A Random Forest Model To Predict Credit Balance Claim Recovery Accounts In Healthcare Sector

Pooja Mittal, Dr. Vinod Kr. Srivastava

Abstract: Credit Balance services play an important role to resolve providers and health care credit balance accounts to recover the overpaid payments and avoid the future errors. This is a critical problem for the health care systems because the accuracy of prediction is very low for identifying the potential overpaid claims where recovery can happen, therefore auditors have to go through all 100% claims, which results in spending lot of time, which cannot yield recovery. To address this problem, this case study will give the ability to health care systems to classify the potential revenue generated claims. We have described and proposed a Random Forest algorithm, which is applied on high dimensional and highly biased data. The proposed Random Forest algorithm has reduced the dimensions by selected the important features and taken care of highly imbalanced data. To classify the overpaid claims, RF has provided a significant improvement over other algorithms such as Decision Tree and Logistic Regression. We have identified the high-ranking features, which influence the credit balance accounts, or claims, which reduce the high dimensionality and enhance performance. This proposed solution offers a new way to help human auditors to focus on revenue generating accounts with high yield and will prevent future errors.

Keywords: Random Forest, Recall, Precision, Confusion Matrix, F-Score

1. INTRODUCTION

In today's world, healthcare industries are facing a major problem of credit balances recovery. This is the problem, which occurred for some areas when the insurance company or service provider use to pay the claim on the prior request basis but after that it came to notice of the company that they have made a wrong payment. Now the retrospective phase or recovery phase arrives in which the payer tries to recover the amount wrongly paid. This is called the credit balance Recovery. The company has to spend a lot of time on one record while dealing with the hospital to check weather a wrong payment has been done or not. As per the study, sometimes it is around three to four hours per record, which is very time consuming and costly process. Apart from that, the chances of errors are very high. The major goal of any firm or company is to minimize the cost and maximize the revenues. Credit balance Services resolve hospital and health system credit balance accounts to recover overpayments and prevent future errors. The queries used to drive workflow were developed in 2006, with very few changes done post that. The accuracy of prediction is ~20%, yet auditors have to work on all accounts that already predicted. As a result, they have to spent lot of time on suspected accounts that cannot yield recovery. The purpose of this study is to

identify those client accounts, which will provide refund to payer from large inventory of the claimed transaction data. This helps auditors to prioritize them for recovery and save on administrative cost. Now auditors can focus their efforts on revenue generating accounts. This will allow our auditors to be more effective in an overall credit balance resolution environment and thus ensuring meeting overall payer objectives. Credit Balance services resolve hospital and health system credit balance accounts to recover overpayments and prevent future errors. Credit Balance resolution and recovery services provide on-site and remote location resources to help providers research and resolve unsolicited and solicited overpaid claims. There are various algorithms to predict credit Balance claims on prior basis but in this study we have used random forest algorithm. Decision trees are the building block of random forest Model. Random forest, as its name depicts, consists of large number of decision trees, which operates as ensemble. Every decision tree gives its own decision and the class, whichever class gets the maximum number of voting, becomes the Model's results. All the trees are uncorrelated and that is the key of this model performs exceptionally well. Uncorrelated models produce ensemble results and those are more accurate than any individual model. Decision Trees (DTs) predicts really well but in case of higher depth, it is prone to over fitting and is explored by random forest Model. Random forest model works on randomly sampled with replacement strategy which is knows as bootstrap samples. These samples are further fed into many DTs with large heights. Each Decision Trees are separately trained on bootstrap samples. Aggregation of decision trees is called random forest Ensemble and the majority of the votes from Decision Tree determine the conclusion.

Two prerequisites of the random forest Model:

- Low correlation between the trees is the key, which produces ensemble prediction that is more accurate than the individual one.
- Actual signal in the features builds better model than random guessing.

Feature Randomness is one of the powerful features where it picks only from the random subset of features in addition to the random samples. This ultimately results in lower correlation across trees and therefore less prone to over-fitting.

It is very important to validate the random forest Model and there are different ways. Validating the model with whole

- *Pooja Mittal is currently pursuing Ph. D. degree program in Computer Science in Baba Mastnath University, India, E-mail: mistletoe241984@gmail.com*
- *Dr. V.K. Srivastava is currently working as Assistant Professor in deptt.of Computer Science in Baba Mastnath University, India, E-mail: srivastava_v_k@yahoo.com*

validation dataset is the most common validation approach. Random forest comes with the “Out of bag” scoring mechanism, which is different from the validation score. E.g. there are 5 decision trees and suppose the training dataset is n. For first Decision Tree out of n observations there are x amount of observations are left out which is called Out of Bag sample and the x rows will not be included for training in first DT. Once all the DTs are built then the x observations will be considered as the unseen data testing. X rows will be passed to all bootstrapped DTs models which do not have these observations. Final prediction is computed based on correct predictions from the out of bag sample. Random forest has outstanding ways of handling Missing values with Replacement. First way is faster where say nth variable is non-categorical and it computes the median of all the values in class I and uses this value to replace all the missing values. In case of categorical variable, replacement is done by the most frequent non missing values. Another way is computationally expensive but usually gives better results. It start with filling the missing values by adhoc and inaccurate values and then run the forest further in order to fill the optimized values. In order to solve the credit balance recovery problem, data is collected through business analyst team, it was decided to do the Exploratory Data Analysis on 2017 data for one of provider and come up with working model to identify the revenue generating accounts or claims.

