Name: Anuj bharat baghel

Email address: anujbagh300@gmail.com

Contact number: 9004672502

Anydesk address: 670717610

Years of Work Experience: fresher

Date: 06-01-2023

Self Case Study -1: *** Predict the potentially fraudulent providers based on the claims filed by them***

Problem link : https://www.kaggle.com/rohitrox/healthcare-provider-fraud-detection-analysis

---

"After you have completed the document, please submit it in the classroom in the pdf format."

Please check this video before you get started:
https://www.youtube.com/watch?time_continue=1&v=LBGU1_JO3kg

Overview

*** Write an overview of the case study that you are working on. ***(MINIMUM 200 words)***
***

INTRODUCTION

What is healthcare fraud?

fraud is nothing but ay deliberate and dishonest act committed by fraudulent that could result in unauthorized benefit to that people who is not proper claimant as per policy terms and condition. This is what we have to detect "Healthcare fraud detection" healthcare fraud detection in which any provider (agent,third_party,etc) fills the claim behalf of beneficiary . In currently medicals fraud is biggest crime problem for insurance sector. Due to increasing the number of fraud claim insurance company will directly effect the common peoples life and their savings which is not good at all. Ans as per policy terms and condition for legitimate client we have to pay amount in 30 days . so we have very less time to perform all investigation on suspicious claim . Due to fraud insurance companies are the major institute which impacted both private ad government insurance companies .

I have personally worked in insurance company domain and have explained this types of form of fraud are :

1. Sometime claimant submits duplicate form or gain submitting for same service they have opted .
2. Sometimes for individual policy holders agent submit form with-out policy-holders information for their personal benefit .
3. Some-times hospital charges more charges than original packages. To overcome this issue insurance companies has make ppn network of hospital and fixed their charges but still frauds happens as large level.
4. Misrepresenting the service provided.
5. Submit some claims that service they have never opted for.

Buisness problem:

1. After having statistical analysis, we can observe that total 15% medical expenses caused due to fraud claims. One of big and major institute which is impact most its insurance company. Due to this unauthorized or invalid practices insurances companies must increase their premium.
2. We have to predict the weather claim document provider is fraudulent or legitimate claimant and we also have to find activity and reason behind so we can prevent this in future .

3. One of major objective of this problem is to protect healthcare system other than financial losses .
4. Also we have to find-out most important features which are critical or imp to find-out fraudulent activity.

ML FORMULATION :

1. We have to build binary classification model for claims provided by provider along with beneficiary data, inpatient data , outpatient data whether they are fraud or not .

BUISNESS CONSTRAINTS:

1. Most importantly cost of misclassification could be very high , so we have to keep in mind FPR and TPR should be less as much as possible . because if any provider is fraudulent then it should not predict as legitimate claims also legitimate claims should not predict as fraud. Then it will predict insurer as fraud that will be huge loss.
2. Insurance company has to pay amount in 30 days. So there is no latency constraints but it should not take more than that . depend upon the model prediction insurance company has to put claim on manual investigation.
3. Model interpretability is also important because we have to justify why we have predicted to be as fraudulent . it should not black box model.

Columns dataset explanation:

Data set is consist of 4 csv files .

1. train csv (train-test both)
2. Train inpatient data (train-test both)
3. Train outpatient data (train-test both)
4. Beneficiary data (train-test both)

(1).Train-1542865627584.csv:

It consists of provider number and corresponding whether this provider is potentially fraud. Provider ID is the primary key in that table.

(2)Inpatient Data (Train and Test):

It consists of the claim details for the patients who were admitted into hospital. So, it consists 3 extra columns Admission date, Discharge date and Diagnosis Group code.

AdmissionDt: It contains the date on which the patient was admitted into hospital in yyyy-mm-dd format.
DischargeDt: It contains the date on which the patient was discharged from the hospital in yyyy-mm-dd format.
DiagnosisGroupCode: It contains group code for the diagnosis done on the patient.

(3)Outpatient Data (Train and Test):

It consists of the claim details for the patients who were not admitted into hospital, who only visited there. Important columns are explained below.

BeneID: It contains the unique id of each beneficiaries i.e patients.
ClaimID: It contains the unique id of the claim submitted by the provider.
ClaimStartDt: It contains the date when the claim started in yyyy-mm-dd format.
ClaimEndDt: It contains the date when the claim ended in yyyy-mm-dd format.
Provider: It contains the unique id of the provider.
InscClaimAmtReimbursed: It contains the amount reimbursed for that particular claim.
AttendingPhysician: It contains the id of the Physicican who attended the patient.
OperatingPhysician: It contains the id of the Physician who operated the patient.
OtherPhysician: It contains the id of the Physician other than AttendingPhysician and OperatingPhysician who treated the patient.
ClmDiagnosisCode: It contains codes of the diagnosis performed by the provider on the patient for that claim.
ClmProcedureCode: It contains the codes of the procedures of the patient for treatment for that particular claim.
DeductibleAmtPaid: It consists of the amount by the patient. That is equal to Total_claim_amount - Reimbursed_amount.

(4) Beneficiary Data (Train and Test): **This data contains beneficiary KYC details like DOB, DOD, Gender, Race, health conditions (Chronic disease if any), State, Country they belong to etc. Columns of this dataset are explained below.**

BeneID: It contains the unique id of the beneficiary.
DOB: It contains the Date of Birth of the beneficiary.
DOD: It contains the Date of Death of the beneficiary, if the beneficiary id deal else null.
Gender, Race, State, Country: It contains the Gender, Race, State, Country of the beneficiary.
RenalDiseaseIndicator: It contains if the patient has existing kidney disease.
ChronicCond_*: The columns started with "ChronicCond_" indicates if the patient has existing that particular disease. Which also indicates the risk score of that patient.
IPAnnualReimbursementAmt: It consists of maximun reimbursement amount for hospitalization anually.
IPAnnualDeductibleAmt: It consists of premium paid by the patient for hospitalization anually.
OPAnnualReimbursementAmt: It consists of maximun reimbursement amount for outpatient visit anually.
OPAnnualDeductibleAmt: It consists of premium paid by the patient for outpatient visit anually.

PERFORMANCE MATRIX:

1. **As we have observe that 90% providers are legitement claims , only 9.5% are them fraud claims . so this is highly imbalanced data . so in this scenario Accuracy cannot be as matrix to check the model performance . we have to check some other matric precision , recall, f1-score, confution matrix . with the help of confusion matrix we have to check the miss-classification points , we have to check FN , FP . FN means claims is legitment but predicted as fraud , FP mean is clais is fraud but predicted as legistment . this is what we have to keep lesss as much as possible .**
2. *So, the performance metrices are:*
   1. Confusion Matrix: It is the table where TP, FP, TN, FN counts will be plotted. From this table we can visualize the performance of the model.
   2. F1 Score: F1 Score = 2(Precision * Recall)/(Precision + Recall) where Precision = TP/(TP+FP) and Recall = TP/(TP+FN). As F1 score consists of both Precision and Pecall it will be correct metric for this problem.
   3. AUC Score: AUC stands for Area Under ROC(Receiver Operating Characteristics) Curve. ROC plots TPR with respect to FPR for different thresholds. The area under the curve depends on the ranking of the predicted probability score, not on absolute values.
   4. FPR and FNR: As cost of misclassification is very high, we need to check the FPR and FNR separately, It should be as low as possible.

Research-Papers/Solutions/Architectures/Kernels

*** https://www.kaggle.com/brandao/diabetes***

1st URL: https://medium.com/analytics-vidhya/healthcare-provider-fraud-detection-analysis-using-machine-learning-81ebf09ed955

HEALTHCARE PROVIDER FRAUD DETECTION ANALYSIS USING MACHINE LEARNING

Problem statement : Build a binary classification model based on the claims filled by the provider along with inpatient data , outpatient data , beneficiary details to protect healthcare provider fraud.

Table of Contents:

1. Introduction
2. Types of Healthcare Provider Fraud
3. Business Problem
4. ML Formulation
5. Business Constraints
6. Dataset Column Analysis
7. Performance metric
8. Exploratory Data Analysis
9. Existing Approaches and Improvements in my model
10. Data Preprocessing
11. Machine Learning Models
12. Final Data pipeline
13. Future work

INTRODUCTION :Fraud is defined as any deliberate and dishonest act committed with the knowledge that could result in the unauthorized benefit to person . we need to build a model to predict whether provider is fraud or not major institute which is impacted by them are insurance companies . due to increase the fraud insurance companies ha to increase their premiums.

(2) TYPES OF HEALTHCARE FRAUD :

1. billing for service that not provided .

2. duplicate submission for same service .

3. misrepresenting the service provided .

4. charging more charged than original package .

