To further prevent medical fraud, we employ a decision tree model to detect anomalies in the data. Setting the random seed to 1000 ensures that running the code multiple times will produce the same results. Given the relatively large dataset of 558,211 data points, careful consideration is given to the data partitioning ratios to balance the requirements of model training and performance evaluation. After careful consideration, we have adopted the following partitioning scheme: approximately 80% of the data is allocated to the training set to ensure thorough model training. About 10% is allocated to the validation set for parameter tuning, and the remaining 10% is reserved for the test set for final performance evaluation. This partitioning strategy maximizes the utilization of the extensive dataset, ensuring comprehensive model training and reasonable performance assessment.

To achieve a better decision tree model. we fine-tune the model by adjusting three key parameters. The cp parameter is a crucial hyperparameter in decision trees. It controls the size of the tree by penalizing its complexity. A higher cp value leads to a smaller tree, while a lower cp value leads to a larger tree. The minsplit and maxdepth parameters also affect the size of the tree. A higher minsplit value leads to a smaller tree, while a higher maxdepth value leads to a larger tree. We employ these parameter adjustments to generate four distinct decision tree models for thorough comparison and to enhance the model's overall performance.

|  |  |  |  |
| --- | --- | --- | --- |
|  | CP | Minsplit | Maxdepth |
| Fit.full | / | / | / |
| Fit1 | 0.01 | 20 | 30 |
| Fit2 | 0.001 | 20 | 8 |
| Fit3 | 0.001 | 5 | 8 |

Table 4.1: Decision Tree Model Parameters Setting

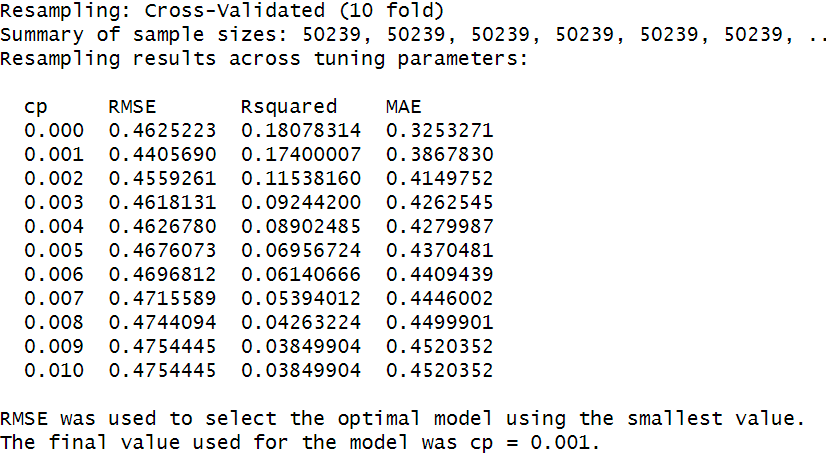
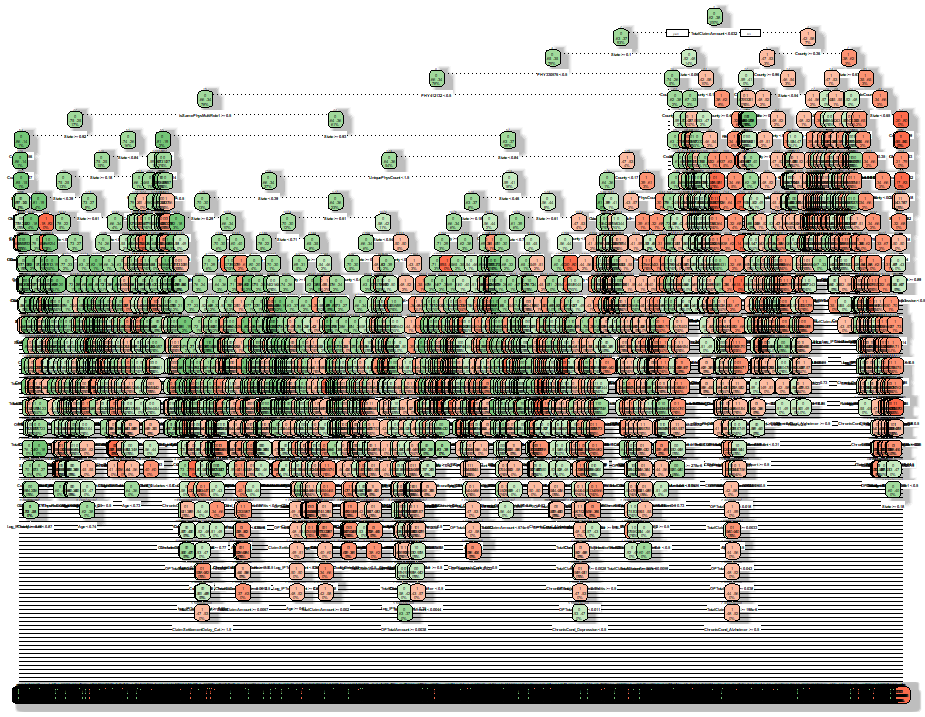


Table 4.2: K-Fold Cross-Validation for Best CP

First, we set up a model with no parameter constraints, allowing it to grow to its maximum extent to achieve optimal performance. Next, we utilized K-Fold Cross-Validation to compare results on the validation set and determined that the best value for the 'cp' parameter is 0.001, while we kept the default values for 'minsplit' (20) and 'maxdepth' (30) to assess the decision tree's performance under the optimal 'cp' value. Subsequently, we employed 'rpart' with a lower 'maxdepth' setting of 8, resulting in the creation of the third decision tree model, allowing us to observe the impact of 'maxdepth.' Finally, based on the third model, we reduced 'minsplit' to 5 to obtain the fourth model and assess the effects of 'minsplit."

