ACTL3142

US Medicare Data Claims Provider Fraud Modelling

Neel Patel

z5205344

Table of Contents

Executive Summary

An analysis into historical datasets can be conducted to learn patterns to predict future outcomes. In this report, we conduct an analysis into the US Medicare Claims Data, to determine factors that can be used to develop statistical models to be able to predict future fraudulent insurance provider. In the analysis we find issues with the data and have to clean it up and balance the data to be able to make better models. Ten different statistical classification models were used to predict with great accuracy, however most had a low true-positive classification rate. Out of the ten models, the we observed that the boosting tree model performs the best due as has a good accuracy rate and a better true-positive rate compared to other models. We also learn that fraudulent providers have similar characteristics to one another due to the boosted tree model.

Data Clean up and Parameter Selection

Initial Data Preparation

The training data for modelling is separated into two separate datasets, the Medicare data per claim that involves all the variables of interest to conduct the study and the providers classified as fraudulent or not. These datasets were merged by ProviderID to make one dataset, so each claim was denoted to have used a provider that is fraudulent or not. In the merging of the data the dates within the claim dataset corrupted and wasn't being read as a date in R. This was fixed by creating a function that read the dates as a string and splitting the string into year, month and days vector. These vectors were converting back into numeric and pasted together and converted back into a date format.

Also, studying the different indicators it was found that the data was improperly entered, the chronic disease indicators were indicated as 1 for if the beneficiary of the claim had the disease and 2 if they didn't have the disease. Similarly, gender was also denoted as 1 and 2 and the renal disease was indicated as Y if the beneficiary had it and 0 if they didn't. These were cleaned up to the binary inputs of 1 indicate existence of the disease and 0 for not and for consistency gender 2 is converted to 0. Another issue with the data is that number of variables have a great number of NAs, so to produce a more interpretable and better model, the physician and claim procedure and diagnosis indicators were removed from the analysis.

Model predictor preparations

The model predictors are prepared by aggregating the variables by ProviderID, (full list of variables used in modelling in Appendix A). However, we see that there is a massive imbalance in the number of fraudulent and non-fraudulent providers. From figure 1, we can observe there is a much greater percentage of non-fraudulent providers compared to fraudulent. Using this dataset

for the model to predict fraudulent providers will be biased towards non-fraudulent, since the training data will have a much greater number of non-fraudulent providers.

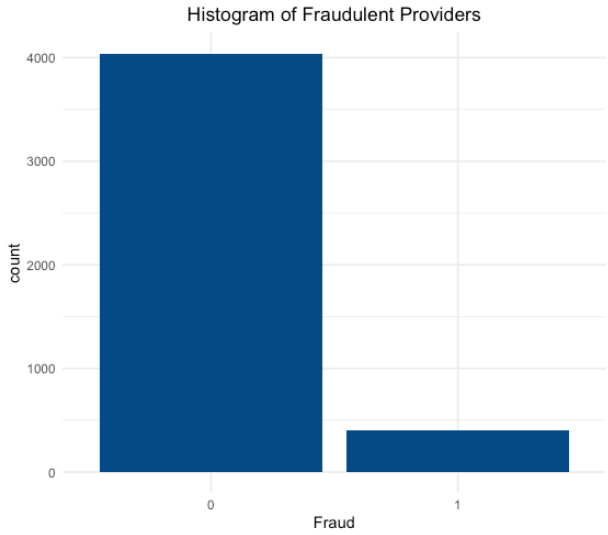

Figure 1: Proportion of fraudulent providers to non-fraudulent providers

To overcome this problem, we can under balance the data, meaning we only use the same number of non-fraudulent data points as there are fraudulent data set or overbalance where you repeat the fraudulent datapoints. This will provide us a less biased model if the unbalanced dataset is used for modelling.

Model prediction

Within this analysis to classify the different providers as fraudulent, ten different classification models were used. The balanced data was used to fit the models and k-fold cross validation was employed to determine the tuning parameters of the models to provide the most efficient fit.