(3) Business constraints : a) The cost of misclassification s very high false negative and false positive should be less as much as possible .(b) if fraudulent providers are predicted as non-fraudulent this is financial losses to legitimate fraudulent . (c) it will cost for investigation and also its matter of reputation of the agency.(d) insurer should pay the claim within 30 days so there is no much business latency constraints .but it should not be more than that .

DATASET DATA ANALYSIS : dataset is given on Kaggle websites .

(1) Train data (train and test)
(2) Out-patient data (train and test)

(3) Inpatient data (train and test)
(4) Beneficiary data

EXPLORATORY DATA ANALYSIS :

(1) Dataset is highly imbalanced dataset.
(2) In beneficiary dataset the gender ratio is gender ratio as follow : male 57% : female43%
(3) In the beneficiary data top 5 states having code 5,10,5,33,39.
(4) Distribution of countries are code 200,0,20,60 and 0 with top 5 state .
(5) In distribution of race 85% beneficiaries belong to race 1.
(6) The total reimbursement amount for inpatient 507162970 and outpatient is 179876080. The inpatient reimbursement amount is 2.81 times higher than outpatient .
(7) The deductive amount for inpatient is 55401242 and 52335131 in both of dataset there exists some outlier with high values with tailed value .

ATTENDING PHYSICIAN (INPATIENT AND OUTPATIENT)

(1) Few physicians are treated 1% and 2% all the patient . that seems suspicious.
(2) The number of claims is less for inpatient data compared to outpatient data.
(3) The number to outpatient data inpatient data but fraudulent activity is more outpatient data. As percentage 57% frauds in inpatient data and 36% frauds in outpatient data . reason behind this is the reimbursement amount for inpatient data Is high than outpatient approximately 35 times higher .

Transaction Details :

(1) 25^{th} ,50^{th} percentiles are very less for claim amount reimbursed .

(2) 75^{th} percentile claims is higher than legitimate claim for insurance claim amount reimbursement.
(3) Age : from the scatter plot we can see that when patient age <60 years and claims period . more than 20 years the probability of transaction is fraudulent is high .
(4) Scatter plot for age vs insurance claim reimbursement : from the scatter plot of patient age vs "inscclaimrehimbursement" we can observed that if the patient age more than 60 then chance of fraudulent activity can be high .
(5) Scatter plot of "inputtotalamountrehimbursment" vs inscclaimrehimbursement : if inscclaimamountrehimbursed >600000 then chance to be fraudulent is very high .

EXISTING APPROCACHES AND IMPROVEMENT IN MODEL .

(1) In the existing approaches were used from sklearn no proper strategy were taken for data imbalance in the blog actor has used . random sampling using SMOTE to handle data imbalance . Along with first cut approaches have used ensembles model by custom change to get better performance.

DATA PREPROCESSING

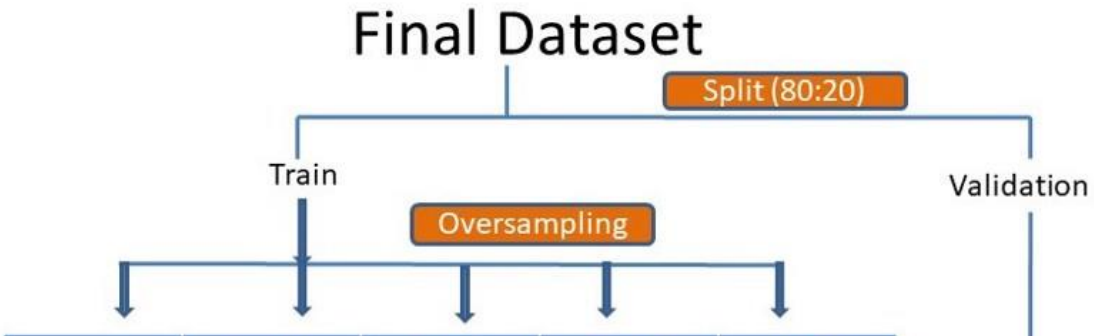
(1) Firstly author has calculated the age with the help of 2 features DOB and DOD .
(2) Calculated claim duration and hospitalization duration because if the number of days claimed for inpatient treatment is more than no of days hospitalized is suspicious so author added this features .
(3) Merge all the dataset : we have 4 different datasets which are interconnected by foreign keys. I need to merge them using foreign keys to get overall dataset.
(4) Merged inpatient and outpatient data based on column .

(5) Merge beneficiary details with inpatient data outpatient data on beneid.
(6) Merge provider details with previously merged data with provider id .
(7) Created total reimbursement amount inpatient and outpatient .
(8) Beneficiary are also associated with the fraud activity so for that author has computed the mean if that is higher than amount then its suspicious .
(9) The last operation also performed for attending physician, operating physician, other physician and take the mean of that if that higher than amount than physicians are also suspicious .
(10) Sometimes Providers along with physicians, beneficiaries, and sometimes diagnosis and procedures are also associated. So take another feature with provider id and group by. Take count after that.
(11) After these all the procedures authors has removed the features which not more useful to solve this problem .
(12) Now will standardized the dataset .

MODELING

(1) After this now author has performed the modeling and since this is imbalanced dataset so author has used the strategy of SMOTE sampling .

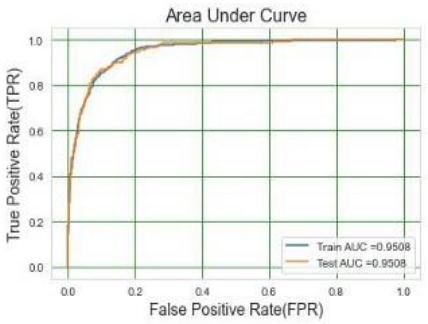
(2) Using the smote sampling these below result we have got acu score .

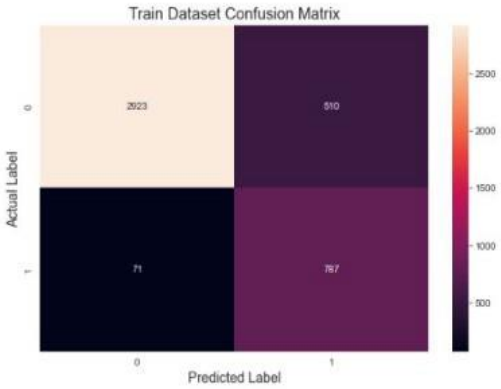


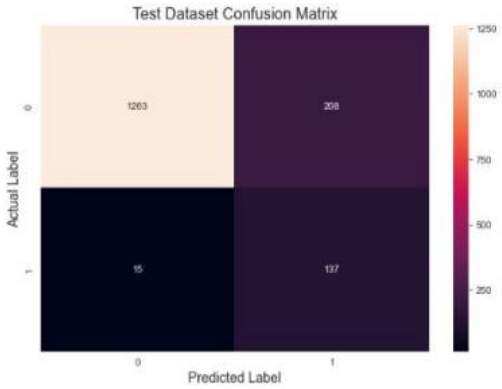
	80:20	75:25	65:35	50:50
LR	0.9508	0.9502	0.9501	0.9405
DT	0.9054	0.9185	0.9268	0.9108
SVC	0.9399	0.9368	0.9399	0.9345
NB	0.8793	0.8816	0.8820	0.8814

Pick the best one with Maximum AUC

(3) After performing the smote sampling the logestic regression has performed the best model with smote sampling of 80:20 auc 0.9508.



