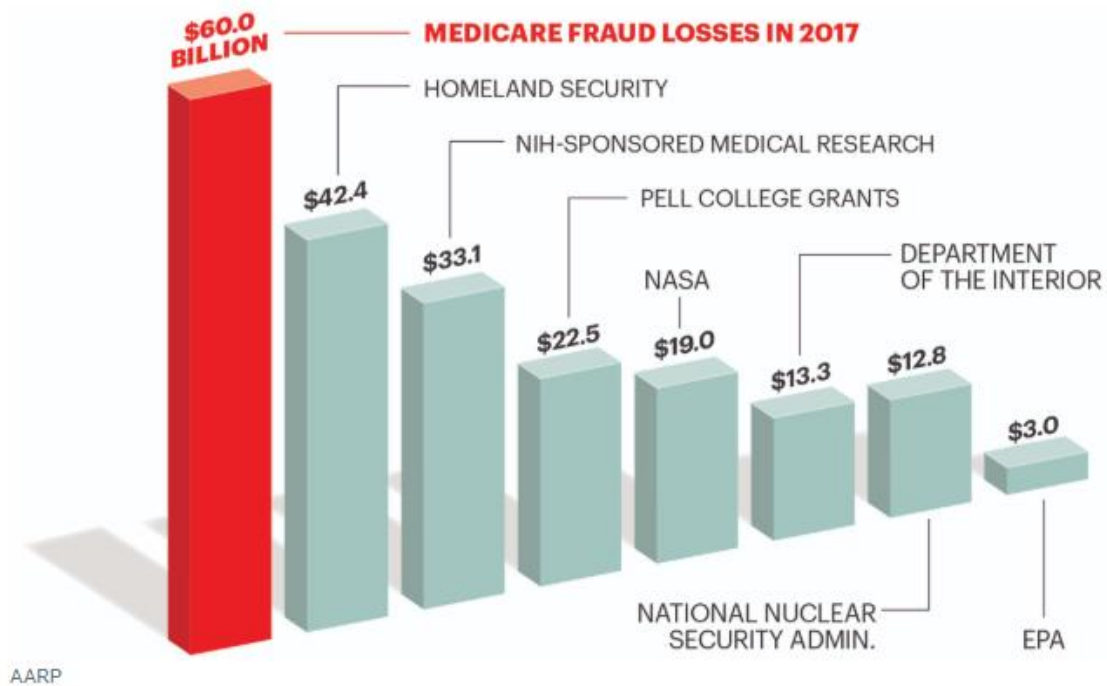


# Analysis of US Medicare Fraud

ActI3142 Assignment – Part B

Z5194905



<sup>1</sup> <https://www.aarp.org/money/scams-fraud/info-2018/medicare-scams-fraud-identity-theft.html>

### Executive Summary:

For any model there are two main factors to consider— interpretability and predictive power. Throughout this report, Area Under the ROC Curve (AUC) will be used as the rating criteria for predictive power as it removes some of the bias of unbalanced data. This report will generally consider predictive power a higher priority than interpretability to create a better fraud detection system.

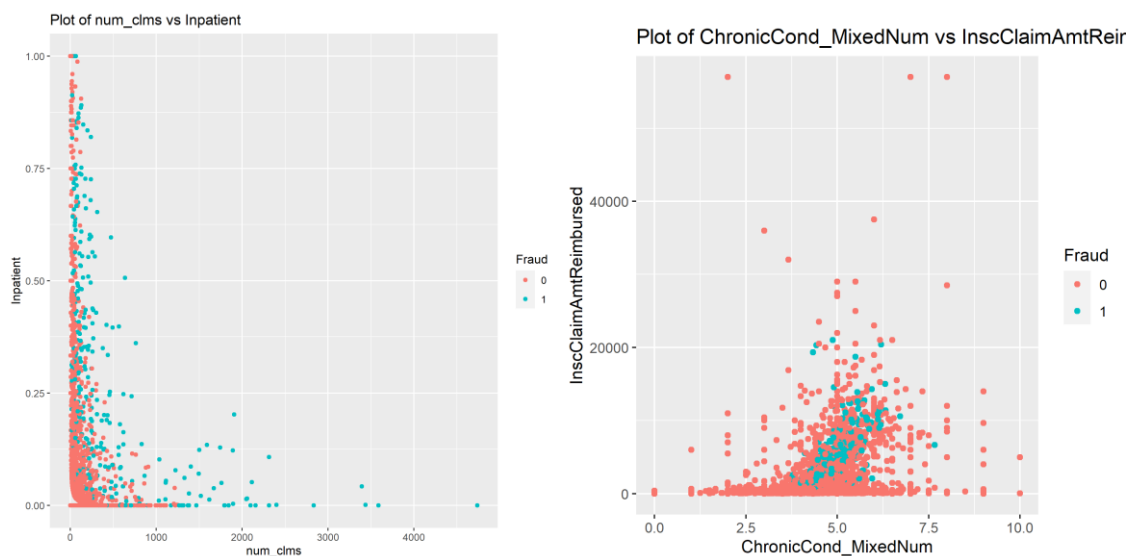
This report has aggregated the US Medicare data by providers and fit a series of classification models to identify future fraudulent claims. These models include logistic regression, discriminant analysis, KNN, classification trees and support vector models. The report summarises the findings of each model by describing common characteristics amongst fraudulent providers. These characteristics include; high reimbursements, many claims, county location and proportion of inpatients. Flexible models were better in almost all cases because of the large data set. More detailed analysis and the methodology of fitting these models can also be found in the appendix.