2. LITERATURE REVIEW

RF classifier is the collection of tree-structured classifiers. “The advanced version of bagging is random forest according to Breiman.”[1]. Pal[2] state in his study of “Remote Sensing Classification” that random forest algorithm is well suited when compared with support vector machine. It requires less number of user defined parameters. It is also helpful in handling categorical data with missing values. Apart from that, it is also beneficial to use random forest with highly biased data. Cafri et al.[3]explain that random forest is a machine-learning model developing approach, which is based on decision trees forest. While predicting the risk, it can be found better due to its nonparametric nature. He defines how a group of decision trees can predict more accurately as compared to a single decision tree. Talab et al.[4]define that random forest algorithm have many attractive features by doing the study on classification problem. It is a non-parametric and nonlinear algorithm in handling big data sets. It contains numerical data as well as categorical data. It is also in better position where observations are less but predictors are more. It includes the interactivity among various predictors. Ok et al.[5]explain how random forest method can be more promising along with parcel based post classification policy. However, when different parametric combination or test conditions are applied to the random forest algorithm, again it will provide the results that demonstrate the feature of consistency of random forest. It can be a reliable way for predictions. Evan et al. [6] suggest that during preliminary test, random forest produce more accurate output. However, random forest is less efficient for nonlinear regression. It has been widely used in different ranges of fields like Mao et al. [7]used random forest in “Cultural Modeling for Behavior Analysis and Prediction”. Carvajal et al.[8]have used random forest in “Intelligent Digital Oil and Gas Fields”. random forest was used in “Development of a model to identify combined use in residential water end use events” in International symposium on Process System Engineering by Eduardo S. Soares et al.[9]. Conjeti et al.[10]built a “Domain Adapted Model in Vivo Intravascular Ultrasound Tissue Characterization” since oblique random forest has the ability to use oblique diseases splits inspire of the traditional axis align decisions boundaries at split nodes. Iadanza et al.[11]widely used random forest and decision support system in Healthcare and in clinical engineering. As per them, random forest train and provide information to:

1. “Cardiologist” in patient setting, to help managing the patients.
2. Random forest helps to access the severity of runtime state of the patient, which helps nurses who perform the home visit. Apart from that, Scholars have also optimized the random forest algorithm efficiency to improve the predictive analysis.

3. METHODOLOGY

3.1. Random Forest

One of the supervised machine learning algorithms, which is based on ensemble learning, is random forest. A number of decision trees are formed and then merged to create a “forest”. It is based on the assumption that rather than using a single decision tree, it is better to use a collection of DTs to take decision with higher accuracy. It is able to handle both regression and classification problem. If the target variable is continuous then random forest regression is used and for discrete values random forest classifier is used. There are many studies, which have proven that random forest has high accuracy with high tolerance. It is able to handle noise and highly biased data due to its two phase process. It is an example technique and follows the two steps process:

Step 1: It creates some samples from the original data using bootstrap sampling techniques.

Step 2: It runs the decision tree for each sample since it creates the decision tree for each sample. The majority voting of decision trees predictions which decide the final classification result.

Random forest training process and predictions is covered in 3 steps[12] :

1. Training site selection: It uses the bootstrap random sampling technique to build and training site from the actual data set where each training set has the same size as the actual data site. E.g. k<=K
2. Building the random forest model: It builds the N decision tree models for each of bootstrapped training data sets, which form the forest from these trees. It does not select the most contributing or the features with highest information gain rather it uses the random k number of features.
3. Voting: Training process of the N decision trees is independent which gives the random forest a capability to create the decision trees in parallel,

which saves lot of training time. Combining the decision trees forms a random forest model. When there is an unknown input samples is fed to predict the class then each trees votes for class individually. As a result, random forest chooses the class, which has the highest number of votes. Additionally, it has a great ability to process the bulk data sets as compared to other methods. Decision tress has been proven very successful in solving the classification problem of statistical learning.

Credit Balance problem is resolved and implemented using random forest algorithm. In the study of credit balance problem in healthcare, it has highly imbalanced dataset. This study uses the random forest algorithm CV (random forest cross validation) into two steps. First, to identify the features those are important to reduce the dimensionality. In the second step, important features are used to build random forest algorithm and the model performance is measured by using Confusion Matrix and classification report.

***Table 1** Confusion Matrix*

	Predicted		
Actual		Primary	Secondary
	Primary	TP	FP
	Secondary	FN	TN

3.2. Classification Evaluation Index

Confusion Matrix is the most common way for the performance measurement of machine learning classification problem. It is a four different combination of predicted and actual values. It is extremely useful to measure the Precision, Recall, Accuracy, F-score and most importantly the AUC curve.

In this study, our focus is whether the claim will be potentially recovered or not. TP, FP, TN and FN are crucial factors for the final algorithm performance evaluation.

TP denotes claim is positive predicted and it is true.
FP denotes claim is positive predicted and it is false.
TN denotes claim is negative predicted and it is true.
FN denotes claim is negative predicted and it is false.

Recall

It denotes the TPR (True Positive Rate) i.e. out of all positive classes how much we predicted correctly.

$$TPR = Recall = \frac{TP}{TP + FN} \quad (1)$$

Precision

It denotes, out of all the positive classes we have predicted correctly how many are actually positive.