Model	**Accuracy**	**Positive True Rate**
Logistic Regression	82.26 %	70.09%
K Nearest Neighbours	79.17%	71.79%
Logistic Ridge Regression	85.49%	57.26 %
Logistic Lasso Regression	84.74%	57.26 %
Linear Discrimination Analysis	83.76%	64.96 %
Classification Tree model	88.12%	62.39%
Random Forest Tree model	82.78%	82.05 %
Bagging Tree Model	92.71%	49.57%
Boosting Tree Model	83.61%	83.76%
Support Vector Machine	27.52%	62.39%

Table 1: The accuracy and positive true rates for each model

From the modelling process it was found that the Boosting tree model was the most optimal model to predict the fraudulent model. This is because it has the highest true positive rate out of all the models. Some models have the accuracy in the 90% area however, it was found that the great proportion of that accuracy was its ability to predict non-fraudulent providers far better than it was able to predict the fraudulent. This can be put down due to the data being unbalanced and hence this was to be predicted that model which are accurate will be due to its ability to predict non-fraudulent providers much better compared to fraudulent.

The boosting tree model fits a decision tree uses current residuals rather than the outcome as the response. Then new decision tree branches is fitted in order to update the residuals, each of these new trees can be smaller and with few terminal nodes. There is another shrinking parameter that slows down the process even further. This way, the model is able to perform better than other models and avoids overfitting which is seen in other models as they have a low true positive rate. The model's tuning parameters are found using 10-fold cross validation, so the hyperparameters are chosen such that the model has the lowest training error.

	Reference	
Prediction	**0**	**1**
0	1014	19
1	199	98

Table 2: Confusion Matrix for Boosted Tree Model

From the confusion matrix we can see that out of the 117 fraudulent providers, it predicted 98 of them correctly. However, it also predicted 199 other providers as fraudulent that aren't, which is the reason for its drop-in accuracy.

Fraudulent Providers Characteristics

We can see from the boosting tree model which of the variables has the most influence in predicting a provider as fraudulent or not. From figure 2, we can see that the total amount of claims, average DeductibleAmtPaid and average InsClaimAmtReimbursed have the greatest influence on the model prediction.

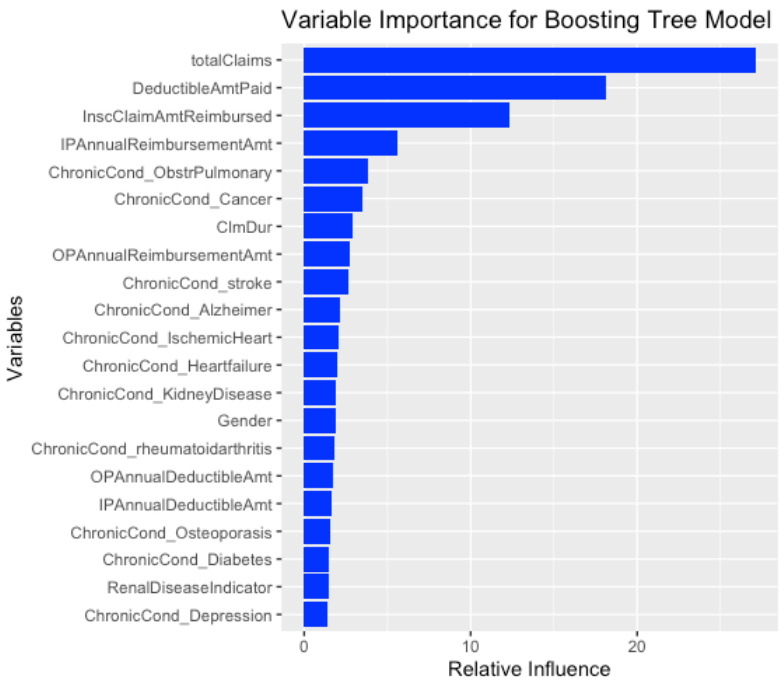

Figure 2: Relative influence of each variable in the boosted tree model

This can also be seen from the logistic regression, conducting the hypothesis test where the null hypothesis is $\beta_I = 0$, and we see that for a significance level of 5%, InsClaimAmtReimbursed, DeductibleAmtPaid, IPAnnualReimbursementAmt, OPAnnualReimbursementAmt,

ChronicCond_ObstrPulmonary , ChronicCond_stroke , ClmDur and totalClaims reject the null hypothesis suggesting these variables influence the models ability to predict.