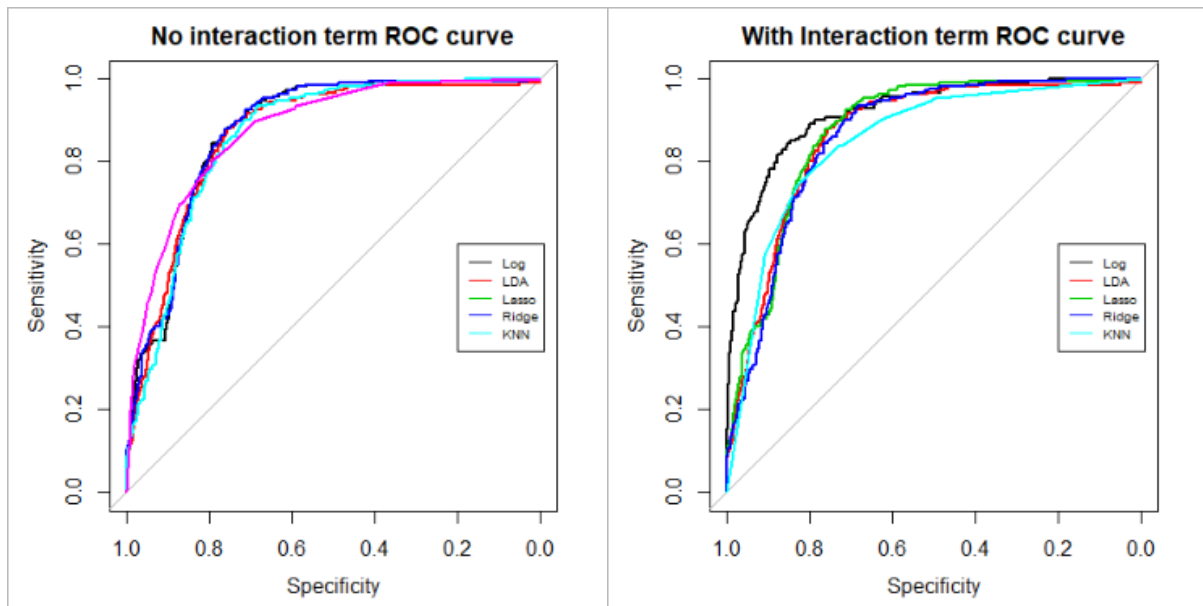
From the analysis, a Random Forest was the best model, and, while less interpretable than other methods such as logistic regression, prediction power was prioritized to reduce future fraud costs.

### Model Selection:

#### Inclusion of interaction terms

From the below figures, it was clear that some variable interactions could be indicative of fraudulent behavior among providers. Introducing the below predictors as interaction variables improved the predictive behavior of all models, suggesting there is an underlying relationship for fraudulent providers.



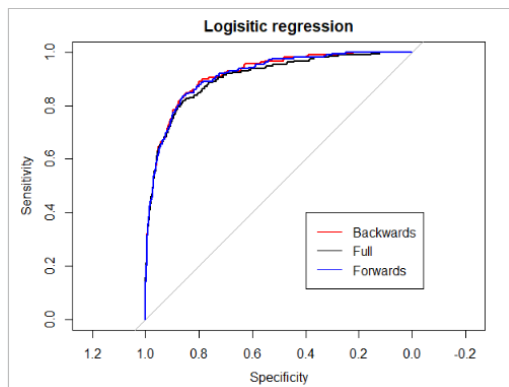


Therefore, these interaction terms will be considered for all following regression analysis.

### Logistic regression:

#### Stepwise selection

Since the dimensionality of the data is too large for best subset selection, a model was chosen from forward stepwise selection and backward stepwise selection. Although backward stepwise selection had the larger AIC of the two, (813.5 compared to 812.5), it was chosen since it has slightly higher predictive power.



#### Provider Characteristics from logistic regression

More Fraudulent	Less fraudulent
<ul style="list-style-type: none"> <li>High reimbursement amounts</li> <li>If the most common beneficiary race is "3"</li> <li>More claims per beneficiary</li> </ul>	<ul style="list-style-type: none"> <li>Long hospital stays</li> <li>Higher proportion of beneficiaries with Chronic kidney disease</li> </ul>

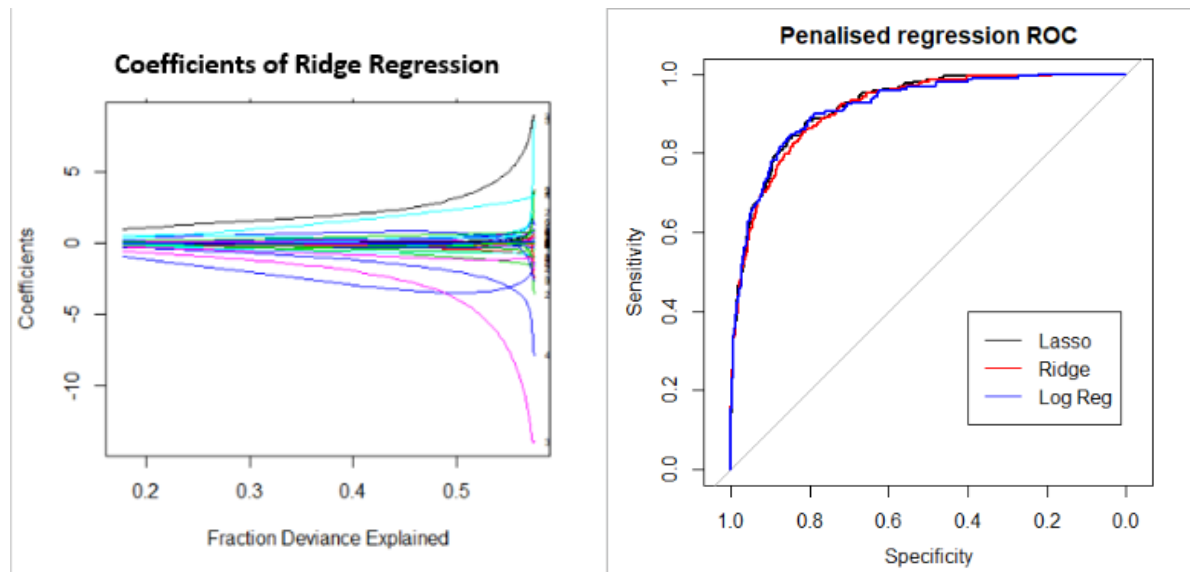
The model reveals that providers who have higher claims and have more inpatients are certainly much more likely to be fraudulent. For future risk management strategies, the table above shows other significant factors which indicate fraud according to the logistic model.

