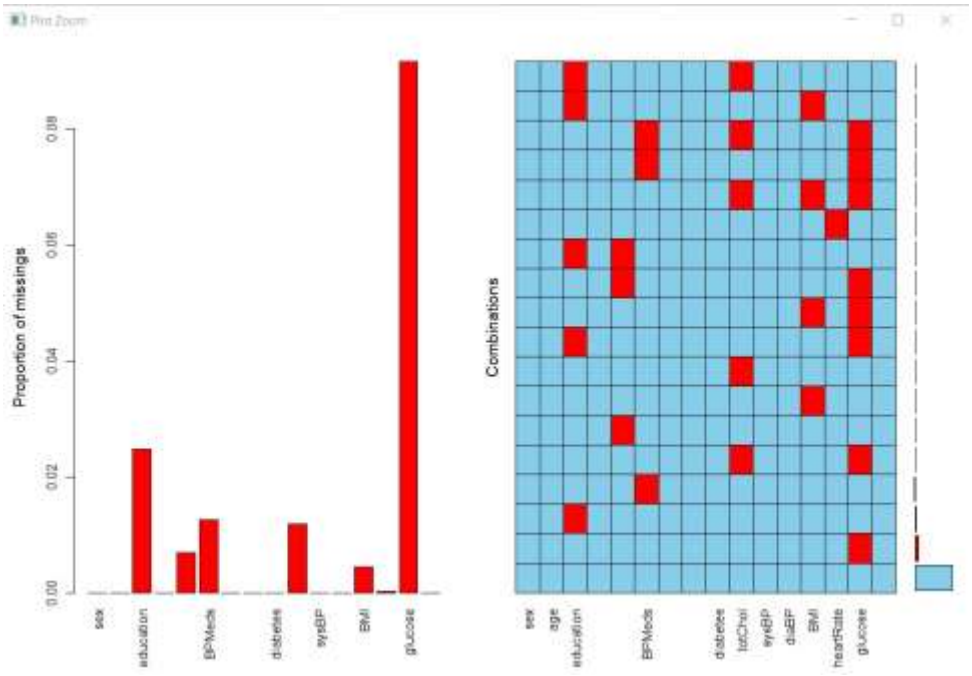

Fig.4.1 Missing values of the Framingham dataset

ISSN- 2394-5125 VOL 7, ISSUE 08, 2020

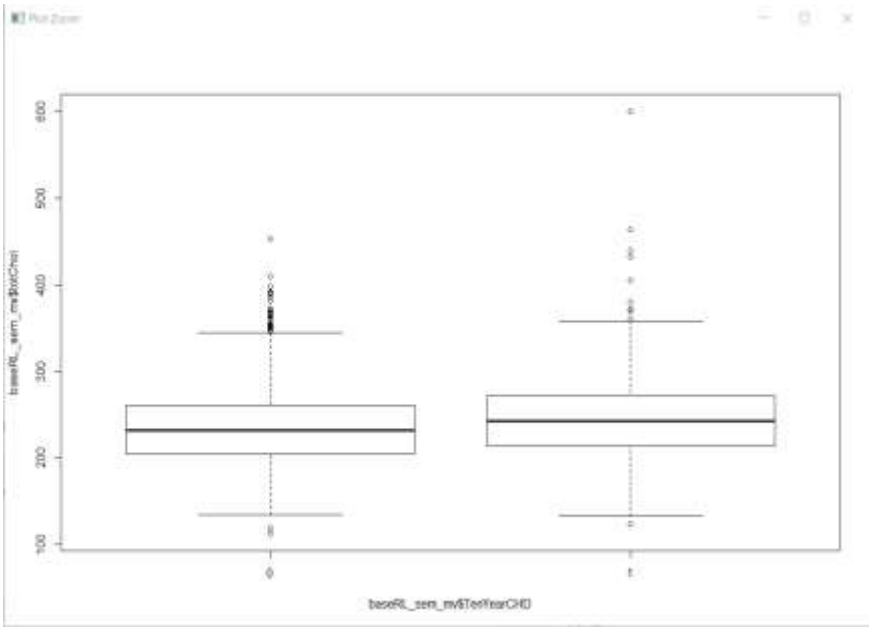

Fig.4.2 Identifying Outliers by using Boxplot

```
> table(new.framingham$TenYearCHD)

   0    1
3596  644
>
```