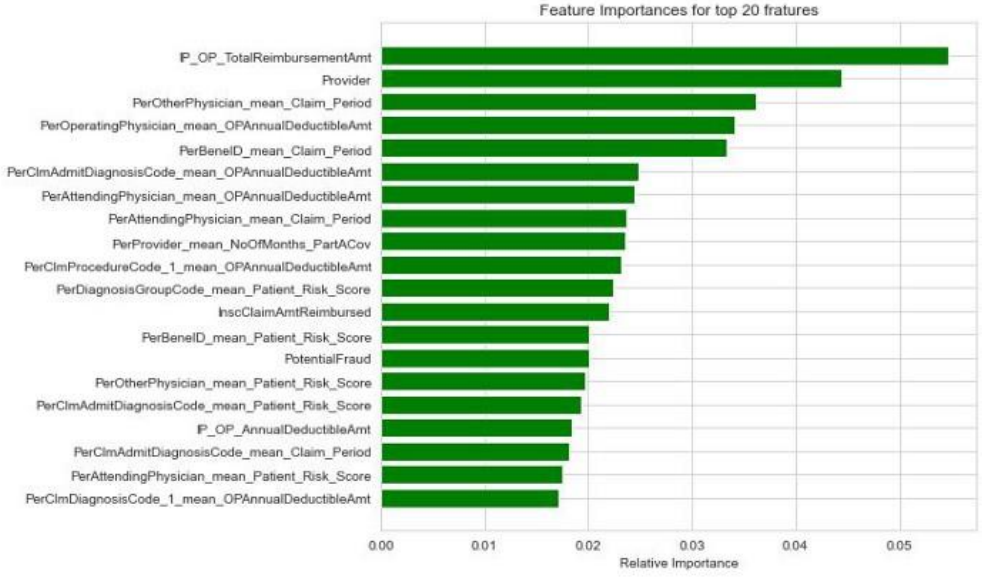




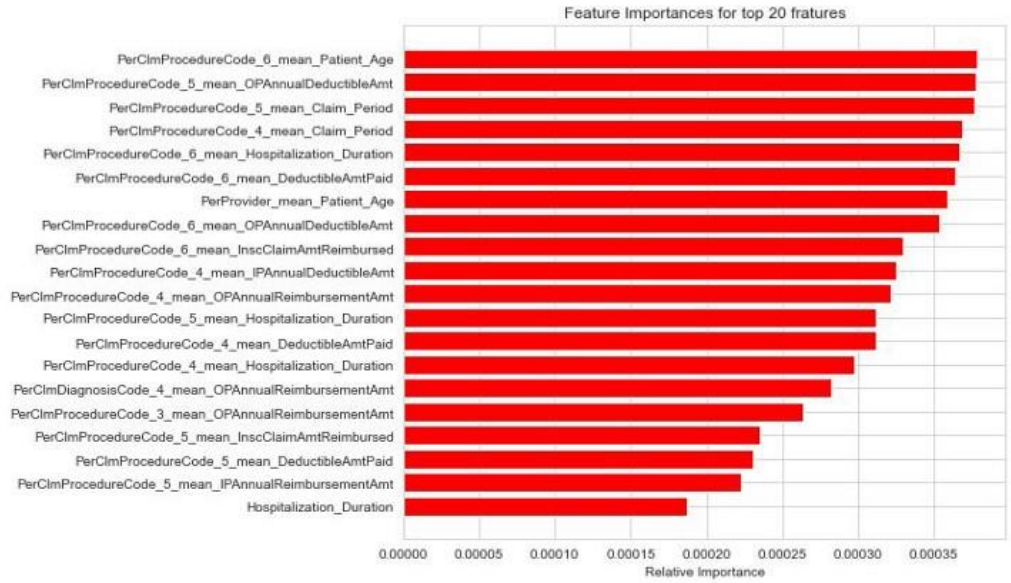

```
Best Threshold = 0.1105
Model AUC is : 0.9508
Model F1 Score is : 0.5513
```

(4) In next approach author is used random forest model with gini impurity to find which are features is most important to get best auc score .

(5) These are the most imp 20 features



(6) Thses are below lest imp features



(7) Using the these most imp features top 160 features this is below model performance author got

Sampling Ratio	Model	Features	AUC	F1 Score
80:20	Logistic Reg	All Features	0.951	0.551
80:20	Logistic Reg	Important Features	0.942	0.56
80:20	Random Forest	All Features	0.943	0.62
80:20	Random Forest	Important Features	0.943	0.634

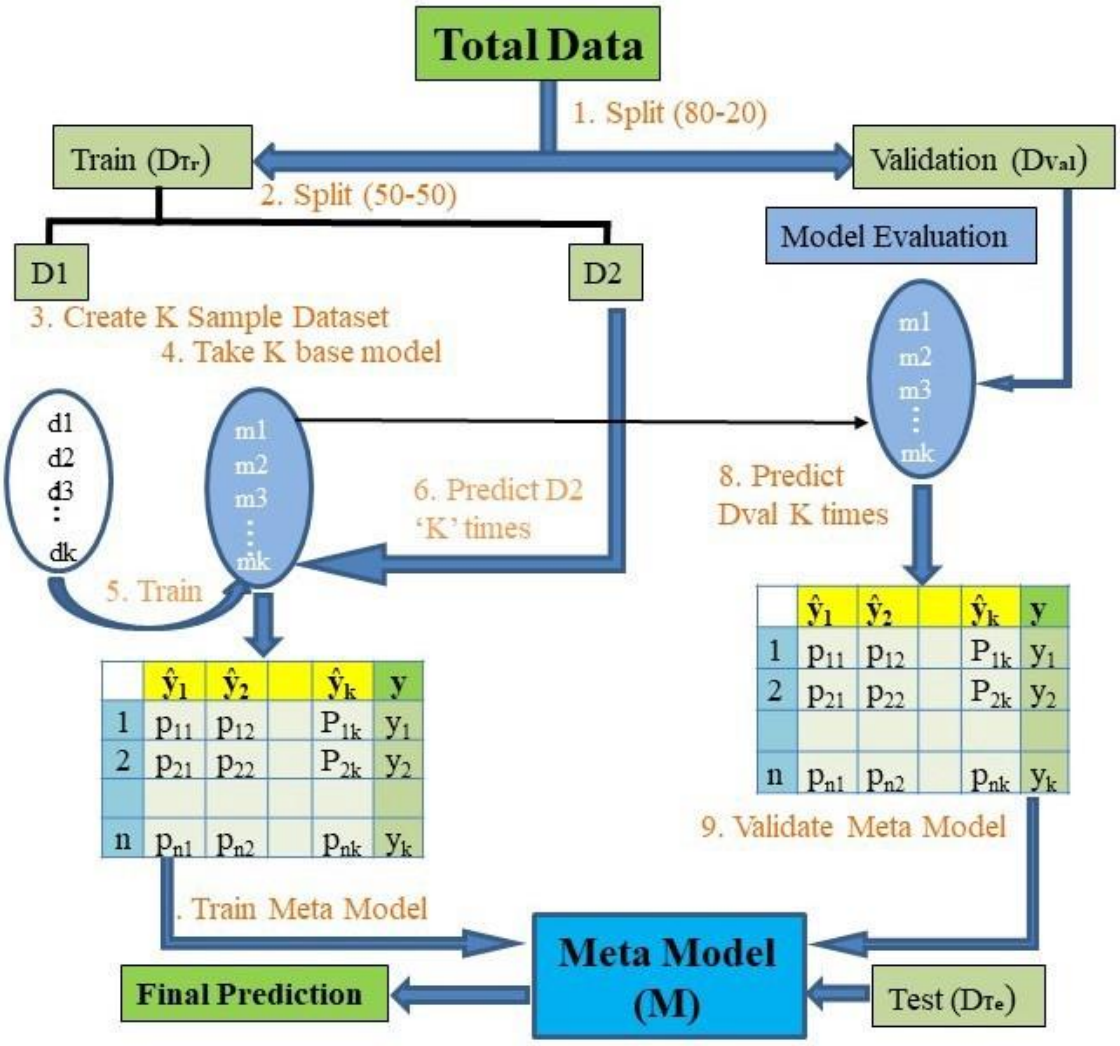
(8) Now at this time author has observed that logistic regression model has performed better than random forest but when we have draw confusion matrix than random forest performing better .

(9)

(10) Observation : when we have draw the confusion matrix with features of most imp features still there no such big improvement in plot . even we have used only imp features false negative rate is increased which is not good at all. As this medical problem we cannot afford the false negative rate increase so at this point according to author logistic regression is best model for this problem .

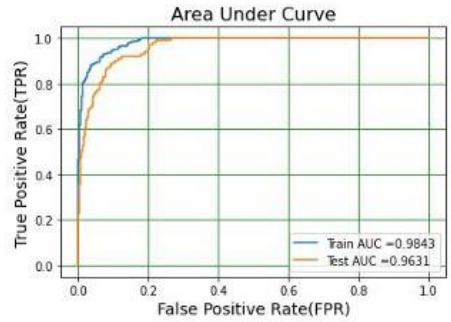
APPROCAHES 3

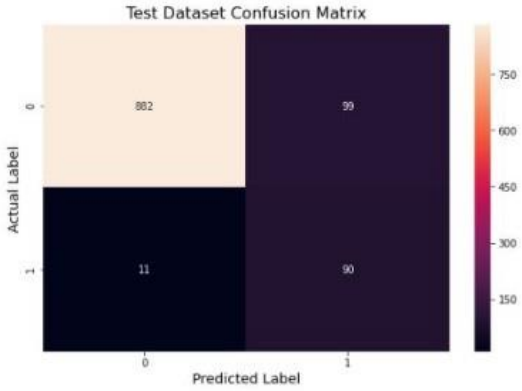
(1) In this approaches author has decided to go with ensemble techniques along with the meta data to get the better performance .
(2) These below is architecture that author has followed .



Meta Model	AUC	F1 Score
Logistic Regression	0.958	0.522
Decision Tree	0.94	0.517
SVC	0.956	0.491
RF	0.963	0.621

(3) OBSERVATION : Using the custom implementation of ensembles using 50 different combination random forest worked really well . Test AUC = 0.9631






```
Best Threshold = 0.1797
Model AUC is : 0.9631
Model F1 Score is : 0.6207
```

APPROCHES 4

(1) In this approaches author has used StackingCVClassifier from mlextend but this is not worked really well as for train and test auc .