Claims from states 33, 39, 45, 49 and 31 were shown to have a significant effect on the probability of fraud and so stricter surveillance may need to be introduced in these areas. Other factors such as high rates of obstructive pulmonary disease and missing diagnosis codes were shown to be indicators of honest providers, however, not at a statistically significant level.

This model is simple and very interpretable and a surprisingly good fit, as seen from the ROC curves above.

Ridge regression:

The ridge model loses some interpretability by including all predictors in modelling. The plot below highlights that most coefficients have values close to 0.



The  $\lambda$  chosen from CV was 0.00012, indicating that there is a relatively small parameter penalty. Since this penalty value of predictors is so low, the ridge model is practically reduced back to a normal OLS optimization problem. Therefore, we can conclude that a large set of predictors are useful.

Lasso:

Lasso has similar predictive power to the ridge model, but it is much more interpretable by reducing the number of predictors. From CV, the best lambda was 0.00248 and the best number of predictors were 20 predictors (can be found in the appendix). We can see from the plot that small  $\lambda$ 's are generally better and the best value is  $\approx 0$ . This suggests that most predictors are good indicators of fraud as few get eliminated by the penalty, therefore, it is unsurprising that its ROC curve is similar to logistic regression.

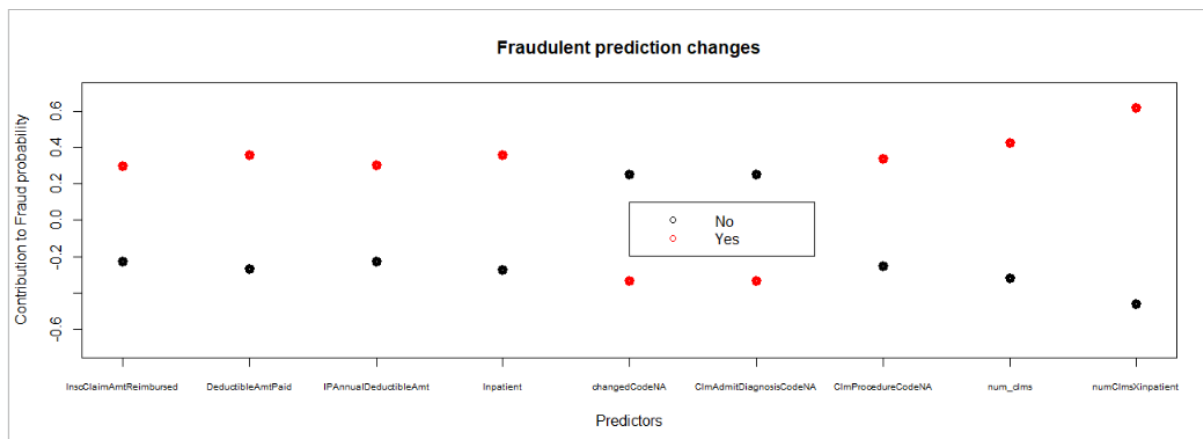
Summary logistic regression

Ridge regression has similar predictive power to backward stepwise logistic regression and since it is much less interpretable it is discarded as the worst model of the three. Lasso has a slightly better predictive power than backward selection (AUC is 92.55% compared to 91.85 %) but it is a more complex model (20 vs 16 predictors). Therefore, backward selection is chosen as the best logistic model.

## Discriminant Analysis

Linear discriminant Analysis:

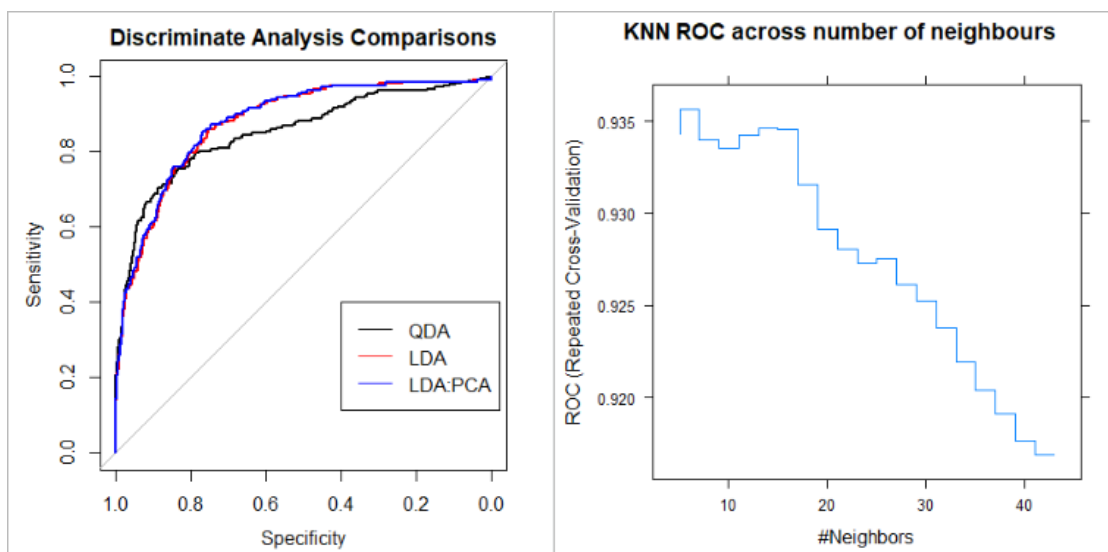
LDA is not an improvement over the logistic regression methods, however, it does give insight into the behaviour of fraudulent providers. The graph below shows the effects of the 10 most powerful predictors according to LDA.