Fig.4.3 Ratio of the Framingham dataset

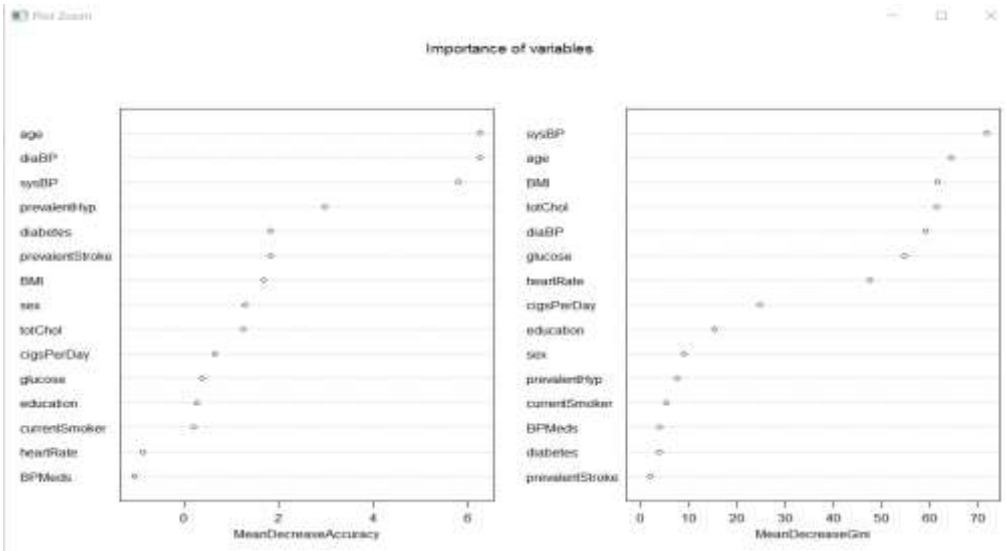

Fig.4.4 Variable Importance

ISSN- 2394-5125 VOL 7, ISSUE 08, 2020

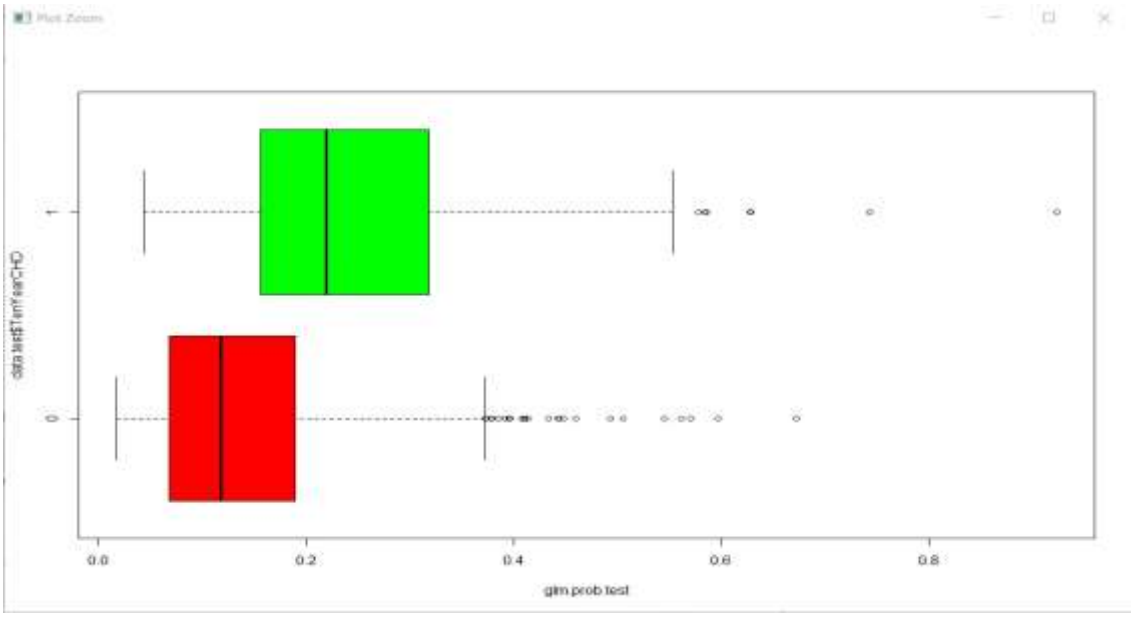

Fig.4.5 Boxplot for Logistic Regression on test samples