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

F-Measurement

$$F - Measurement = \frac{2 * Recall * Precision}{Recall + Precision} \quad (3)$$

It is difficult to compare the models with low Precision and high recall and vice-versa. Therefore, F-Score is useful to compare models. The AUC denotes the area under curve and it is useful to find the model performance. Higher the value of AUC curve, better the model is. X co-ordinates of the ROC curve are FPR and Y coordinates is recall. We will focus only in credit balance problem, which has highly balanced classes with the smaller number of positive classes. The goal of this study is to primarily focus onto avoid misclassification of positive classes or the high precision for positive classes. Therefore, Precision, Recall, F-Score and AUC curve will be used for model performance evaluation.

3.3. Feature Importance

In most of the business cases, it is equally important to have the inter process model as the model accuracy. Therefore, we may have to compromise little bit of model accuracy for the sake of interoperability. For example, when a bank rejects a loan application, it is important to explain the reason, which can be presented to the customer. In credit Balance case study, the data is comprised of different categories: Payer related information, Provider related information, Customer Insurance information, payment information. By keeping all the information for algorithm building can be expensive in terms of time and space which can significantly impact the model performance. Additionally, it is not easy to focus on the useful features and understanding the model's logic and forest has the inbuilt capability to rank the features based on the importance. As a result, it is helpful to reduce the dimensionality by removing the less important features. The main idea is to reduce the performance by adding noise to the features. The feature importance calculation is identified as below:

1. For each decision tree, only a subset of decision trees is used for determining the OOB score. Mostly, the details that is not enough to train and test the model. Here, OOB provides a good trade off by keeping around 36.8 % of total trading data aside for validation. Data is never used in the process of training the decision trees. Additionally, OOB can be used to evaluate the decision tree performance and predictions the error rate is considered as OOB error. For example, OOB1err.

Adding the random noise in feature "A" for all OOB data and then calculate the OOB error again and keep it as OOB2 error.

2. Check Feature Importance: Considering the N number of decision trees in random forest, feature importance can be calculated by the following equation:

$$IMPORTANCE = \sum_{i=1}^{n} \frac{(OOB2err - OOB1err)}{N} \quad (4)$$

If accuracy of outer bag is reduced by adding a random noise to the feature, this indicates the feature having high correlation with the target variable or it has a high degree of importance.

In this case study, feature solution has been used due to the following benefits:

1. Model is computationally less expensive with fast performance.
2. Model can be easily explained to the business.

3. Shows the high correlation of independent features with target variable.
4. Model becomes more accurate with fewer numbers of variables with high predicting score.

The following steps are taken to identify the features with high importance:

1. Run the model with total number of m features.
2. Check the accuracy of the model.
3. Find the high performance index features by sorting the features with the weight index.
4. Select the top features explain the high variability ratio.
5. Build and run the model with selected features and compare the accuracy.
6. Follow this process until models give the required performance with the minimum number of important features.

3.4. F-Score of Random Forest Algorithm

Usually decision trees have equal weightage of random forest. In case of highly imbalanced data, this does not work well. Weighted F-Score measurement is used in order to generate the better predictive model by giving different methods to decision trees. In our case, “Payer refund” is primary category and “Non payer refund” is the secondary case. Model accuracy is not sufficient if only the accuracy is calculated. For example, positive cases in our study are 98% and secondary cases are 2% only. If all primary cases are falsely classified, however, accuracy is 98%. However, it is not good from the business point of view where all the potential claims will be put as lower priority and cases with non-revenue generating claims will be put on higher priority. In this research, both Precision and Recall are considered and F-Score measures are used to evaluate model performance. This way the model performance is improved. F-score combines the Precision and Recall and determines the model performance. If the F-Score is higher, model performance is better. Its formula is shown as below:

$$F - Measure = \frac{2 * recall * precision}{recall + precision} \quad (5)$$

$$F - Measure = \frac{2TP}{2TP + FP + FN} \quad (6)$$

In credit balance case study, algorithm building and testing follows:

Step 1: Divide the datasets in three parts- training set, testing set, and validation set. Out of N data, 70% samples are considered as training set and 30% samples are kept as validation set. Samples data are not drawn in training and validation are called testing set which are kept as unseen dataset for final model evaluation.

Step 2: Build the random forest model by giving 70% input training set and use the number Decision Tree estimators to build the classifier.

Step 3: Input the validation data to the classifier to measure the weight value by calculating F-Score. Each sample is classified by each decision tree in the forest as an independent classifier and based on majority voting the decision is made. This calculates the F-Score by getting TP, TN, FP and FN for the final classifier decison.

Step 4: Input the unclassified test set to evaluate the random forest model performance. Model performance is measured by F-Score Weight Index which is the final model performance for credit balance classification problem that classifies whether the claim will generate the potential recovery or not.