We see that OPAnnualReimbursementAmt, ChronicCond_ObstrPulmonary , ChronicCond_stroke , ClmDur are also influential for the boosting tree model however the three main totalClaims, DeductibleAmtPaid and InsClaimAmtReimbursed for each provider greater influence. There is a common trend for each of these variables, from figures 3,4 and 5 we can see that fraudulent providers tend to have a higher value for each of the variables.

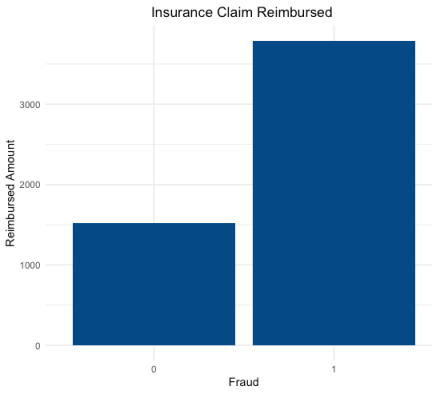

Figure 3: Average insurance claim amount reimbursed

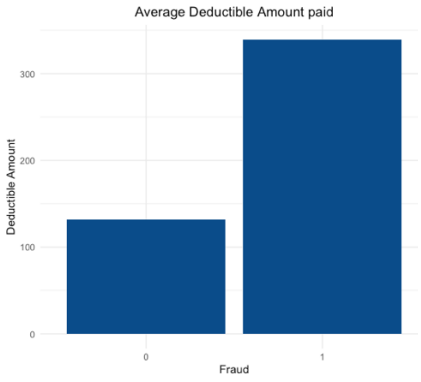

Figure 4: Average Deductible amount

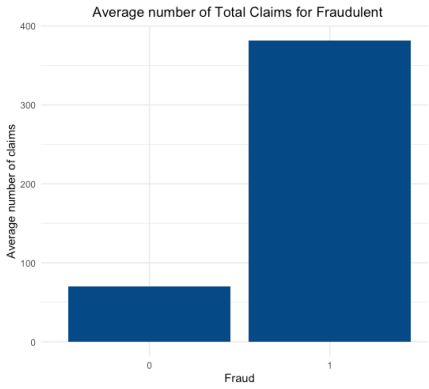

Figure 5: Average number of claims for providers

Limitation

There are many limitations to the analysis, mainly being a big trade-off between getting accurate model predictors and time, the dataset being unbalanced, and the more accurate models have greater complexity. To conduct such analysis, the r-code written takes a long amount of time to make accurate models with a large dataset. However, there are functions that reduce the time taken but at a trade-off with accuracy of the model. Also, having the dataset being unbalanced, transforming the training data to be balanced by either reducing the number of non-fraudulent providers in training set or increasing the fraudulent by repeating the fraudulent cases in the training set causing some data to be left out or repeated. This results in the model being underfitted or overfitted which could become an issue for future prediction accuracy. Lastly, the models such as boosted tree diagrams are very complex statistical algorithms which require higher knowledge and aren't easy to provide simplified explanations for. Thus, more accurate models tend to be more complex and hence are less interpretable.

Conclusion

We conclude from the analysis conducted that the boosting tree model is the best classification model to predict if insurance providers are fraudulent or not. It performs much better than other predictors due to its greater true-positive rate, meaning it is able to predict fraudulent providers more time than not. However, as a result of the imbalance in the data, the model is overfitted and results in it predicting more fraudulent providers than there are, hence it has a lower accuracy rate. But from the boosted tree model we see that the three variables totalClaims, DeductibleAmtPaid and InsClaimAmtReimbursed are great indicators for predicting if a provider is fradulent or not.