The larger the difference, the better the predictor. And so, providers with large number of claims and high proportion of inpatients were shown to be the best indicator of fraudulent behaviour followed by deductible amount paid. Interestingly, providers who had higher missing admit diagnosis codes may be an indicator that the provider is honest.

#### LDA : PCA

To reduce computational complexity whilst retaining modelling power, principle component analysis (PCA) was used to reduce dimensionality. This loses nearly all the interpretability of the model, however sometimes it can identify trends in the data based on predictor correlation. In this case, PCA LDA was a worse predictor (with an AUC of 88.02%, compared to 88.27%).



#### Quadratic discriminant Analysis:

QDA could only be used with PCA since there were collinear predictors. It was not a strong model in terms of predictive power or interpretability with an AUC of 85.18%

#### Summary Discriminant analysis:

Normal LDA is the best of these models as it remains interpretable.

#### AUC for Discriminant Analysis models

	LDA (%)	LDA: PCA	QDA
AUC (%)	87.68	86.7	85.18

## KNN

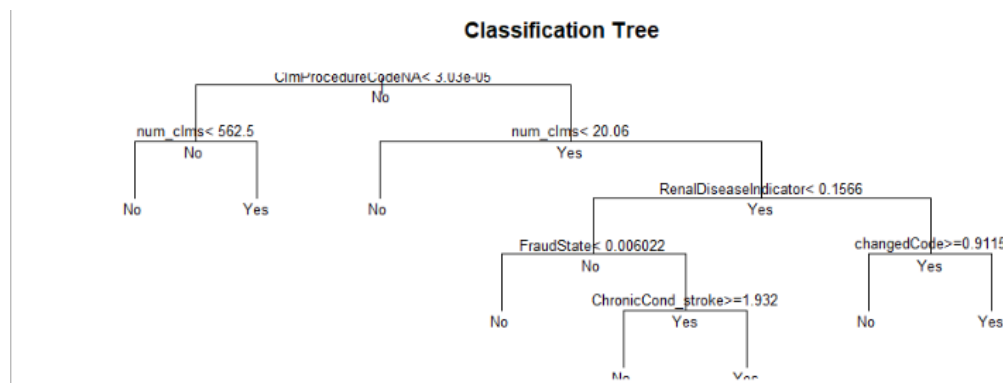
K-Nearest-Neighbors was not a particularly good model with an AUC of 85.26%

7 neighbors were the optimal number to be considered and, clearly from the graph above, a more flexible model was preferred.

## Tree based methods

### Classification Trees:

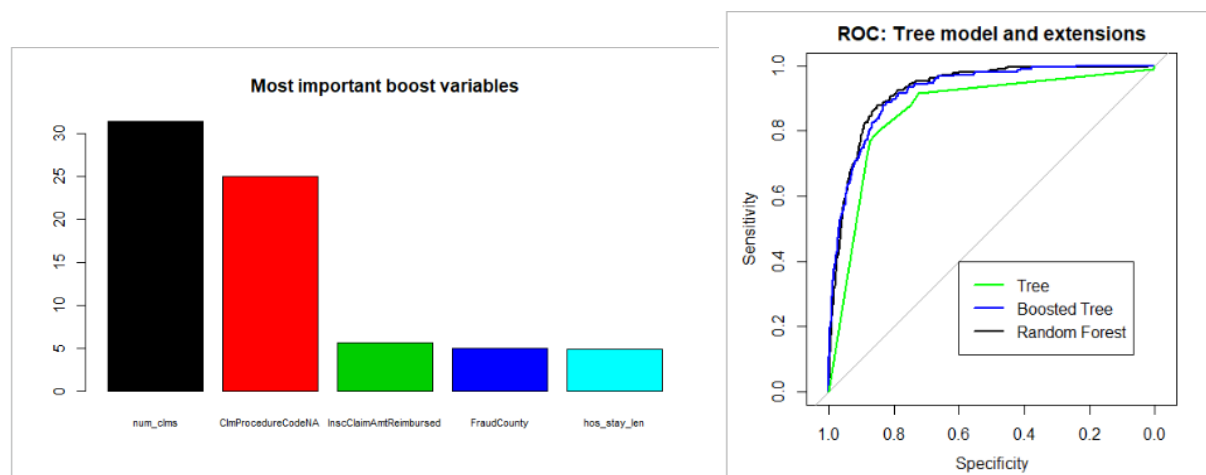
Classification trees were a good combination of interpretability and predictive power. The tree is shown below.