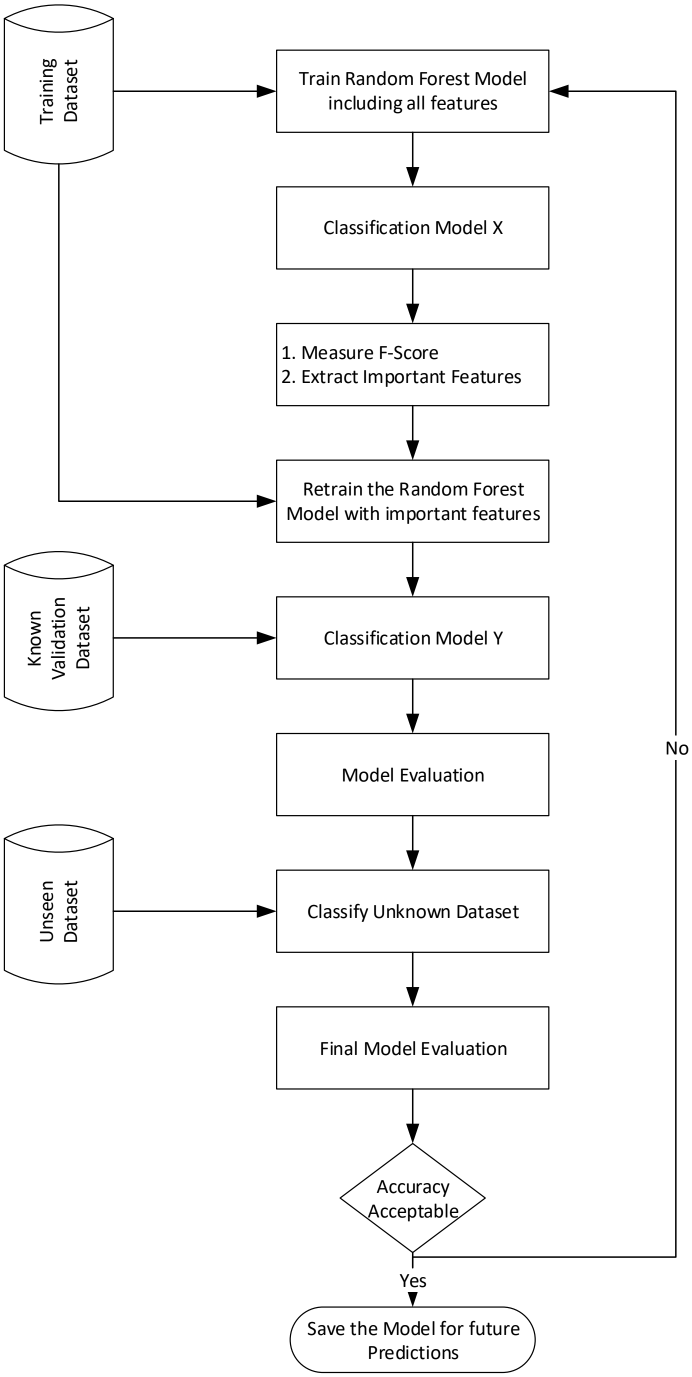


Figure 1 *Classification Approach*

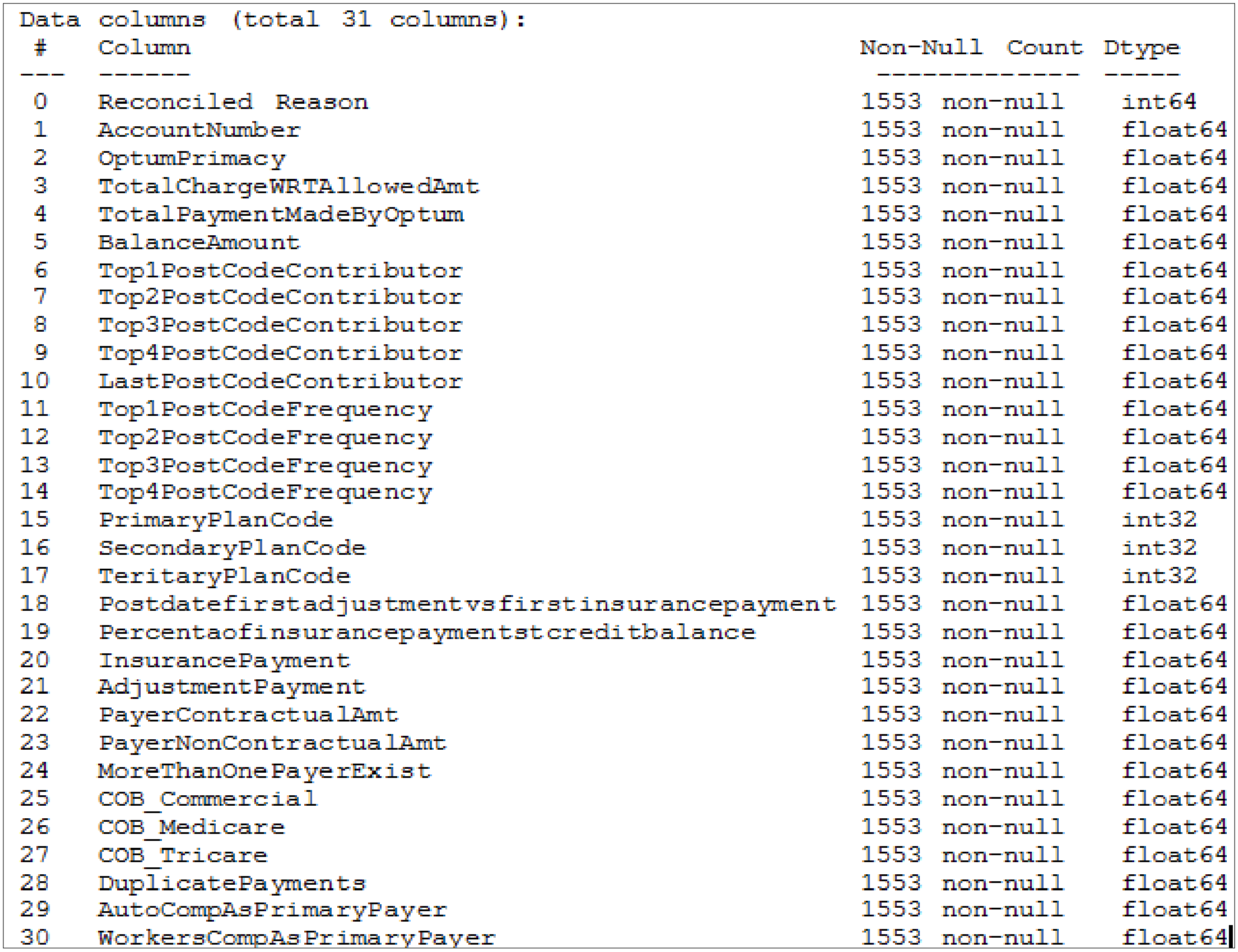
.