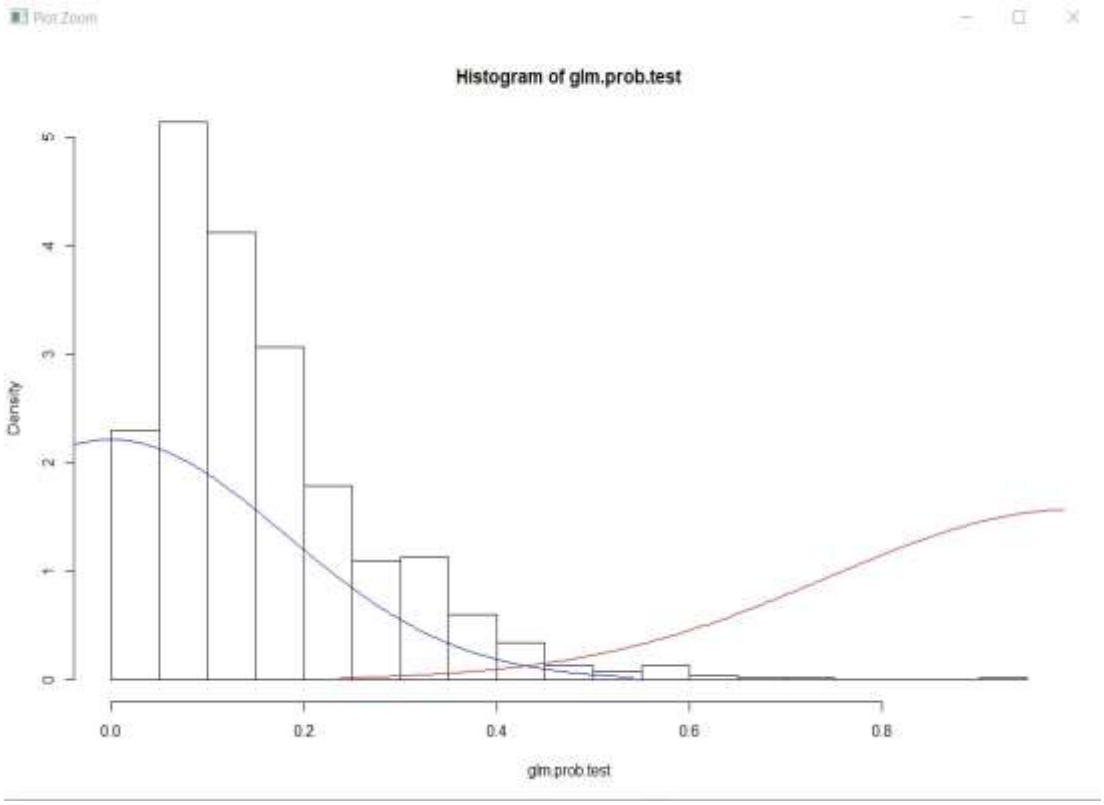

Fig.4.6 Histogram and Density plot are shown on Logistic Regression

ISSN- 2394-5125 VOL 7, ISSUE 08, 2020

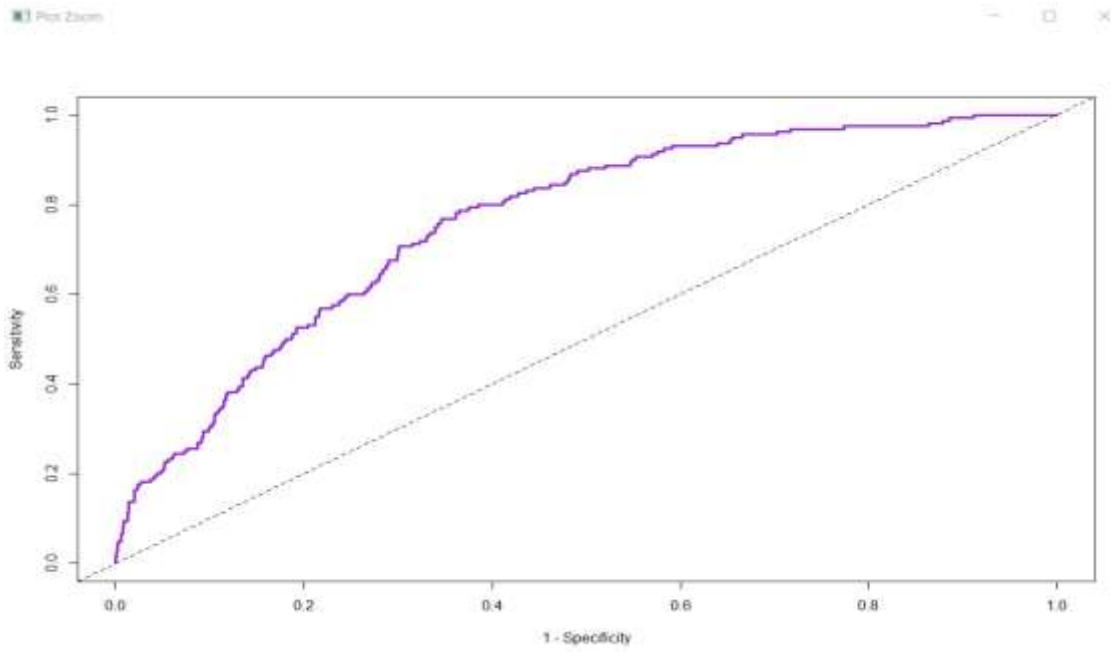

Fig.4.7 Plotting the ROC curve for the Logistic regression between Sensitivity Vs Specificity