Conclusion

Overall conclusion is construct a research to develop the model for healthcare fraudulent detection which model should be scalable , accurate , fast . scalable refer to a in future we can work with the high volume data , accurate refers to model should give high f1score value and less false positive rate and false negative rate , fast refers to it should produce the good result the in read world data .

AUTHORS FINAL APPROCHES : **so now finally authors has choose the ensemble technique by custom implementation using different combination of data for final approaches** .

How this paper useful in my case study

1. Brief overview : I have need to merge all 4 csv files to get the brief idea of the features for patient and providers.
2. Handle imbalanced dataset : I have a got the idea from this blog to handle the imbalanced dataset using smote splitting using different combination .
3. How to fill missing value: In this blog author has used model based imputing to handle missing data ,null data . such as total reimbursed amount , total deductive amount ,insurance amount .
4. Selected features : in this case I will use co-relation technique to check the correlation technique between the features. And according to that will remove the more related features.
5. Data transformation : in this case I will after merging the dataset we can find out the average of the amount reimbursed per user and total amount reimbursed per provider .
6. Model selection: I will use supervised machine learning model such as logistic regression , svm , knn, decision tree . and performance matric such as false positive rate , false negative rate, f1 score . after that for getting better model performance I will use custom implementation of ensemble techniques.

URL 2: https://towardsdatascience.com/for-real-auto-insurance-fraud-claim-detection-with-machine-learning-efcf957b38f3

PROBLEM STATEMENT

The goal of this project to build a model that can detect the auto insurance fraud . the chance behind this problem to reduce the loss of insurance company and deal with highly imbalance data. Because fraudulent claims are far less than legitimate claims .

RELAVANCE TO BUISNESS

As we know that in the real world data its very common to work on imbalanced data. Where the majority class is high data compare with minority class. These are some classification problem example given in blog below For instance, classification of other types of frauds, classification of defective goods, classification of at-risk teenagers, identifying high potential employees, identifying people of interest such as terrorist,

Criteria for success

Since this is imbalanced dataset so according to the author want to classify the claim if that is fraud or legitimate claims . since this is highly imbalanced data classification problem so author has decided to go with f1 score as performance matrix because he want to analyze the harmonic mean of precision and recall. Secondly author has decide to go with roc-auc curve . and according to author auc value should be more 0.50 since this is Mediclaim problem . for this task author has decided to keep over 0.70 auc value.

BACKROUND OF INSURANCE FRAUD

According to author Insurance fraud is a deliberate deception perpetrated against or by an insurance company or agent for the purpose of financial gain. And according to author fraud can take place by anyone like claim-provider, hospital, third party people, insurance agent. And these people can submit the claim for that never occurred for like accident , damage.

According to FBI insurance fraud is about 40 billion . excluding the healthcare insurance. Auto insurance fraud is most of are misrepresenting the information of vehicle damage, accident .

Below info taken from blog

Fraud accounted for between 15 percent and 17 percent of total claims payments for auto insurance bodily injury in 2012, according to an Insurance Research Council (IRC) study. The study estimated that between $5.6 billion and $7.7 billion was fraudulently added to paid claims for auto insurance bodily injury payments in 2012, compared with a range of $4.3 billion to $5.8 billion in 2002.

EXECATIVE SUMMARY

The summary of this project to build a classification model that can detect the claim is fraud or not . in this project major issue is that frauds are far less than legitimate claims in numbers . so this is highly imbalanced dataset problem .

In this project author has tried different method to handle the data imbalanced . the best model that fitted to problem is using ensemble techniques .

According to author final model that fitted and performed well is weighted xg-boost . which gave the f1 score of 0.72 with auc-roc score 084. Which was far better than simple baseline model 0.397 and auc score 0.7. that was best score and model performance that author got from xg boost . and according to author model is successfully able to distinguish between if claim is fraud or not . before doing the modeling author has done some good amount of pre-processing on dataset . that we will discuss in more details in next page .

DATASET OVERVIEW

Data source: https://www.kaggle.com/roshansharma/insurance-claim

This dataset s containing the n = 1000 samples . since this not a big dataset. With that we have to make model to classify the frauds . so this is only about 1000 frauds . since this is small data but still author believes that with dataset he can make the robust model .

According to author in 2015 MIT REVIEW the professor called andrew ng has said that there are many companies who does have the big or huge to make the model for production . but still companies tries to make model that perform well in production . and whenever companies has big data they makes the model with huge data so this not a big problem according to author

EXPLORATERY DATA ANALYSIS

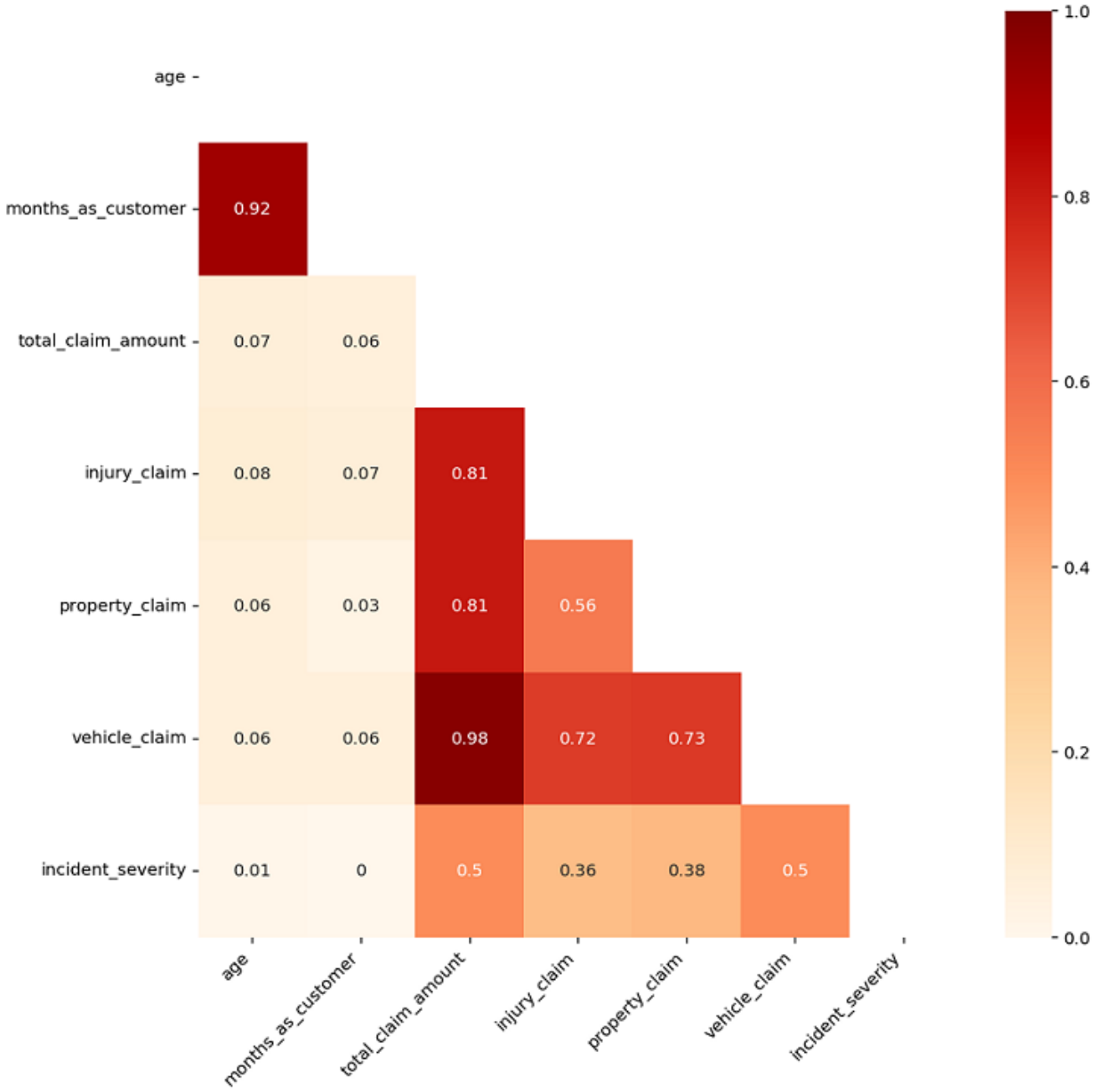
DEPENDENT VARIABLE : Eda were started from dependent variable on claim data . first author checked that how many are of them are fraud so among them 247 are fraud and 753 providers are not fraud . only 24% provider are fraud . and 75% are the non fraud.

Correlation in data : author has check the correlation between independent features and he plotted heatmap for 0.3 pearson correlation matrix . month and age has correlation of 0.92 this is because the drive buy vehicle from that month .