4. EXPERIMENTS AND RESULTS:

4.1 Data Preprocessing and Feature Selection:

First, the health raw data is received from the business teams, which contained raw demographics, raw insurance data and transaction data. Then, the raw data is imported to RDBMS system in order to do initial data preprocessing. A correct mix of data was prepared of different target classes i.e. Payer Refund and Non Payer Refund etc. Feature Selection was done followed by data cleaning and pre-processing and below features were identified.

Table 2 *Identified features post data-cleaning and preprocessing*



```
Data columns (total 31 columns):
 #    Column                                          Non-Null  Count  Dtype
---   ------                                           -------------  -----
 0    Reconciled Reason                               1553 non-null    int64
 1    AccountNumber                                   1553 non-null    float64
 2    OptumPrimacy                                    1553 non-null    float64
 3    TotalChargeWRTAllowedAmt                        1553 non-null    float64
 4    TotalPaymentMadeByOptum                         1553 non-null    float64
 5    BalanceAmount                                   1553 non-null    float64
 6    Top1PostCodeContributor                         1553 non-null    float64
 7    Top2PostCodeContributor                         1553 non-null    float64
 8    Top3PostCodeContributor                         1553 non-null    float64
 9    Top4PostCodeContributor                         1553 non-null    float64
10    LastPostCodeContributor                         1553 non-null    float64
11    Top1PostCodeFrequency                           1553 non-null    float64
12    Top2PostCodeFrequency                           1553 non-null    float64
13    Top3PostCodeFrequency                           1553 non-null    float64
14    Top4PostCodeFrequency                           1553 non-null    float64
15    PrimaryPlanCode                                 1553 non-null    int32
16    SecondaryPlanCode                               1553 non-null    int32
17    TeritaryPlanCode                                1553 non-null    int32
18    Postdatefirstadjustmentvsfirstinsurancepayment  1553 non-null    float64
19    Percentaofinsurancepaymentstcreditbalance       1553 non-null    float64
20    InsurancePayment                                1553 non-null    float64
21    AdjustmentPayment                               1553 non-null    float64
22    PayerContractualAmt                             1553 non-null    float64
23    PayerNonContractualAmt                          1553 non-null    float64
24    MoreThanOnePayerExist                           1553 non-null    float64
25    COB_Commercial                                  1553 non-null    float64
26    COB_Medicare                                    1553 non-null    float64
27    COB_Tricare                                     1553 non-null    float64
28    DuplicatePayments                               1553 non-null    float64
29    AutoCompAsPrimaryPayer                          1553 non-null    float64
30    WorkersCompAsPrimaryPayer                       1553 non-null    float64
```

4.2 Data Distribution: Further research found that the target class is highly imbalanced as below:

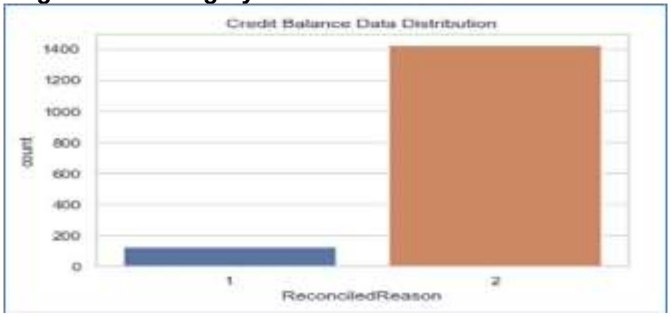
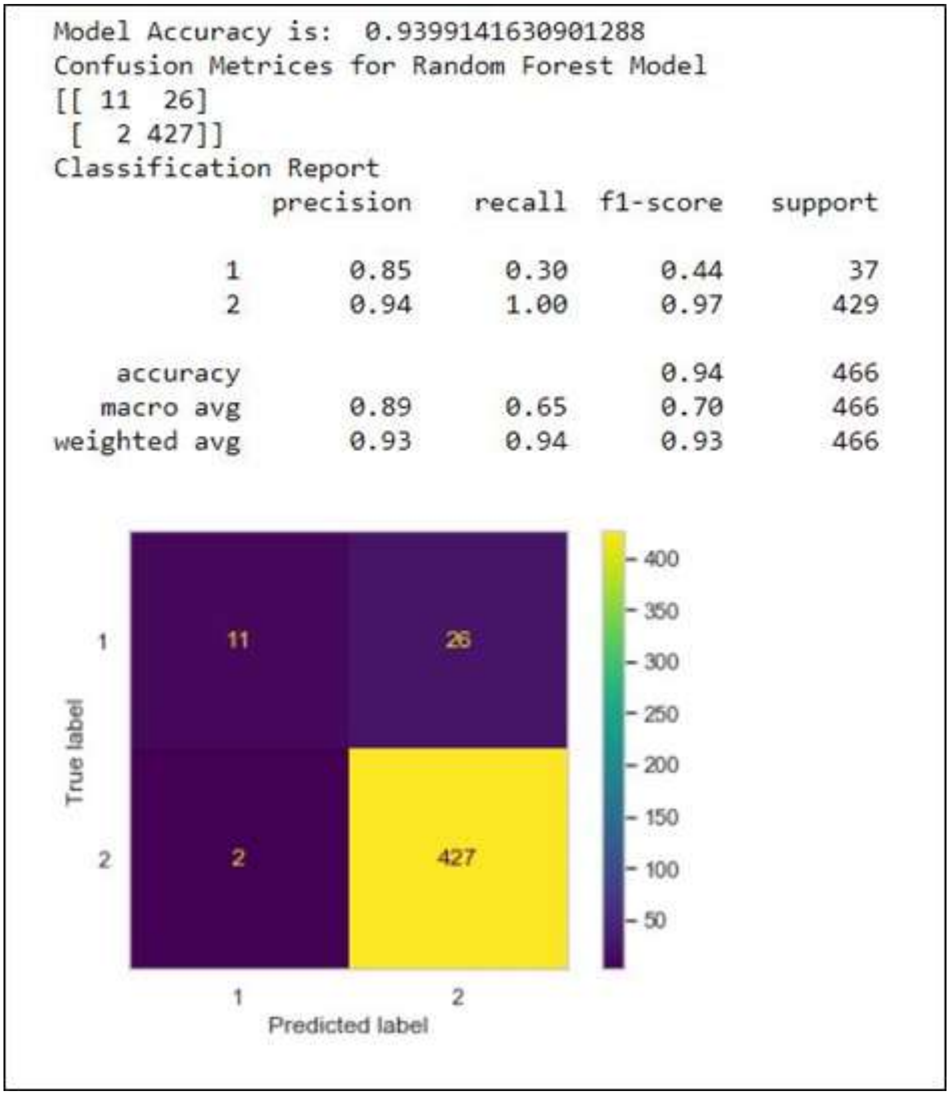