One of the advantages of classification trees is their interpretability, each provider can be assessed through the simple tree branching process. However, due to this simplicity they miss some of the trends in the data. Again, for risk managing strategies, providers with large number of claims should be held to a higher scrutiny as all providers with less than 20 claims are deemed honest. Similarly, if claim procedure code records are complete, they are much less likely to be fraudulent. Not only is this tree an intuitive blueprint for identifying potential fraudulent providers, it also has low test errors with an AUC of 86.47%

### Boosted Trees

Boosting trees reduces interpretability, however they gain much more predictive power with an AUC of 92.21%.



We see that the best indicators of fraud are a provider's number of claims, proportion of missing procedure codes, insurance claim amount, hospital stay length and whether the county with the most claims was either 160, 130, 590, 620 or 170.

### Random Forest

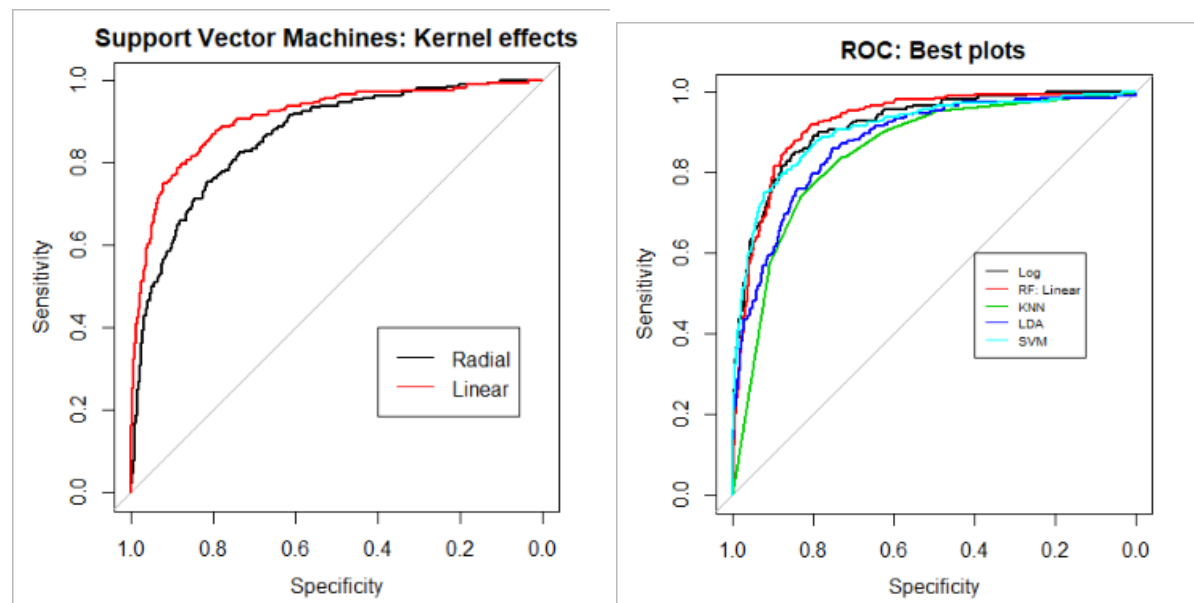
The last tree extension that will be used to model is random forest. It again improves on the boosted model, slightly with an AUC of 92.55%. From the ROC curve above, we can see that the random forest model is better than any other tree model at nearly all probability thresholds.

### Tree models summary:

Although less interpretable than the tree model, the random forest can predict fraudulent providers much more accurately and so it is selected as the best tree extension.

### Support Vector Machines:

Both radial and linear SVM's were fit to the data and each performed well. Whilst the radial had strong predictive properties the linear SVM is a better predictor at every probability threshold level. The C value that minimized CV error for the linear SVM was 1.684.



### Area under ROC curve of best models

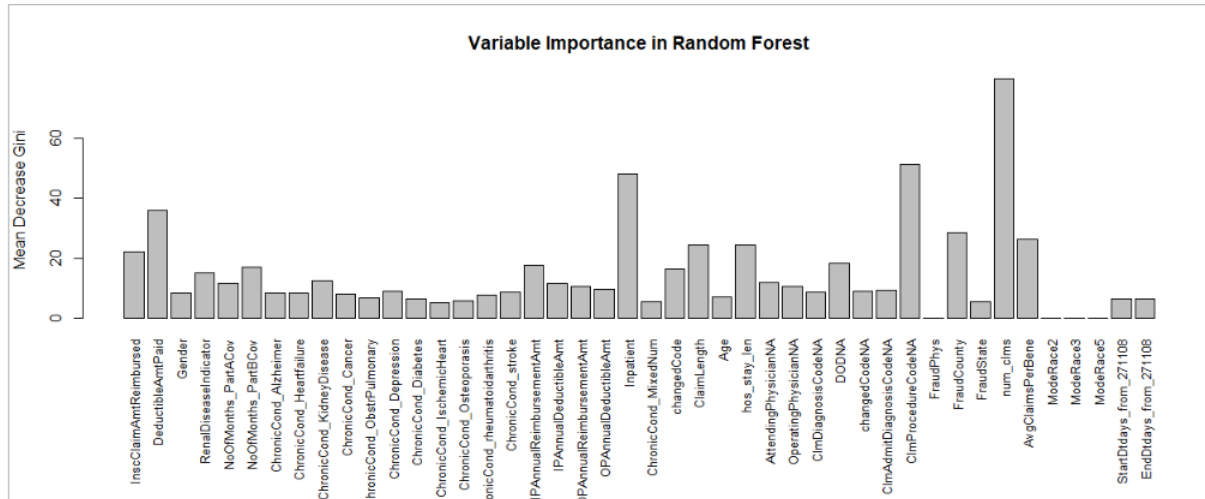
	AUC (%)
Log	91.85
Random Forest	92.55
KNN	85.26
LDA	87.68
SVM	90.89

### Determining best model

Random Forest is chosen as the best predictive model to classify fraudulent providers. Random forest has the highest AUC and a great ROC curve.