Appendix

A: List of predictors

Variables	**Description**
InscClaimAmtReimbursed	Average Insurance Claim Reimbursed
DeductibleAmtPaid	Average Deductible amount paid
IPAnnualReimbursementAmt	Average Inpatient Annual Reimbursed
IPAnnualDeductibleAmt	Average Inpatient Annual Deductible amount
OPAnnualDeductibleAmt	Average Outpatient Annual Deductible amount
OPAnnualReimbursementAmt	Average Outpatient Annual Reimbursed
Gender	Proportion of claims that are gender 1
RenalDiseaseIndicator	Proportion of claims that have Renal Disease
ChronicCond_Alzheimer	Proportion of claims that have Alzheimer
ChronicCond_Cancer	Proportion of claims that have Cancer
ChronicCond_Heartfailure	Proportion of claims that have Heart failure
ChronicCond_KidneyDisease	Proportion of claims that have Kidney Disease
ChronicCond_Depression	Proportion of claims that have Depression
ChronicCond_stroke	Proportion of claims that have Stroke
ChronicCond_IschemicHeart	Proportion of claims that have Ischemic Heart
ChronicCond_Diabetes	Proportion of claims that have Diabetes
ChronicCond_ObstrPulmonary	Proportion of claims that have Obstructed Pulmonary
ChronicCond_Osteoporasis	Proportion of claims that have Osteoporasis
ChronicCond_rheumatoidarthritis	Proportion of claims that have Rheumatoidarthritis
ClmDur	Total duration of the claim
totalClaims	Total amount of claims for provider

Table 3: Variables description

B: Technical Models

Logistic Regression

```
Call:
glm(formula = Fraud ~ ., family = "binomial", data = x_trn_over_df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.8637  -0.6796  -0.2317   0.7834   1.9730

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -1.741e+00  3.123e-01  -5.576 2.46e-08 ***
InscClaimAmtReimbursed         1.572e-04  3.255e-05   4.831 1.36e-06 ***
DeductibleAmtPaid              1.286e-03  3.159e-04   4.070 4.71e-05 ***
IPAnnualReimbursementAmt       2.350e-05  1.143e-05   2.056   0.0398 *
IPAnnualDeductibleAmt          3.285e-05  1.133e-04   0.290   0.7718
OPAnnualDeductibleAmt         -3.232e-04  2.255e-04  -1.433   0.1518
OPAnnualReimbursementAmt       1.264e-04  5.709e-05   2.215   0.0268 *
Gender                         1.275e-03  2.527e-01   0.005   0.9960
RenalDiseaseIndicator          4.878e-01  3.506e-01   1.391   0.1641
ChronicCond_Alzheimer         -6.719e-01  2.736e-01  -2.456   0.0141 *
ChronicCond_Cancer            -2.662e-01  3.483e-01  -0.764   0.4447
ChronicCond_Heartfailure       5.833e-02  2.883e-01   0.202   0.8397
ChronicCond_KidneyDisease      4.260e-01  3.060e-01   1.392   0.1638
ChronicCond_Depression        -7.328e-04  2.604e-01  -0.003   0.9978
ChronicCond_stroke            -4.052e-01  4.000e-01  -1.013   0.3111
ChronicCond_IschemicHeart     -4.593e-01  3.339e-01  -1.376   0.1690
ChronicCond_Diabetes          -3.183e-01  3.081e-01  -1.033   0.3015
ChronicCond_ObstrPulmonary     6.595e-01  2.910e-01   2.266   0.0234 *
ChronicCond_Osteoporasis      -1.467e-01  2.730e-01  -0.537   0.5911
ChronicCond_rheumatoidarthritis -2.812e-03  2.712e-01  -0.010   0.9917
ClmDur                         5.859e-02  2.878e-02   2.036   0.0418 *
totalClaims                    7.354e-03  2.872e-04  25.606  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7824.2  on 5643  degrees of freedom
Residual deviance: 5323.7  on 5622  degrees of freedom
AIC: 5367.7

Number of Fisher Scoring iterations: 6
```