Figure 2 Data Distribution of Overpaid Positive and Negative Classes

Total Observations: 496 accountsPayer Refund (+Positive cases): 7% (37 accounts) only. Figure 2 shows the credit Balance highly biased data distribution where the ratio of True Positive and True Negative are 7% and 93% espectively. To build the working model on highly imbalanced data, it needs special ensemble technique like Bagging and Boosting data. Therefore, random forest technique was applied with hyper-parameters and high-ranking features which gave the promising results. 4.3 Random Forest Model: Figure 3 shows the snapshot where 93% overall accuracy was calculated with precision of 85% on identifying potential overpaid claims for testing Model. Figure 3 Confusion Matrix shows that the random forest algorithm is capable of effectively classifying the 1 and 2 categories and specifically the higher precision of category 1. High precision of category 1 meets the primary objective of this modeling to predict the potential overpaid claims in order to prioritize the potential claims, which will enable auditors to focus on revenue generating accounts.



```
Model Accuracy is:  0.9399141630901288
Confusion Metrices for Random Forest Model
[[ 11  26]
 [  2 427]]
Classification Report
              precision    recall  f1-score   support

           1       0.85      0.30      0.44        37
           2       0.94      1.00      0.97       429

    accuracy                           0.94       466
   macro avg       0.89      0.65      0.70       466
weighted avg       0.93      0.94      0.93       466
```

Figure 3 Confusion *Matrix and Classification Results*

As mentioned above, Grid Search Cross Validation was used to identify the hyper-parameters and model was retrained with the identified hyper-parameters. Top ten features were extracted by trained model with best parameters property.

Figure 4 shows the top 10 features with cumulative weightage with the target variable. Figure5 shows the ROC curve to show the model capability of distinguishing between different classes. Our Model has the 89% ROC value which depicts the model generates the good quality of classification results.