## Random Forest Model – in depth

This Random Forest fits many different trees to the data. Then each tree “votes” on whether an inputted claim is fraudulent or not, and the majority decision becomes the estimate. For this model, it has the lowest CV error with 500 trees and the number of variables tested at each branch was 7. This model is too complex to show each separate tree structure, however, we can represent the important of each predictor shown below.



We see again that number of claims, inpatient status and missing claim procedure codes are good predictors. The deductible amount paid, claim amount reimbursed and average claims per beneficiary also are significant. This also shows that providers that have more claims in counties 160, 130, 590, 620 and 170 should also be monitored more closely.

Ultimately, Random Forest is more difficult to interpret than other modelling techniques such as logistic regression. But accuracy was prioritized over interpretability to produce the best fraud detection system.

### Recommendation for further analysis

For future modelling, more variables should be created to identify any other missed relationships, for example, a generalized linear model of the expected cost of a claim. More interaction effects between variables should also be incorporated into the models.

Imputation of missing values should be done by KNN rather than using column means, however, the computational complexity of this task was out of the range of this report.

More oversampling and under sampling techniques, such as ROSE, CBO and ADA-SYN should be considered to alleviate the unbalanced data issue.

### Conclusion

It is clear from the different models that a provider’s number of claims and proportions of inpatients are indicative of fraudulent behavior. Other factors such as missing procedure codes, claim amount reimbursed, and the county of claim were also good indicators. Random Forest had the lowest test errors and so this model should be used for future fraud detection.

## Appendices:

### Appendix A – Assumptions

- If there were modal ties for most common beneficiary race, the first entry was taken

### Appendix B – Data Preparation

The data had several issues that made modelling difficult.

#### Imputation of missing data

Missing data can sometimes be informative, for example, a fraudulent provider may be more likely to not enter an individual's diagnosis. For this reason, new NA indicator variables were created for each column.

Claim procedure codes 5 and 6 were completely missing and were dropped from the dataset and missing values from other columns were then imputed by the average of the column so bias was not introduced.

#### Variable creation

Some new variables were created which would be helpful for later analysis including –

1. Number of claims from an individual
2. Age
3. Claim length
4. Number of chronic conditions
5. Whether diagnosis code changed
6. Hospital stay length

The data was then aggregated by provider and a final new variable added

7. Average claims per beneficiary

During this process, some variables could not be averaged such as race, state, county or physicians. For each provider, the modal race was recorded and whether a provider came from one of the top 5 fraudulent physicians. These columns were also created for states and counties.

#### The Unbalanced Data Problem

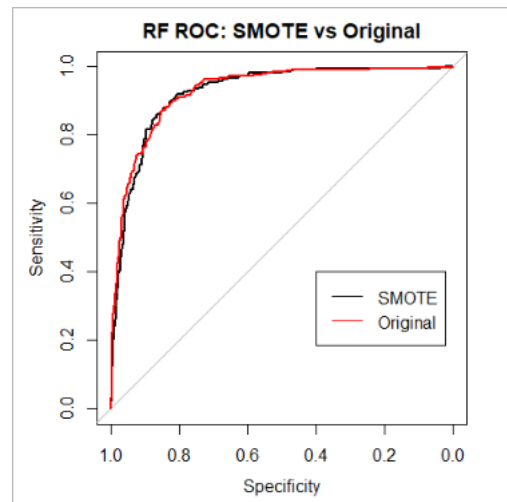
There are 4036 non-fraudulent providers and only 400 who are fraudulent. This introduces a problem for selecting a probability threshold for any model. If it is selected by minimizing the total error, the majority class is prioritized as all providers can be deemed innocent whilst total error is only 9.9%. Therefore, synthetic datapoints are introduced to the training dataset prior to performing CV through the SMOTE algorithm. This increases the number of fraudulent providers in the training data as shown below-

	<b>Not Fraudulent</b>	<b>Fraudulent</b>
<b>Original training data</b>	2030	188
<b>SMOTE training data</b>	752	564

Although some data points have been removed, this rebalancing allows CV (which minimizes error) produce more appropriate output. This increase in performance can be shown empirically by the slightly better ROC curve in random forest.

SMOTE is used here to increase the minority sample size, however, the SMOTE algorithm is not perfect, it has reduced the number of data points in our training set and can lead to over-generalization and variance.<sup>2</sup> SMOTE will have better effects on SVM and random forest.<sup>3</sup>

Finally, before modelling, the data was split into testing and training partitions.



## Appendix C – Modelling

The basis for modelling was the caret package since it unifies the conventions of its dependent packages allowing for neater code. Most of the caret results were verified first from the base package used.

The ROC curves and AUC values came from the package pROC.