	Reference	
Prediction	**0**	**1**
0	1012	35
1	201	82

Table 4: Confusion Matrix for Logistic Regression

K nearest neighbours

	Reference	
Prediction	**0**	**1**
0	1012	35
1	201	82

Table 5: Confusion Matrix for K nearest neighbours

The tuning parameter K was found using cross validations

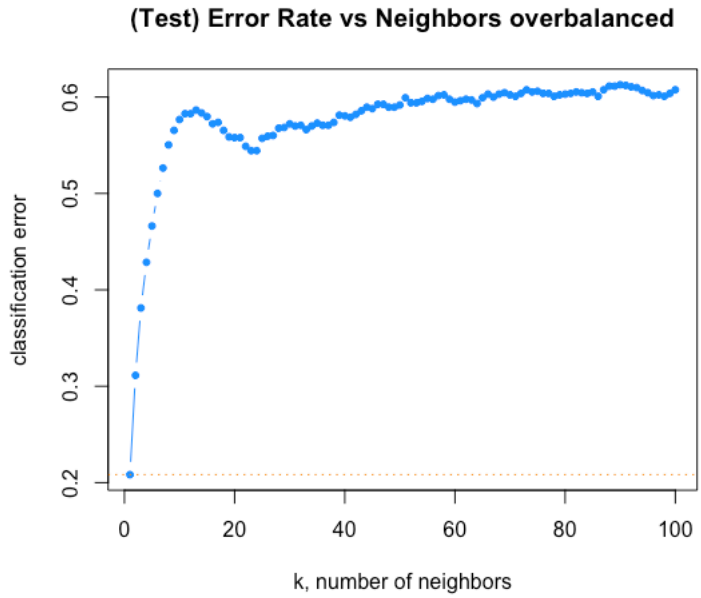

Figure 6: Test error rate for each k using cross validation

Logistic Ridge

	Reference	
Prediction	**0**	**1**
0	1070	50
1	143	67

Table 6: Confusion Matrix for Logistic Ridge

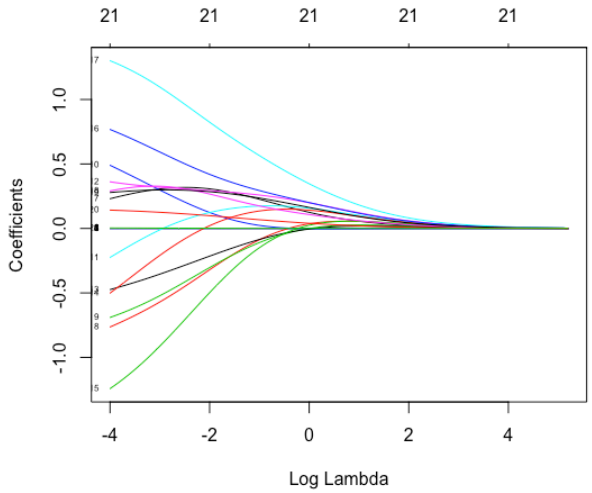

Figure 7: Coefficients for logistic regression based on lambda

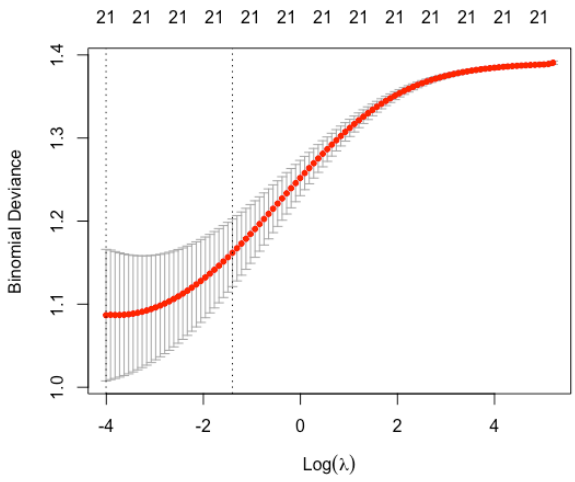

Figure 8: Cross validation to choose hyperparameter

Logistic Lasso

	Reference	
Prediction	**0**	**1**
0	1060	50
1	153	67

Table 7: Confusion Matrix for Logistic Lasso

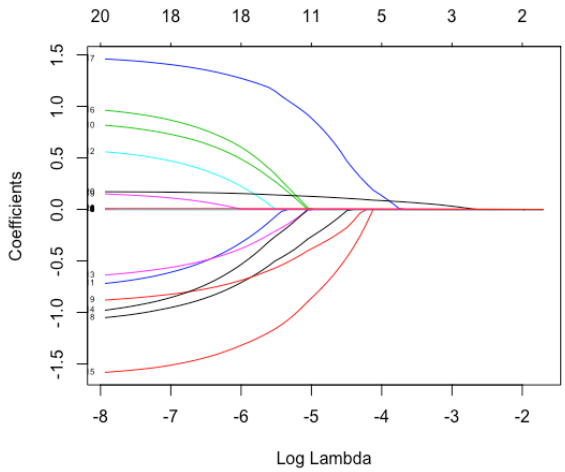

Figure 9: Coefficients for logistic regression based on lambda

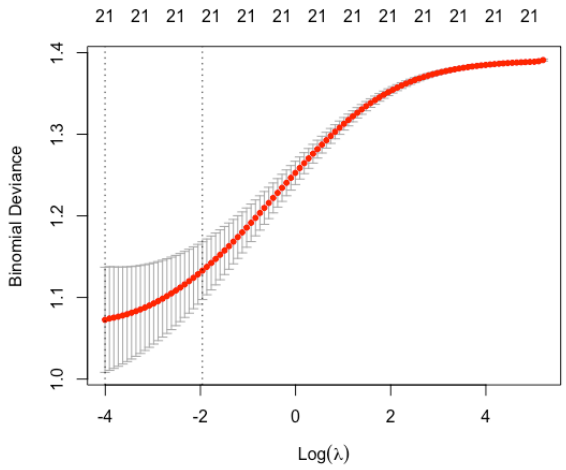