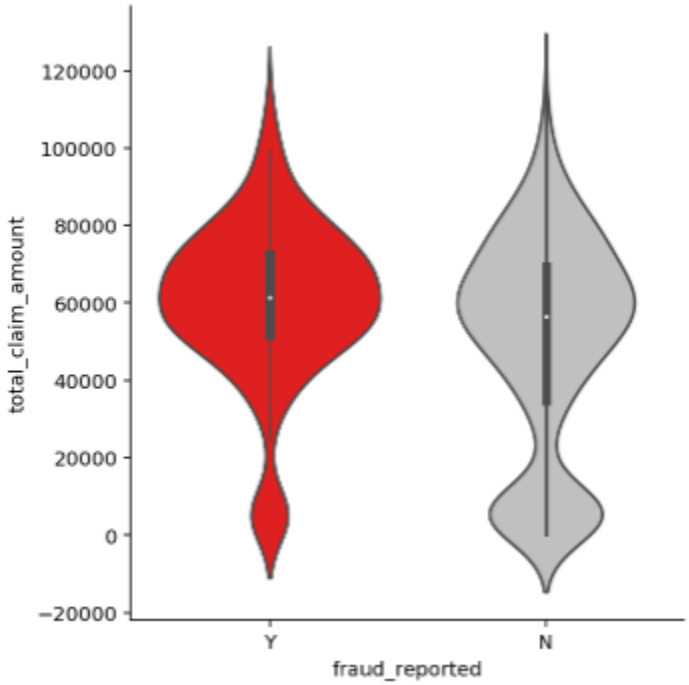
Apart from this author has not found much correlation from data . if data was highly correlated then there may be the multicollinearity problem .

Refer this

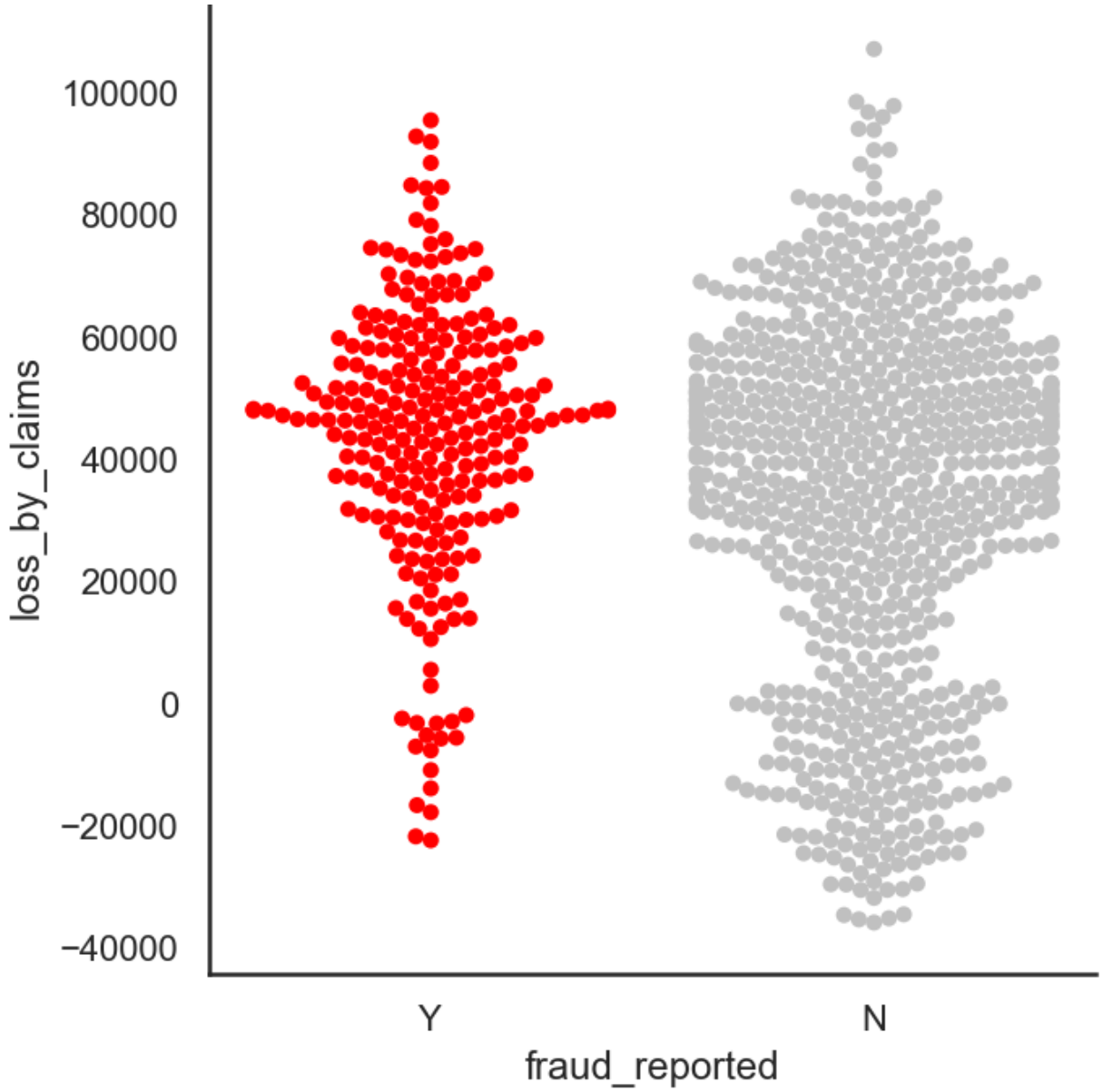
below image for correlation



below is the image of mean of fraud and non fraud claims



losses by



BASELINE SCORE

Since this is highly imbalanced dataset problem , so accuracy cannot use as matric because if we use the accuracy as a matric then result will be biased to majority class . and it will give the accuracy of 75% for non fraud because that is contribution in data of non frauds . this is not good at all.

If we make a naive prediction that all claims are frauds, so that no frauds escape our watch, we will have a score as shown below:

- Sensitivity: 1.0
- Specificity: 0.0
- Precision: 0.248
- F1 score: 0.397
- ROC AUC Score: 0.50

As you prior goal to predict the number of frauds as possible . so we will use the f1 score as out matric for this problem due to imbalanced in the problem . and also for know how much they are fraud and non fraud we will also use the auc as matric . where usually we checks 0.50 dataset . but for this problem author take took the limit of 0.70.

Modeling

Five different classifiers were used in this project:

- logistic regression
- K-nearest neighbours
- Random forest
- XGBoost
- AdaBoost

1. In this author has done the hyperparameter search for the parameters . so here he has used the rendomized search cv over the grid search cv because random search cv works faster than gridsearch .
2. After this author as computed the mean accuracy score with the random search cv on hyperparameter parameters on train and test dataset . also computed the f1 score , auc-roc score

1. Model with class weighting and hyperparameter tuning.

Below explanation taken from blog

A best practice for using the class weighting is to use the inverse of the class distribution present in the training dataset. For example, the class distribution of the test dataset is a 1:100 ratio for the minority class to the majority class. The inversion of this ratio could be used with 1 for the majority class and 100 for the minority class; for example:{0:1.0, 1:100.0}. In our case, class weights were {0:0.246667, 1:0.75333}. In XGBoost, class weights are defined differently. XGBoost uses scale positive weight which is the total negative examples divided by the total positive examples. For an imbalanced binary classification dataset, the negative class refers to the majority class (class 0) and the positive class refers to the minority class (class 1). The scaled positive weight in this analysis is 3.054.

2. Modeling with Oversampling using SMOTE

As author as used the smote technique to for sampling using this technique we have 568 point in training dataset for both class . this is how we use smote to handle imbalanced dataset .

For more explanation given in blog

The five classifiers were running on a SMOTE data set, with hyperparameter tuning. SMOTE works by selecting examples that are close in the feature space, drawing a line between the examples in the feature space and drawing a new sample at a point along that line. Specifically, a random example from the minority class is first chosen. Then k of the nearest neighbors for that example are found (typically k=5). A randomly selected neighbor is chosen, and a synthetic example is created at a randomly selected point between the two examples in feature space.

3. Modeling with Oversampling using ADASYN

In this techniques of adasyn author has used to handle the imbalanced dataset. This technique is almost similar to smote technique just a small correction adding some value to the training dataset , so the dataset should be like more realistic . so the both class have 565 points in training dataset as we have introduce the some noise in data using ADASYN technique using hyperparameter search .