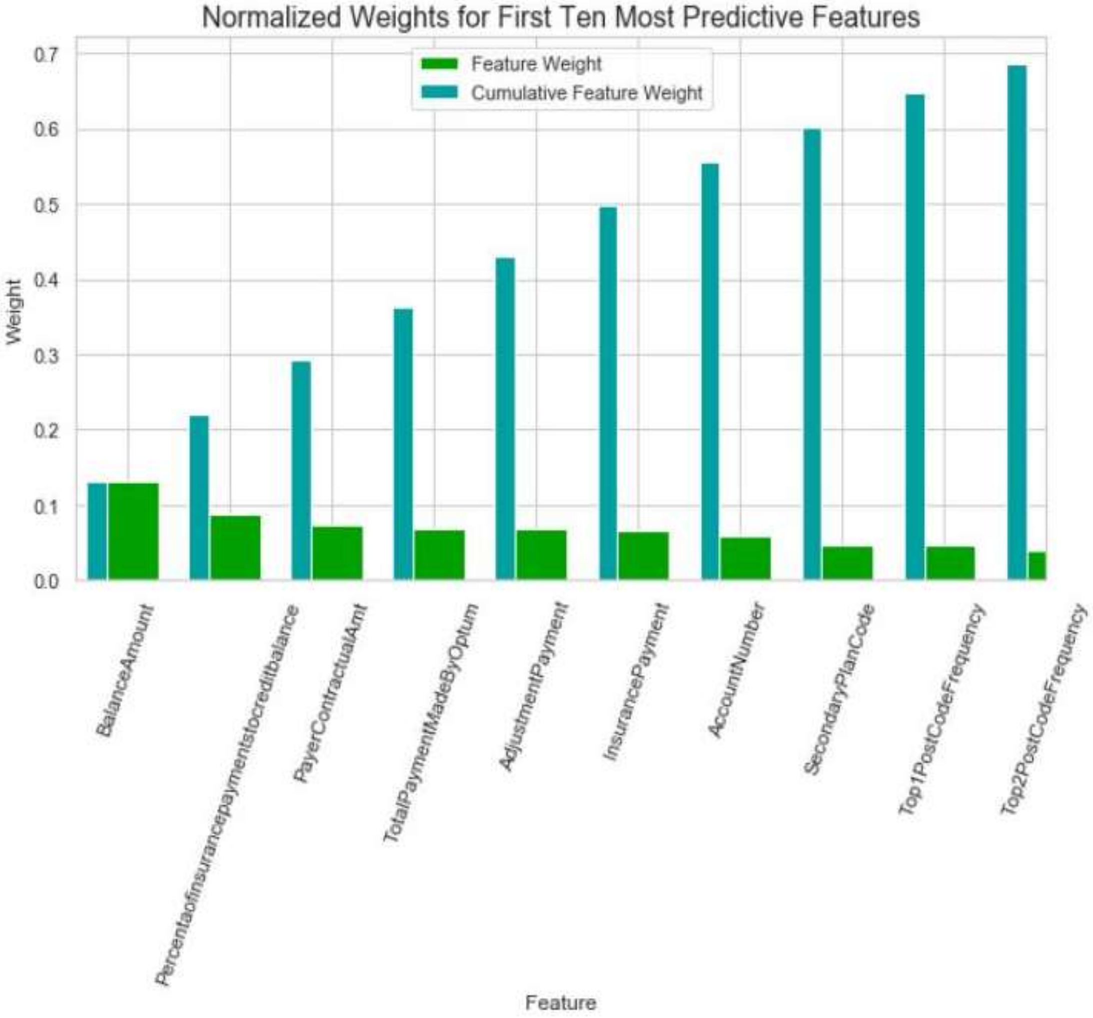
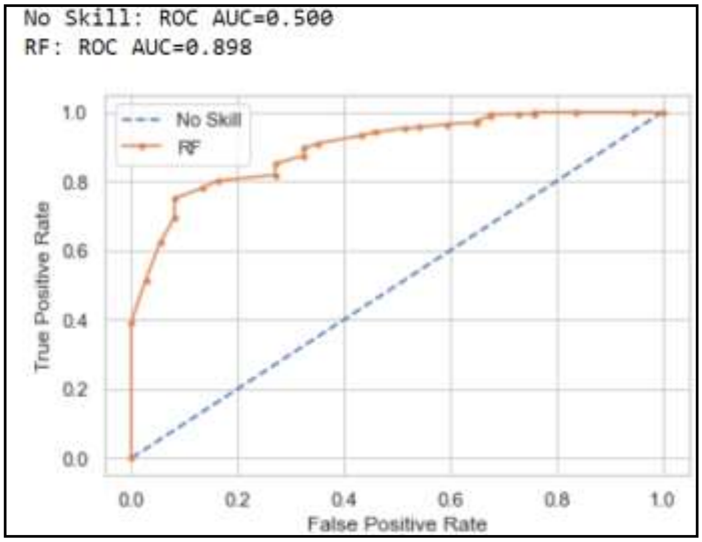


Figure 4 *Import Features Weightage with Target Variable (Reconciliation Reason)*



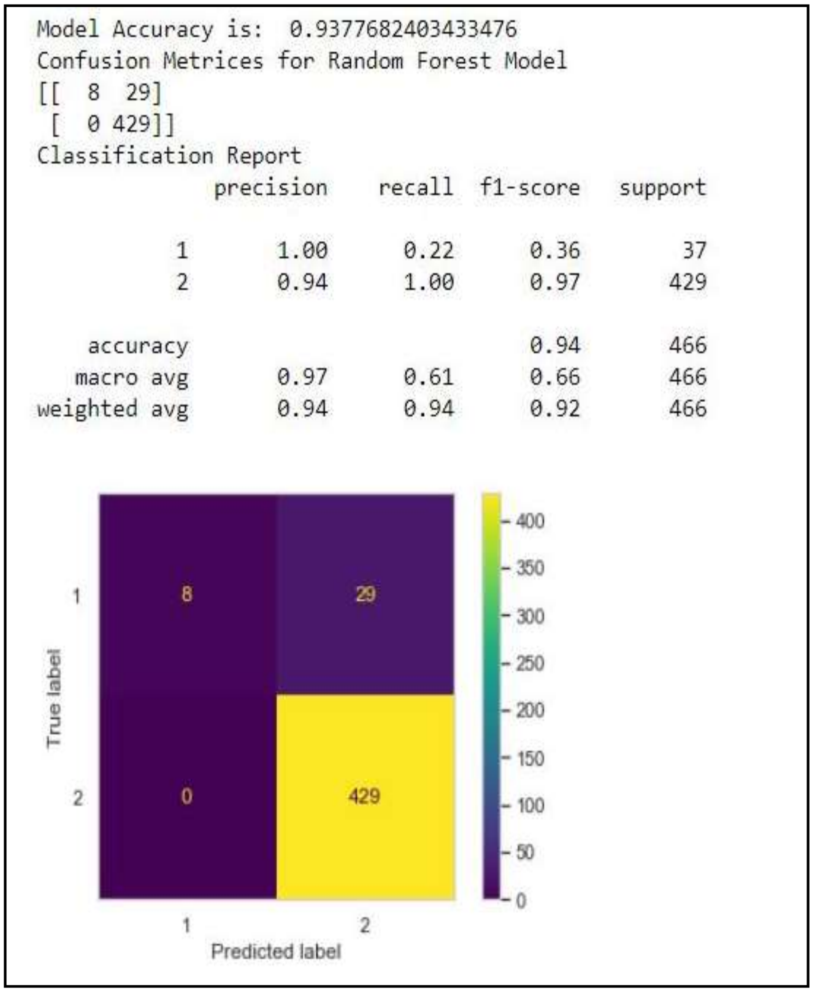
```
No Skill: ROC AUC=0.500
RF: ROC AUC=0.898
```