Figure 10: Cross validation to choose hyperparameter

Linear Discrimination Analysis

	Reference	
Prediction	**0**	**1**
0	1038	41
1	175	76

Table 8: Confusion Matrix for Linear Discrimination Analysis

Classification Tree Model

	Reference	
Prediction	**0**	**1**
0	1099	44
1	114	73

Table 9: Confusion Matrix for Classification Tree Model

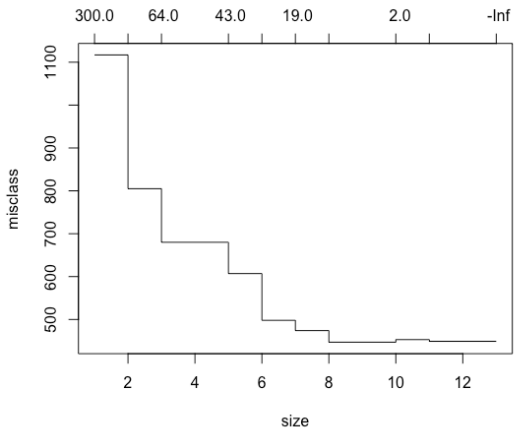

Figure 11: Cross validation to choose hyperparameter

Bagged Tree Model

	Reference	
Prediction	**0**	**1**
0	1175	59
1	38	58

Table 10: Confusion Matrix for Bagged Tree Model

Random Forest Tree Model

	Reference	
Prediction	**0**	**1**
0	1005	21
1	208	96

Table 11: Confusion Matrix for random Forest Tree Model

Boosting Tree Model

	Reference	
Prediction	**0**	**1**
0	1005	21
1	208	96

Table 12: Confusion Matrix for Boosting Tree Model

Support Vector Model

	Reference	
Prediction	**0**	**1**
0	293	44
1	920	73

Table 13: Confusion Matrix for Support Vector Model