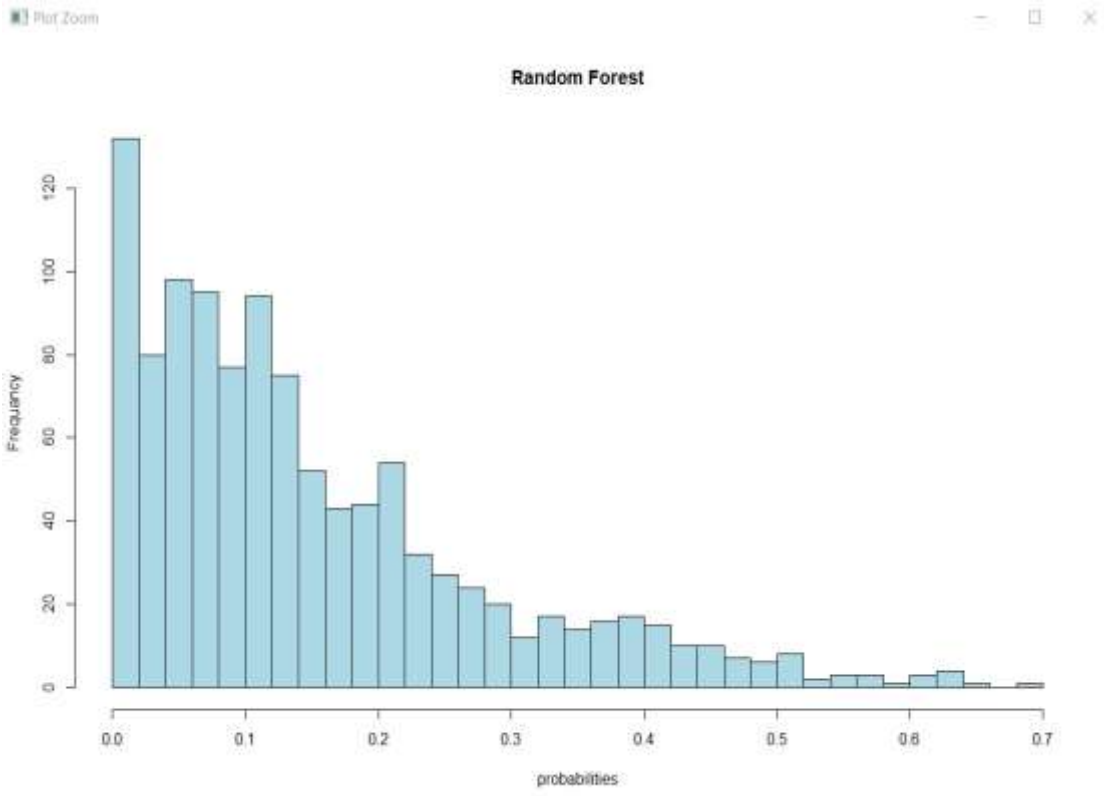

Fig.4.8 Histogram for Random Forest probabilites with the frequency

ISSN- 2394-5125 VOL 7, ISSUE 08, 2020

```
> boost.test$metrics
               H      Gini       AUC      AUCH        KS       MER       MWL Spec.Sens95
scores 0.07927721 0.2784152 0.6392076 0.6554829 0.2389608 0.1449407 0.1896197  0.07684098
       Sens.Spec95        ER   Sens      Spec Precision Recall    TPR        FPR         F
scores     0.09375 0.1969006 0.1625 0.9124867 0.2407407 0.1625 0.1625 0.08751334 0.1940299
          Youden TP FP  TN  FN
scores 0.07498666 26 82 855 134
>
```

Fig.4.9 Performance Metrics for boosting

```
> glm.train$metrics
               H      Gini       AUC      AUCH        KS       MER       MWL Spec.Sens95
scores 0.1838808 0.4514892 0.7257446 0.7341667 0.3592086 0.1472081 0.1678708   0.1977819
       Sens.Spec95       ER       Sens      Spec Precision     Recall        TPR         FPR
scores   0.2115869 0.14877 0.06801008 0.9949168 0.7105263 0.06801008 0.06801008 0.005083179
               F    Youden TP FP   TN  FN
scores 0.1241379 0.0629269 27 11 2153 370
```