### Logisitic regression

The full and empty logistic models were created using the base `r glm` function. Forward and backwards regression was then performed using the base `r “step”` function. The best backwards step-wise selection model was the following –

Coefficients:

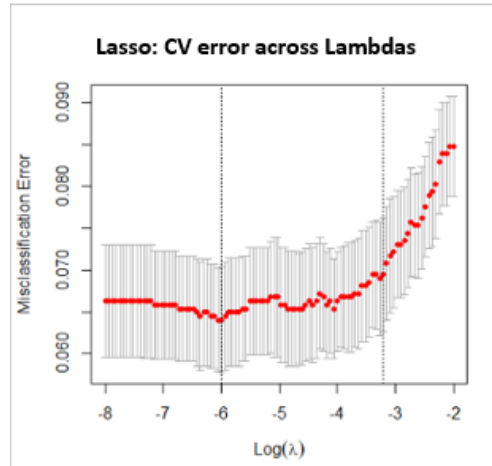
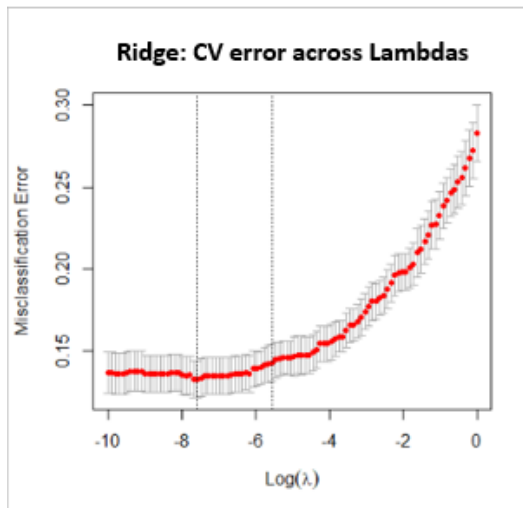
	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-3.983e+00	3.733e-01	-10.669	< 2e-16 ***
InscClaimAmtReimbursed	5.116e-05	3.474e-05	1.473	0.141
RenalDiseaseIndicator	8.426e-01	5.466e-01	1.542	0.123
Inpatient	1.967e+00	4.215e-01	4.666	3.07e-06 ***
ClaimLength	2.663e+00	2.683e+00	0.993	0.321
hos_stay_len	-2.646e+00	2.683e+00	-0.986	0.324
num_clms	5.103e-03	4.169e-04	12.242	< 2e-16 ***
ModeRace2	-1.466e+01	4.455e+02	-0.033	0.974
ModeRace3	8.107e-01	1.154e+00	0.703	0.482
ModeRace5	-1.410e+01	1.582e+03	-0.009	0.993

Choosing lambdas –

The best lambda for lasso and ridge was chosen through a search grid from  $e^{-8}$  to 1. The CV error rose with higher lambdas as all points were classified as non-fraudulent. The lambda which minimised CV error for ridge and lasso was 0.00012 and 0.00038 respectively. These models can be found below-

<sup>2</sup> He and Garcia – Learning from unbalanced data (2009)

<sup>3</sup> [https://link.springer.com/chapter/10.1007/978-3-540-77046-6\\_43](https://link.springer.com/chapter/10.1007/978-3-540-77046-6_43)



Ridge Model –

	1
(Intercept)	-2.385215e+01
InscClaimAmtReimbursed	2.101056e-04
DeductibleAmtPaid	8.755023e-04
Gender	-8.828194e-01
RenalDiseaseIndicator	8.809168e-01
NoOfMonths_PartACov	1.514677e+00
NoOfMonths_PartBCov	7.507895e-01
ChronicCond_Alzheimer	1.921173e-01
ChronicCond_Heartfailure	-1.200750e+00
ChronicCond_KidneyDisease	-1.722828e+00
ChronicCond_Cancer	-1.473916e-01
ChronicCond_ObstrPulmonary	-1.130703e+00
ChronicCond_Depression	-9.403087e-03
ChronicCond_Diabetes	3.662955e-01
ChronicCond_IschemicHeart	4.049670e-01
ChronicCond_Osteoporosis	-6.596607e-01
ChronicCond_rheumatoidarthritis	-4.416406e-01
ChronicCond_stroke	5.624409e-01
IPAnnualReimbursementAmt	1.755684e-06
IPAnnualDeductibleAmt	-1.321865e-05
OPAnnualReimbursementAmt	-6.849408e-06
OPAnnualDeductibleAmt	-4.016096e-04
Inpatient	1.591650e+00
changedCode	3.327704e-01
ClaimLength	2.455986e+00
Age	-3.971932e-02
hos_stay_len	-2.209393e+00
AttendingPhysicianNA	-1.703850e+00
OperatingPhysicianNA	2.116271e+00
ClmDiagnosisCodeNA	-1.389429e+01
DODNA	8.665796e+00
changedCodeNA	-9.546024e-01
ClmAdmitDiagnosisCodeNA	2.062876e+00
ClmProcedureCodeNA	-2.297412e+00
FraudPhys	5.015415e+00
FraudCounty	8.851936e-02
FraudState	1.348926e-01
num_clms	4.823943e-03
AvgClaimsPerBene	8.387295e-01
ModeRace2	-6.632889e+00
ModeRace3	3.331923e+00
ModeRace5	-1.399887e+00
StartDtdays_from_271108	-6.796714e-03
EndDtdays_from_271108	1.294317e-03
CondNumXInscClaimReimbursed	-2.984213e-05
numClmsXinpatient	1.278939e-01