4 Modeling with Oversampling using bootstrapping

As we have already used the technique of smote and adasyn to the training data . so the training data has equal values 565 in both classes . then the these values are already in distribution which were add by these technique But now we are going to do this same thing but using bootstrap technique randomly we will add the values to the biased will be less . and values will not be repeated more Bootstrap was only done on the training set. The five classifiers were running on the bootstrapped data set, with hyperparameter tuning**.**

5. Approaches followed for modeling .
    1. Author has used the technique of ensemble .in this technique uses the multiple base learns to improve the overall model performance using voting . before ran this ensemble Techni author has checked the correlation , so got that xg boost, random forest has high correlation . The ensemble model will use the best logistic regression, KNN and the best of XGB, Random forest and AdaBoost (ensemble 3 model), based on F1 scores, from the models with class weighting, models with oversampling by SMOTE, ADASYN and Bootstrapping**.**

       The tree models selected are:
       - Logistic regression with SMOTE (F1: 0.41)
       - KNN with bootstrapping (F1: 0.42)
       - Weighted XGBoost (F1: 0.72)
    2. Overall the author has the ensemble modeling with sampling of smote and adasyn with different approaches . sometimes he used the without oversampling and sometime with oversampling

This is overall observation of modeling

	No oversampling					With Smote					With ADASYN					With Bootstrap				
	LogReg	KNN	RanFor	XGB	AdaBoost	LogReg	KNN	RanFor	XGB	AdaBoost	LogReg	KNN	RanFor	XGB	AdaBoost	LogReg	KNN	RanFor	XGB	AdaBoost
CV scores:	0.77	0.77	0.80	0.84	0.81	0.83	0.54	0.86	0.88	0.87	0.83	0.53	0.85	0.87	0.87	0.73	0.81	0.92	0.93	0.90
Train score:	0.84	1.00	0.98	0.97	0.90	0.83	1.00	0.99	1.00	0.89	0.72	1.00	0.98	1.00	0.86	0.80	1.00	1.00	1.00	0.95
Test score:	0.71	0.74	0.78	0.84	0.78	0.71	0.30	0.78	0.82	0.80	0.71	0.31	0.79	0.81	0.80	0.62	0.65	0.79	0.78	0.76
Sensitivity:	0.34	0.16	0.53	0.81	0.42	0.40	0.95	0.52	0.58	0.55	0.39	0.97	0.50	0.58	0.57	0.48	0.52	0.57	0.53	0.55
Specificity:	0.82	0.93	0.87	0.86	0.89	0.81	0.09	0.87	0.90	0.88	0.81	0.10	0.89	0.89	0.88	0.67	0.69	0.87	0.87	0.83
Precision:	0.38	0.44	0.57	0.65	0.57	0.41	0.26	0.57	0.67	0.60	0.41	0.26	0.60	0.63	0.60	0.33	0.36	0.58	0.57	0.52
F1:	0.36	0.24	0.55	0.72	0.48	0.41	0.40	0.54	0.62	0.57	0.40	0.41	0.54	0.61	0.58	0.39	0.42	0.57	0.55	0.53
ROC AUC:	0.64	0.67	0.83	0.84	0.83	0.64	0.55	0.81	0.83	0.80	0.63	0.57	0.80	0.82	0.80	0.62	0.60	0.83	0.82	0.79

	No oversampling	Oversampled
	Max voting	Max voting
CV scores:	0.81	0.91
Train score:	1.00	1.00
Test score:	0.78	0.76
Sensitivity:	0.37	0.57
Specificity:	0.92	0.82
Precision:	0.59	0.51
F1:	0.46	0.55
ROC AUC:	0.80	0.77

	No oversampling	Oversampled
	Blending	Blending
CV scores:	0.98	0.98
Val score:	0.98	0.99
Test score:	0.74	0.60
Sensitivity:	0.42	0.47
Specificity:	0.85	0.65
Precision:	0.48	0.31
F1:	0.45	0.37
ROC AUC:	0.78	0.74

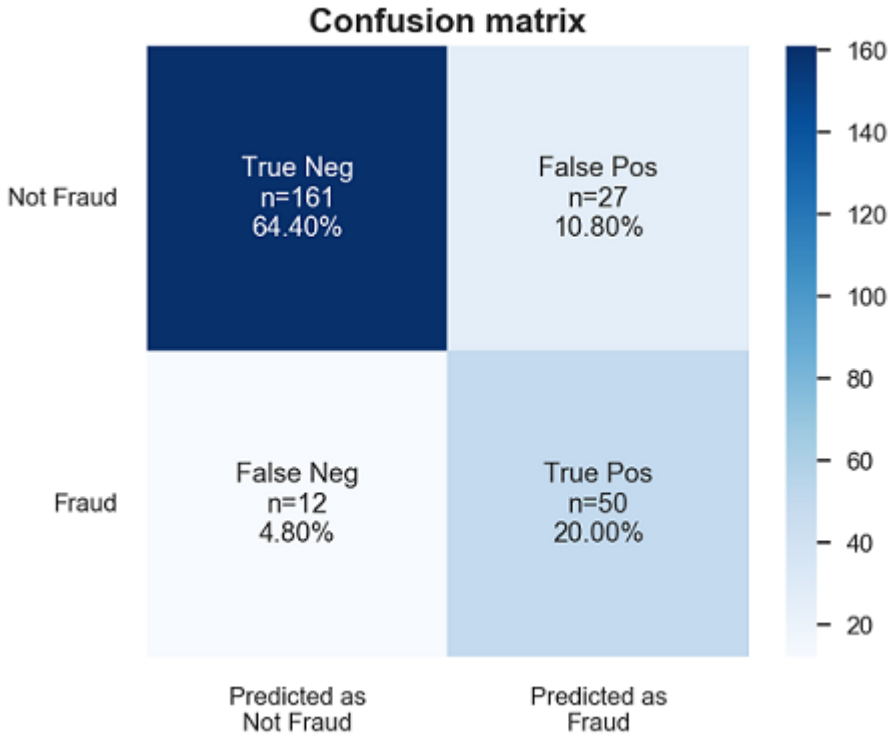
Final model choosed with thses paramters using hyperparamter search

The final fitted model were used that is weighted xg boost with these below hyperparameters

- Scale_pos_weight: 3.0540540540540 54,
- Reg_lambda (L2 regularization weight): 0.1,
- Reg_alpha (L1 regularization weight): 0.05,
- N_estimators: 550,
- Max_depth: 6,
- Gamma: 3,
- Eta: 0.05

Final fitted model performance

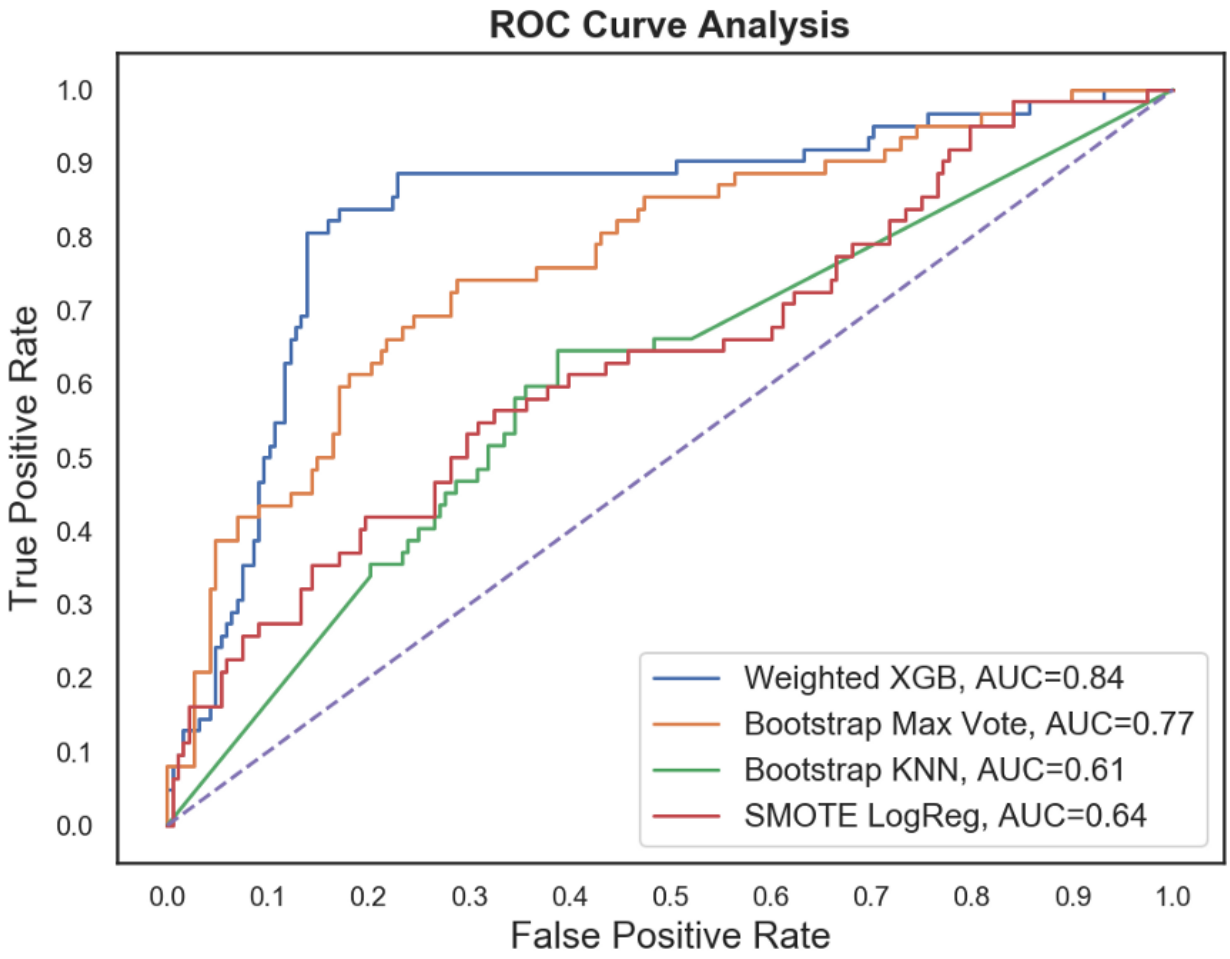
1. These model were giving the train accuracy of 0.96 and test 0.84
2. An accuracy score difference of 0.123 between train and test is relatively small. Thus, this model can be said to have low variance and is generalizable on unseen data.