Fig.4.10 Performance Metrics for Logistic Regression

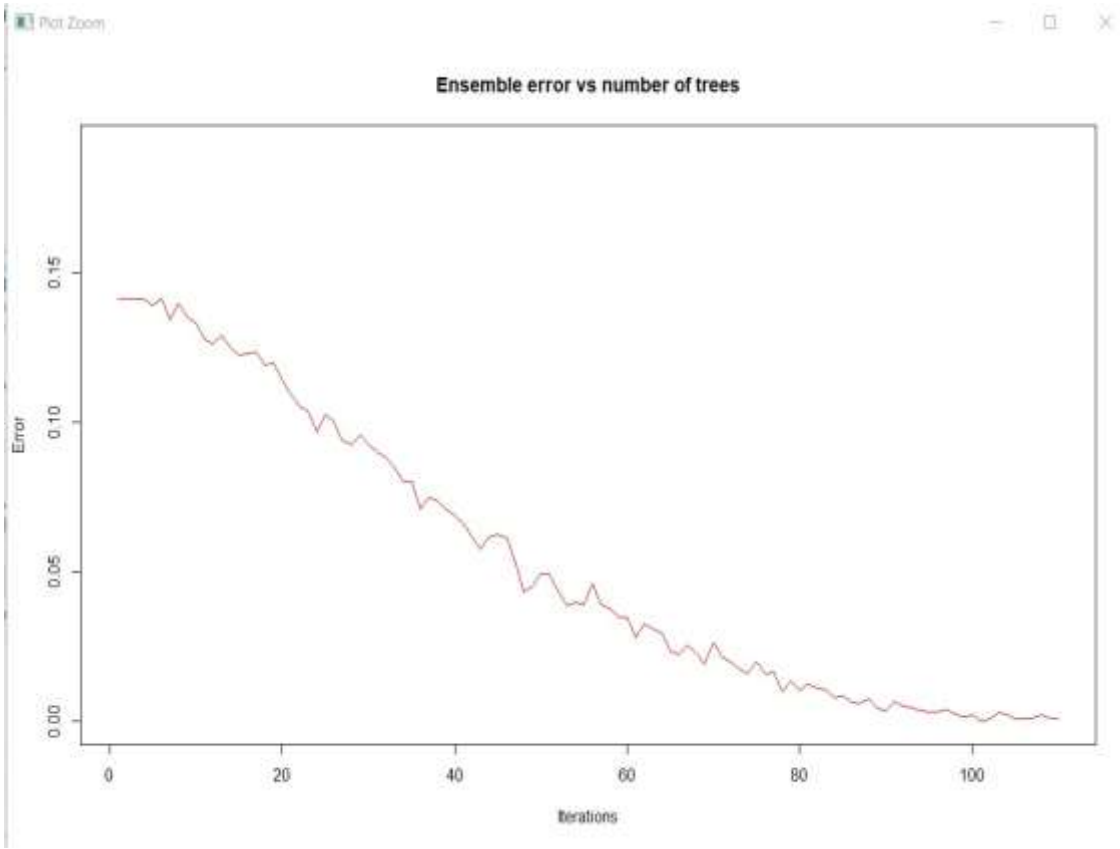

Fig.4.11 Error measurements according to the number of trees generated

ISSN- 2394-5125 VOL 7, ISSUE 08, 2020

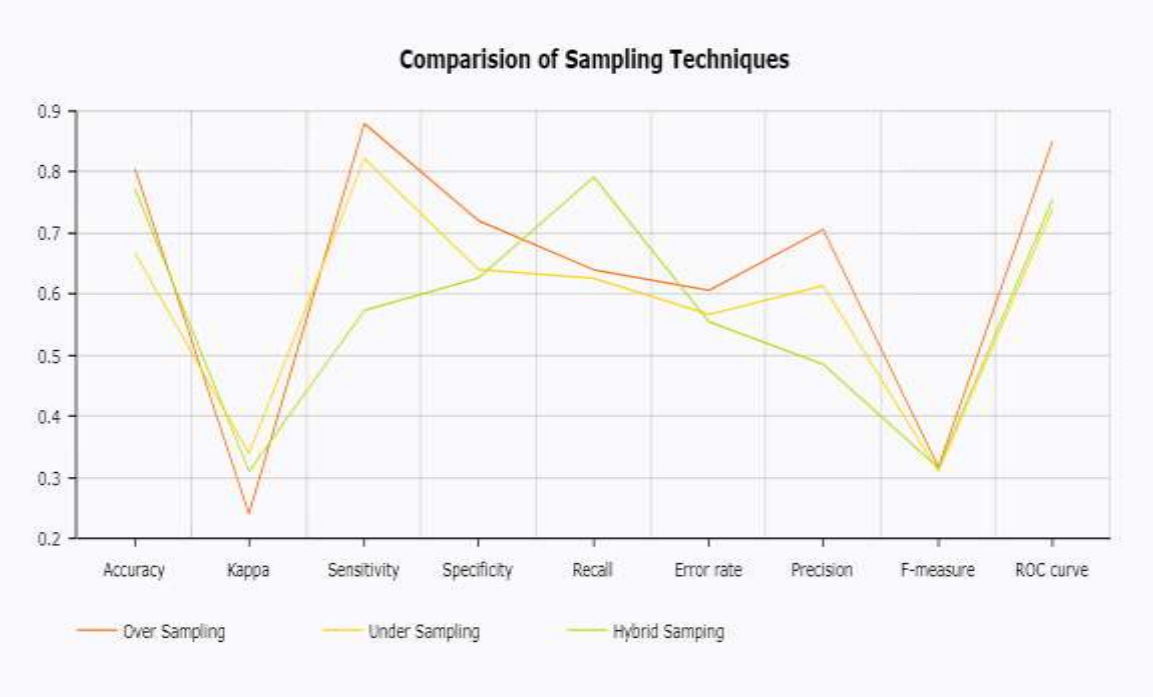

Fig.4.12 Comparison of Sampling Techniques for Framingham dataset

Measures	LR	DT	BOOSTING	RF
Accuracy	0.7257	0.8541	0.6392	0.8953
Kappa	0.3592	0.3682	0.2389	0.3732
Sensitivity	0.6801	0.5732	0.0937	0.6875
Specificity	0.9849	0.6285	0.0768	0.9839
Recall	0.6801	0.626	0.1625	0.0687
Error Rate	0.14877	0.0732	0.1969	0.1494
Precision	0.7105	0.640	0.2407	0.4230
Fmeasure	0.1241	0.333	0.1940	0.0687
ROC curve	0.808	0.817	0.782	0.815