Lasso model -

	1
(Intercept)	-2.385215e+01
InscClaimAmtReimbursed	2.101056e-04
DeductibleAmtPaid	8.755023e-04
Gender	-8.828194e-01
RenalDiseaseIndicator	8.809168e-01
NoOfMonths_PartACov	1.514677e+00
NoOfMonths_PartBCov	7.507895e-01
ChronicCond_Alzheimer	1.921173e-01
ChronicCond_Heartfailure	-1.200750e+00
ChronicCond_KidneyDisease	-1.722828e+00
ChronicCond_Cancer	-1.473916e-01
ChronicCond_ObstrPulmonary	-1.130703e+00
ChronicCond_Depression	-9.403087e-03
ChronicCond_Diabetes	3.662955e-01
ChronicCond_IschemicHeart	4.049670e-01
ChronicCond_Osteoporosis	-6.596607e-01
ChronicCond_rheumatoidarthritis	-4.416406e-01
ChronicCond_stroke	5.624409e-01
IPAnnualReimbursementAmt	1.755684e-06
IPAnnualDeductibleAmt	-1.321865e-05
OPAnnualReimbursementAmt	-6.849408e-06
OPAnnualDeductibleAmt	-4.016096e-04
Inpatient	1.591650e+00
changedCode	3.327704e-01
ClaimLength	2.455986e+00
Age	-3.971932e-02
hos_stay_len	-2.209393e+00
AttendingPhysicianNA	-1.703850e+00
OperatingPhysicianNA	2.116271e+00
ClmDiagnosisCodeNA	-1.389429e+01
DODNA	8.665796e+00
changedCodeNA	-9.546024e-01
ClmAdmitDiagnosisCodeNA	2.062876e+00
ClmProcedureCodeNA	-2.297412e+00
FraudPhys	5.015415e+00
FraudCounty	8.851936e-02
FraudState	1.348926e-01
num_clms	4.823943e-03
AvgClaimsPerBene	8.387295e-01
ModeRace2	-6.632889e+00
ModeRace3	3.331923e+00
ModeRace5	-1.399887e+00
StartDtdays_from_271108	-6.796714e-03
EndDtdays_from_271108	1.294317e-03
CondNumXInscClaimReimbursed	-2.984213e-05
numClmsXinpatient	1.278939e-01

### LDA and QDA –

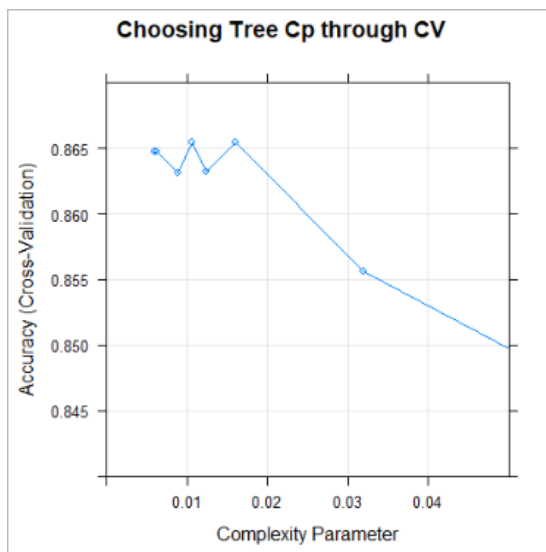
The LDA model was first scaled through caret and then fit using the MASS package. For PCA, the predictors were formed into principle components first and then fit.

### KNN –

The data was first normalized and then the number of neighbors was chosen by minimizing the CV error. 5 to 20 neighbors were considered and 7 had the lowest error.

### Trees –

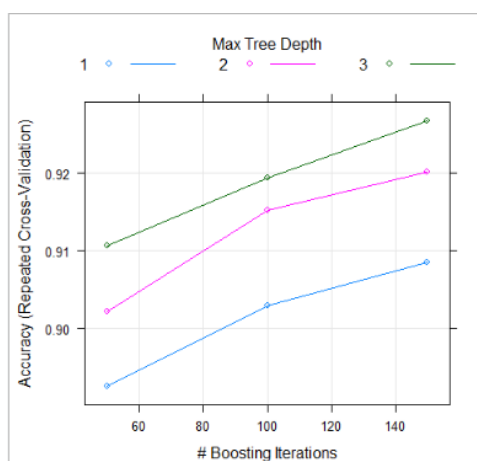
The complexity parameter was chosen through cross validation and yielded the best results at 0.0106383, showing again that flexible models best suit this data.



### Boosting trees –

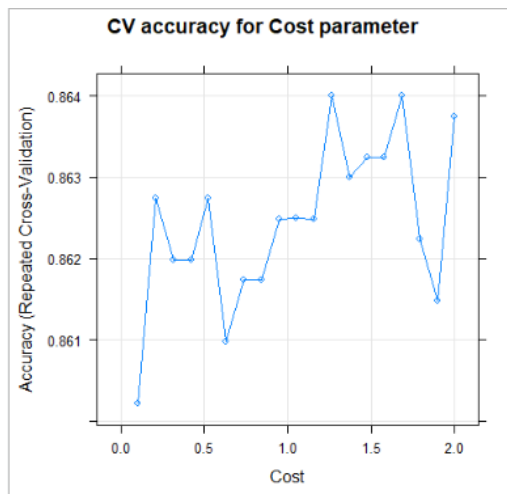
The GBM package was used to fit the model, and again, more flexible models are preferred. The trees with higher depths and more boosting iterations produce the lowest CV errors. The parameters which optimised performance were –

- 150 Trees
- 0.1 shrinkage
- Depth level of 3
- A minimum of 10 observations per node



## Support Vector Machines-

Both linear and radial boundaries were used to predict fraud. Linear was picked for its higher AUC. The “cost” parameter of 1.684 was picked to minimize the CV error.



## References:

1. He and Garcia – *Learning from unbalanced data* (2009)
2. Padmaja T.M., Dhulipalla N., Krishna P.R., Bapi R.S., Laha A. (2007) An Unbalanced Data Classification Model Using Hybrid Sampling Technique for Fraud Detection. In: Ghosh A., De R.K., Pal S.K. (eds) Pattern Recognition and Machine Intelligence. PReMI 2007. Lecture Notes in Computer Science, vol 4815. Springer, Berlin, Heidelberg.  
[https://doi.org/10.1007/978-3-540-77046-6\\_43](https://doi.org/10.1007/978-3-540-77046-6_43)