The summary of the classification report is presented below.

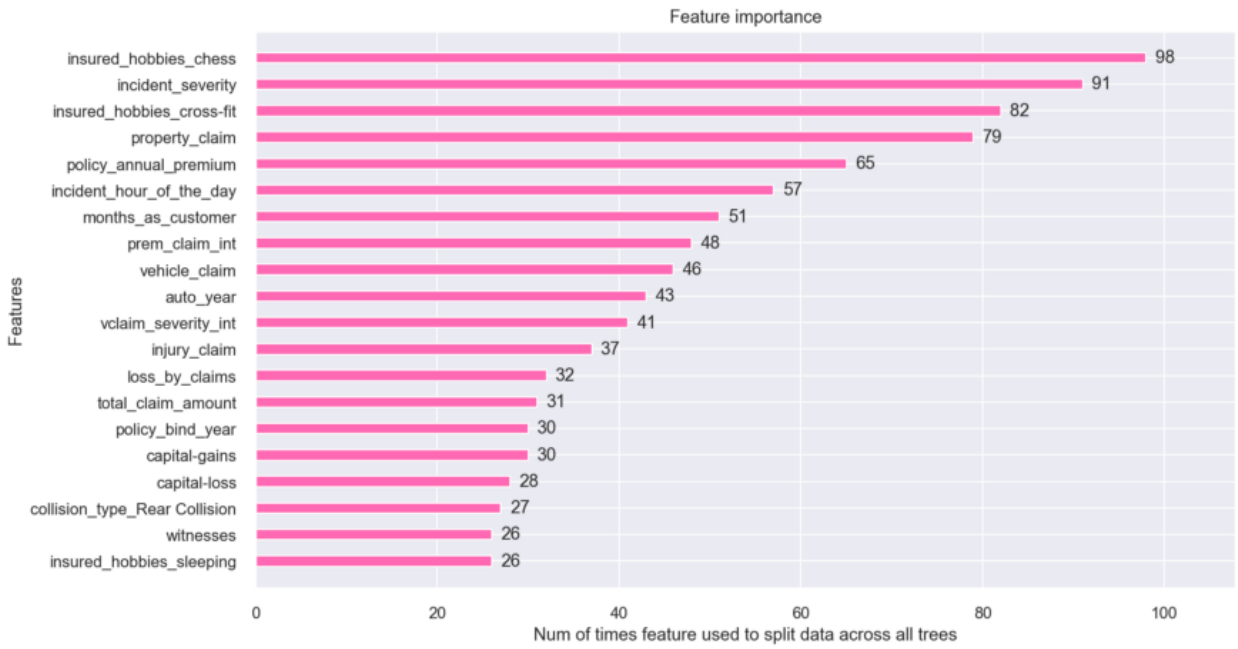
1. Sensitivity for the final model is 81%.
2. Specificity for the final model is 86%
3. Precision of fraud detection is 65%.
4. Precision of non-fraud detection is 93%.
5. The F1 score of the model is 72%.
6.

	precision	recall	f1-score
Not Fraud	0.93	0.86	0.89
Fraud	0.65	0.81	0.72


ROC Curve Analysis
True Positive Rate
False Positive Rate
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Weighted XGB, AUC=0.84
Bootstrap Max Vote, AUC=0.77
Bootstrap KNN, AUC=0.61
SMOTE LogReg, AUC=0.64

7.

Top features using weights



Overall observation and conclusion

1. This project has built a model that can detect auto insurance fraud. In doing so, the model can reduces loses for insurance companies. The challenge behind fraud detection in machine learning is that frauds are far less common as compared to legit insurance claims.
2. Five different classifiers were used in this project: logistic regression, K-nearest neighbors, Random forest, XGBoost, AdaBoost.
3. In this project author has used these 4 techniques to handle imbalanced dataset weighting , oversampling with smote , oversampling adasyn , sampling using bootstrap .
4. The best and final fitted model was a weighted XGBoost what yelled a F1 score of 0.72 and a ROC AUC of 0.84. In conclusion, the model was able to correctly distinguish between fraud claims and legit claims with high accuracy.

How this blog usefull for my casestudy

1. Firstly I was dealing with only healthcare fraud problems in my case study but after studying the this case study I have got the idea how to handle overall some other fraud related problems related to same insurance domain.

2. In this blog most imp learning for me was the learn diff techniques to handle imbalanced dataset .

3. In this blog I have learned technique like feature weighting , oversampling with SMOTE, oversampling with ADASYN, bootstrapped sampling .

4. In this I have also learned that what is better way to use the ensemble technique using different sampling classifiers .

5. In this also author has focus more on f1 score and auc score for performance matric .

6. In this blog I have learned that how sensitivity and specificity is imp in performance matric for imbalanced dataset .

URL 3 : https://cpb-us-w2.wpmucdn.com/sites.gatech.edu/dist/4/216/files/2015/09/p70-Statistical-Methods-for-Health-Care-Fraud-Detection.pdf

A survey on statistical methods for health care fraud detection

Jing Li, Kuei-Ying Huang, Jionghua Jin, Jianjun Shi

INTRODUCTION :

Since 1980 healthcare sector is most evolving field in US . so that's 'why its also most targeted sector for the purpose of fraud.

OBJECTIVE

1. Main objective is to reduce the loss on insurance companies so they can provide quality service to legitimate providers.
2. Indirectly we also reduce the loss of medical expenses .
3. We want to protect the insurance provider from loss .

CLASSIFICATIONOF FRADULENT BEHAVIOURS

Main these below parties involved in fraud .

1. Patient claim providers , agents .
2. Hospital, doctor, ambulance , medical , laboratories .
3. Insurance subscribers : providers , patient , employers, corporate hr

These are the parties are involved most in doing insurance fraud , among them we need to detect the fraud claims .

(4)DATA PRE-PROCESSING:

Mainly insurance claims participant involved in insurance subscriber and service , provider the data has 2 characteristics .

1. First table which contain the information of insurance subscriber provider .
2. Second table contains the details of subscriber id and other details .

We have merge these 2 table to get the global view of data .

4.1. GOAL SETTING

First of fall we have to decide the main fields where frauds occurs most in insurace sector .

1. Ambulance Services
2. Skilled Nursing Facilities
3. Laboratory Services
4. Psychiatric Services
5. Home Health Services
6. New or expanded benefits under the Balanced Budget Act of 1997.

These are points has been selected with help of domain expert from thi research paper .

4.2 DATA CLEANING

Data should be clean and in the format of able to perform the operation on it irrespective to the given input . step are varies on the how the data format and its shape

4.3 HANDLING MISSING VALUES

Sometimes in data there are lot of missing value and we are discussing about the healthcare sector . in this type of data missing value could be occur due to the irrelevance, inapplicability , omission , ignorance etc. the reason behind why we are carrying most due to operation that we want to perform statistical operation ,because in stats data should not contain the missing value to perform the stats operation of distribution or calculation .

These are two techniques used in this research paper to handle missing value .

1. hot-deck imputation: It fills in the missing values in an incomplete case using values from the most similar but complete cases of the same dataset.
2. regression imputation: A regression model is fitted for each variable with missing values with other variables without missing values as input variables(features)

4.4 DATA TRANSFORMATION

So now we have to convert the data in respective views to detects frauds. We have to merge the columns using the primary key and foreign key .these all features are dependent upon the type of fraud want to detect . for eg: if we want to detect the fraud of claim document provider , we need to calculate average claim amount per provider, average number of claims per provider, total % of top service codes are claimed by the provider etc for a specific time period. Now in the new dataset there will be a unique primary key corresponding to the fraud type to be detected(if Provider fraud needs to be detected, the table should have provider id as primary key).