Figure 5 *ROC Curve*

In the final step, random forest model was trained with top features and hyper-parameters, which resulted an efficient mathematical algorithm and produced high precision results for predicting the true positives as displayed in Figure 6. In

order to train and validate the model, data was split in 70/30 ratio respectively, which provided the best accuracy. The hyper-parameters identified by Grid Search
Cross validation as below where 50 estimators were created as mentioned here:

Model Best Parameters :
{'max_depth': None,
'min_samples_leaf': 2,
'min_samples_split': 2,
'n_estimators': 50}



```
Model Accuracy is:  0.9377682403433476
Confusion Metrices for Random Forest Model
[[  8  29]
 [  0 429]]
Classification Report
              precision    recall  f1-score   support

           1       1.00      0.22      0.36        37
           2       0.94      1.00      0.97       429

    accuracy                           0.94       466
   macro avg       0.97      0.61      0.66       466
weighted avg       0.94      0.94      0.92       466
```

Figure 6 *Confusion Matrix with hyper - parameters*

Table 4 clearly states the better performance of positive classes of RF algorithm than other algorithms like Logistic Regression, Decision Tree.

Table 4 *Performance Results of different algorithms*

11Algorithms	Precision	Recall	Accuracy
Decision Tree	55	16	92%
Logistic Regression	18	89	67%
Random Forest	85	30	93%

CONCLUSION AND FUTURE SCOPE

Working on the identifying potential overpaid claims requires lot of manual efforts since auditors have to go through all the 100% cases whereas there are only 7% true positive cases. We have put efforts to overcome this problem by using Machine Learning approach and we have built random forest Classifier with the hyper-parameters with high weightage selected features. We had run the classifier against the validation data and it successfully classified the potential over paid claims, which is a great win. This will enable auditors to focus on revenue generating accounts, which make them more effective in overall credit balance solution, and meeting payer objective. We have seen that the RF classifier has the capability to classify results successfully with high precision. Still, it has the scope of further improvement with respect to improving the recall for the true positive cases. Therefore, it will give the more positive cases in hands to the auditors in the priority and work on them first.

REFERENCES

[1] L. Breiman, "ST4_Method_Random_Forest," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001, doi:

10.1017/CBO9781107415324.004.

[2] M. Pal, “Random forest classifier for remote sensing classification,” Int. J. Remote Sens., vol. 26, no. 1, pp. 217–222, 2005, doi: 10.1080/01431160412331269698.

[3] G. Cafri, L. Li, E. W. Paxton, and J. Fan, “Predicting risk for adverse health events using random forest,” J. Appl. Stat., vol. 45, no. 12, pp. 2279–2294, 2018, doi: 10.1080/02664763.2017.1414166.

[4] K. Taalab, T. Cheng, and Y. Zhang, “Mapping landslide susceptibility and types using Random Forest,” Big Earth Data, vol. 2, no. 2, pp. 159–178, 2018, doi: 10.1080/20964471.2018.1472392.

[5] A. O. Ok, O. Akar, and O. Gungor, “Evaluation of random forest method for agricultural crop classification,” Eur. J. Remote Sens., vol. 45, no. 1, pp. 421–432, 2012, doi: 10.5721/EuJRS20124535.

[6] J. Evans, B. Waterson, and A. Hamilton, “Forecasting road traffic conditions using a context-based random forest algorithm,” Transp. Plan. Technol., vol. 42, no. 6, pp. 554–572, 2019, doi: 10.1080/03081060.2019.1622250.

[7] W. Mao and F.-Y. Wang, “Cultural Modeling for Behavior Analysis and Prediction,” Adv. Intell. Secur. Informatics, pp. 91–102, 2012, doi: 10.1016/b978-0-12-397200-2.00008-7.

[8] G. Carvajal, M. Maucec, and S. Cullick, Introduction to Digital Oil and Gas Field Systems. 2018.

[9] E. S. Soares, K. P. Oliveira-Esquerre, A. M. de Aguiar, G. L. P. Botelho, and A. Kiperstok, Development of a model to identify combined use in residential water end use events, vol. 44. Elsevier Masson SAS, 2018.

[10] S. Conjeti et al., Domain Adapted Model for In Vivo Intravascular Ultrasound Tissue Characterization, 1st ed. Elsevier Inc., 2017.

[11] E. Iadanza, G. Guidi, and A. Luschi, 7 - Decision Support Systems in Healthcare BT - Clinical Engineering. Elsevier Inc., 2016.

[12] X. Gao, J. Wen, and C. Zhang, “An Improved Random Forest Algorithm for Predicting Employee Turnover,” Math. Probl. Eng., vol. 2019, 2019, doi: 10.1155/2019/4140707.