Table I: Performance metrics for LR, DT, Boosting &RF

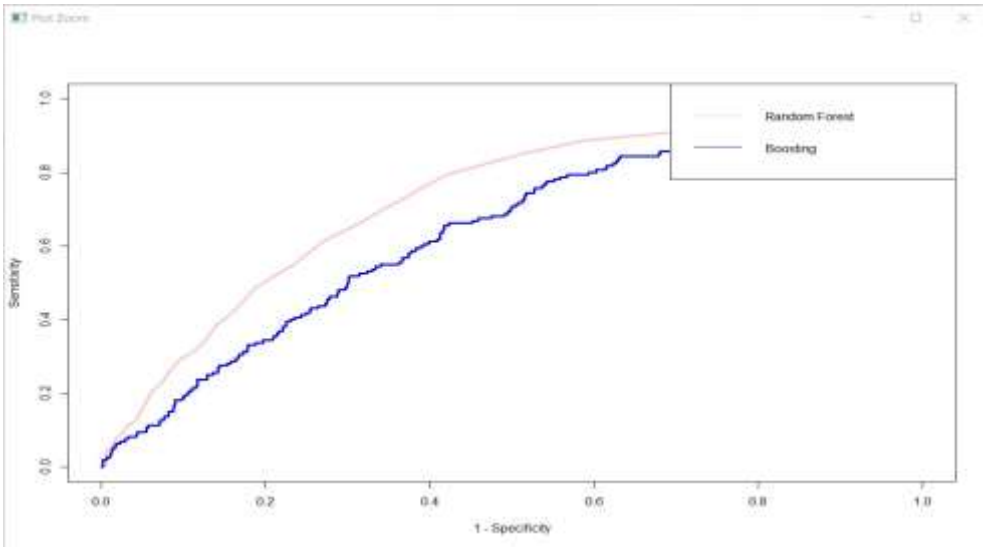

Fig.4.13 Plotting the ROC curve for both Ensemble techniques

V. CONCLUSIONS

We have taken a class imbalanced dataset like Framingham dataset. After performing data preprocessing, we have implemented over sampling, under sampling, and hybrid sampling techniques to re-balance the dataset. Then we proceed with classification using LR, DT, BOOSTING, RF. We can conclude that performance of a prediction model increases after balancing the dataset. Over sampling technique has yield good results among the sampling techniques. Random Forest ensemble classifier has proved to be a better classifier when compared to other algorithms. This work can be extended to focus mainly on ensemble classification algorithms in future.

VI. REFERENCES

[1] Haixiang, Guo & Li, Yijing & Shang, Jennifer & Mingyun, Gu & Yuanyue, Huang & Gong, Bing. (2016). Learning from class-imbalanced data: Review of methods and applications. Expert Systems with Applications. 73. 10.1016/j.eswa.2016.12.035.

[2] Belarouci, Sara & Chikh, Mohammed. (2017). Medical imbalanced data classification. Advances in Science, Technology and Engineering Systems Journal. 2. 116-124. 10.25046/aj020316

[3] Fernández, A., del Río, S., Chawla, N.V. *etal.* An insight into imbalanced Big Data classification: outcomes and challenges. *Complex Intell. Syst.* **3,** 105–120 (2017).

[4] [4] Huda, Shamsul & Liu, Kevin & Abdelrazek, Mohamed & Ibrahim, Amani & Alyahya, Sultan & Al-Dossari, Hmood & Ahmad, Shafiq. (2018). An ensemble oversampling model for class imbalance problem in software defect prediction. IEEE Access. PP. 1-1. 10.1109/ACCESS.2018.2817572.

[5] Leevy, J.L., Khoshgoftaar, T.M., Bauder, R.A. *et al.* A survey on addressing high-class imbalance in big data. *J Big Data* **5,** 42 (2018)

[6] M. A. U. H. Tahir, S. Asghar, A. Manzoor and M. A. Noor, "A Classification Model for Class Imbalance Dataset Using Genetic Programming," in IEEE Access, vol. 7, pp. 71013-71037, 2019, doi: 10.1109/ACCESS.2019.2915611.

[7] Hasanin, T., Khoshgoftaar, T.M., Leevy, J.L. *et al.* Severely imbalanced Big Data challenges: investigating data sampling approaches. *J Big Data* **6,** 107 (2019).

[8] Xie, Wenhao & Liang, Gongqian & Dong, Zhonghui & Tan, Baoyu & Zhang, Baosheng. (2019). An Improved Oversampling Algorithm Based on the Samples' Selection Strategy for Classifying Imbalanced Data. Mathematical Problems in Engineering. 2019.

[9] O.Mrudula, A.Mary Sowjanya, Handling Class Imbalance Using Sampling Techniques in R. proceedings second international conference SMART DSC- 2018. PP.254-26

[10] https://www.kaggle.com/amanajmera1/framingham-heart-study-dataset