FEATURE SELECTION

In this step our task is to add some artificial features with the help of domain expert .and during the process of checking correlation we also have to remove the redundant features, which are highly correlated . ideally we have kept at least 20 to 30 features with the help of domain expert.

DATA AUDINTING

In this step we have to performed some statistical analysis to get the rid of idea behind the data . now we will try to plot some distribution and plots using the probability and statistics .

5. Statistical modelling for fraud detection:

The statistical modeling are mainly two types of supervised and unsupervised modeling .

1. Supervised modeling : these types of statistical modeling we perform when we have labeled data (to detect whether claim are fraud or not).
2. Unsupervised modeling : these types of statistical modeling we perform when we have unlabeled data (we doesn't have the class-label of data)

5.1 Supervised methods

The popular supervised methods which are used for healthcare fraud detection are Decision trees and Neural Networks.

Decision Trees:

Decision trees can be used after doing the feature engineering. It has the ability to generate rules from the input feature and class label. It also has the ability to handle missing values if any. As Decision trees are interpretable it is very much useful for healthcare fraud detection. The only disadvantage is if the dimension of the train dataset is large it generates too many rules, thus tends to overfit. To get rid of that ensemble models like adaptive boosting is incorporated. Using ensemble models various classifiers are constructed sequentially with each focused on those training cases that have been misclassified by previous classifiers.

Neural network

Neural network is a techniques of deep learning , even it ca handle the non-linear data as well . but the problem with the neural network is to avoid the overfitting, its overfits very fast . so to avoid that issue we uses the early stopping.

Unsupervised machine learning

Unsupervised machine learning we used when we have unlabeled data . we want to check the behavior of data through clusters . k means , hierarchical clustering and dbscan are the some methods used in unsupervised machine learning

Performance matrix for supervised machine learning

We are dealing with the some healthcare fraud detection problem so we have to use the some performance matric to check the model performance .

Since we are dealing with the highly imbalanced dataset . so that ‘why accuracy cannot be used as matric to check the model performance . so we have to use some other metrics like f1-score and auc-roc score . f1-score we uses for to take the harmonic mean of precision and recall to check the balance performance . and we will also use auc-roc curve as matric to check the false positive rate(fpr) and false negative rate(fnr) . fnr and fpr should as less as possible . because in this problem model should not predict as if any claim is fraud then should not predict as legitimate and is claim is legitimate then is should not predict as fraud that ‘why fnr and fpr should as low as possible .

There are two ways of evaluating performance based on confusion matrix.

1. Cost based method: If the costs of FN and FP are explicitly specified, we can adopt this. FN cost means how much money we are losing if a fraudulent claim can not be detected. FP cost means if a legitimate claim is detected as fraudulent, how much needs to be spent for investigation.

2. Error based method: In that case ROC curve is commonly used, which plots TP against FP rates at different decision making thresholds of a classifier. Then the Area Under that curve (AUC) is calculated which indicates the discriminating power of the classifier. If AUC=1 the classifier works perfectly and if AUC=0.5 if performs randomly.
   Apart from AUC, F1 score can also be taken as performance metric.

Conlusion

So overall goal of out project to make a classification model to detect whether claim is fraud or not . and that model should be scalable , accurate , fast. Scalable refers to in the future model can also handle the high volume of data . accurate refers to model should predict that result as much as accurate and fast refers to model should be able to handle the data in real world data .

How these research paper is helpful for my case-study

1. Handling missing value : with the help this research I got the idea to interact with different technique handle the missing data . we have used two techniques to handle missing data hot deck imputation techniques and regression techniques .
2. Global point of view to handle problem : I got the global view as well how to approach the problem in real world data .
3. Handling imbalanced data: since this healthcare fraud detection problem is highly imbalanced dataset problem , so I also get the rid of problem how to handle the problem and approach that.
4. I also got the good rid of idea how to handle the situation where accuracy as a matric not works well . so which other else matric should we use .
5. I also got the brief overview that how confusion matric is play a big role to avoid the misclassification rate .

First Cut Approach

*** Explain in steps about how you want to approach this problem and the initial experiments that you want to do. ***(MINIMUM 200 words)*** ***

*** When you are doing the basic EDA and building the First Cut Approach you should not refer any blogs or papers ***

DATA CLEANING

I will do the data cleaning operation each steps to get the rid and use full of data from dependent and independent variables from all 4 csv files .

Provider dataset csv file

In this dataset I will check the distribution of data . since there is 90% data is belong to non-fraud class and only 10% providers are fraud . so this is highly imbalanced data set . so I have to use some techniques to handle the imbalanced dataset like oversampling or down sampling since this is big data so I think after reading several blog I should use up sampling with random sampling using sampling techniques like SMOTE sampling .i will check the random down sampling as well to get the rid which one is working better.

Inpatient and outpatient data

1. In this both csv files I have to check that for how many days patient was admitted in hospital .
2. For inpatient data column I will check the total number of days patient got admitted using the column date of admission and data of discharge I since we know that there is many missing values in dataset so I will handle that with the help of domain knowledge.

3. Same this I will also calculate this for outpatient data columns and will impute the missing values using domain knowledge given in previous blogs and research paper.
4. I will also check that total amount that reimbursed and total amount that deducted .
5. I will also check the patient age using patient date of birth and date of death .
6. I will use the count plot to check how many physician treated to patient like attending physician, operating physician and other physician because that is also could be suspicious if most patient getting treated by any particular physician .
7. I think we can add the column here total patient got treated by doing submission of columns attending physician , operating physician and other physician .
8. I will also use pdf and cdf to check the distribution of data of patient and outpatient data .
9. I will also checked the mean of total amount that reimburse and deducted for both inpatient and outpatient if there is big difference between fraud and non frauds provider that would be suspicious .
10. I will also add one more column like hospitalization status using file inpatient and outpatient . I will denote inpatient with 1 and outpatient with 0.

Merge the dataset

1. Now I will merge all 4 csv files by similar column wise using left and group by operation to get the whole or global view of dataset and brief idea.

Feature engineering

1. I will take the average of total amount reimbursed using this : total amount recived by patient / number of claims .
2. I will also calculate the total number of claim days using : total days of claims / number of claims
3. I will also check the total number of claims per patient .

4. I will also check that avg number of days patient got hospitalized using : total number of days patient admitted / number of times hospitalized .

Now I will do the merging of all the columns that we have created artificially just befrore using feature provider id by group by function (avg amount reimbursed, avg days for hospitalization, avg days for columns)

Now I will do feature selection techniques to detect the most important features random forest and I will also check most least important features . then I will try all the dataset like dataset with all features , dataset with important features and dataset with redundant features . obviously dataset with least important features will not performed well but I want to do this just for sake of doing .

Performance matric

1. F1-score
2. Auc-roc curse
3. Will check the FPA AND TPA to check misclassification rate using confusion matrix

Train and test splitting

1. Since we know that the configuration of out dataset 90:10. So this is highly imbalanced dataset , after reading several blogs and research paper I will use the over-sampling with smote techniques with different combination of dataset like 90:10 , 80:20, 60:40 , 50:50 with hyperparameter search and will see the model performance . according will decide the model best split.

My model selection approach will be

1. First I will go with simple linear models like logistic and svm model and knn model . with 10 times k fold cross validation techniques with different combination of data and will check the model performance .
2. After that I will also use the decision tree model to check the performance of tree based model .
3. Now finally for getting better performance I will use the ensemble techniques with hyperparameter search .
4. If tree based model not performed well in ensemble model then I will go for custom implementation of ensemble model using different machine learning models .

Notes when you build your final notebook:

1. You should not train any model either it can be a ML model or DL model or Countvectorizer or even simple StandardScalar
2. You should not read train data files
3. The function1 takes only one argument "X" (a single data points i.e 1*d feature) and the inside the function you will preprocess data point similar to the process you did while you featurize your train data
    a. Ex: consider you are doing taxi demand prediction case study (problem definition: given a time and location predict the number of pickups that can happen)
    b. so in your final notebook, you need to pass only those two values
    c. def final(X):
        preprocess data i.e data cleaning, filling missing values etc
        compute features based on this X

use pre trained model
return predicted outputs
final([time, location])

d. in the instructions, we have mentioned two functions one with original values and one without it
e. final([time, location]) # in this function you need to return the predictions, no need to compute the metric
f. final(set of [time, location] values, corresponding Y values) # when you pass the Y values, we can compute the error metric(Y, y_predict)

4. After you have preprocessed the data point you will featurize it, with the help of trained vectorizers or methods you have followed for your train data
5. Assume this function is like you are productionizing the best model you have built, you need to measure the time for predicting and report the time. Make sure you keep the time as low as possible
6. Check this live session: https://www.appliedaicourse.com/lecture/11/applied-machine-learning-online-course/4148/hands-on-live-session-deploy-an-ml-model-using-apis-on-aws/5/module-5-feature-engineering-productionization-and-deployment-of-ml-models