Figure 4.1: Fit.full grow decision tree model

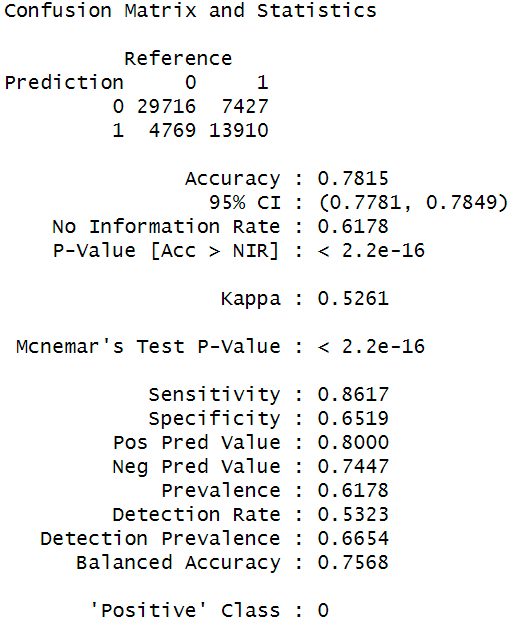
As shown in Figure 4.1. Due to the lack of constraints on any parameters, the chart of the Fit.full model exhibits an exceptionally high level of complexity, making it challenging to read and interpret. This complexity is evident in the depth and structure of the tree, resulting in a decision tree with an extensive number of nodes and branches that almost cover all possible data splitting points. Consequently, the chart is densely packed, and the relationships between nodes become unclear and difficult to comprehend.

Table 4.3: Fit.Full model Confusion Matrix and Statistics

A math equation with numbers and symbols

Description automatically generatedBase on the table4.3. The model exhibits an accuracy of approximately 78.15%, indicating that it correctly classifies nearly 78.15% of instances. The Kappa statistic, with a value of about 0.5261, suggests moderate agreement between the model's predictions and actual classifications. Sensitivity, or recall, stands at around 86.17%, highlighting the model's strong ability to correctly identify instances of the positive class. Precision, which is approximately 80.00%, reveals the proportion of true positive predictions among all positive predictions. Furthermore, the F1-Score, a combination of precision and recall.

In this report, we assume all β =1 then F1-Score = 2 \* (0.80 \* 0.8617) / (0.80 + 0.8617). So, the F1-Score is approximately 0.8278.

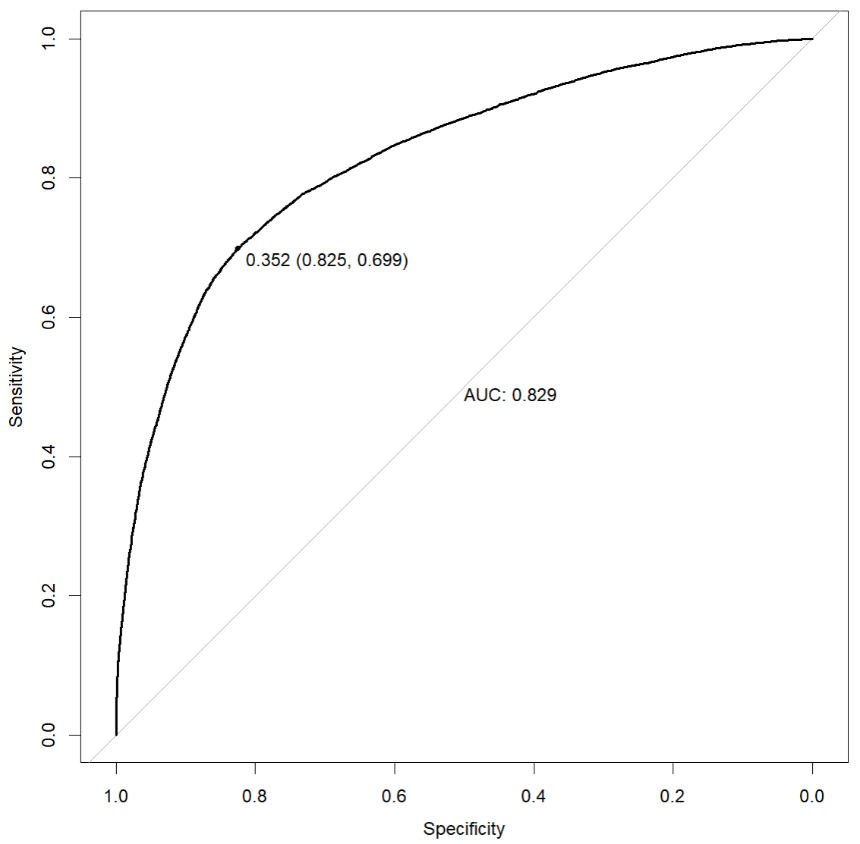
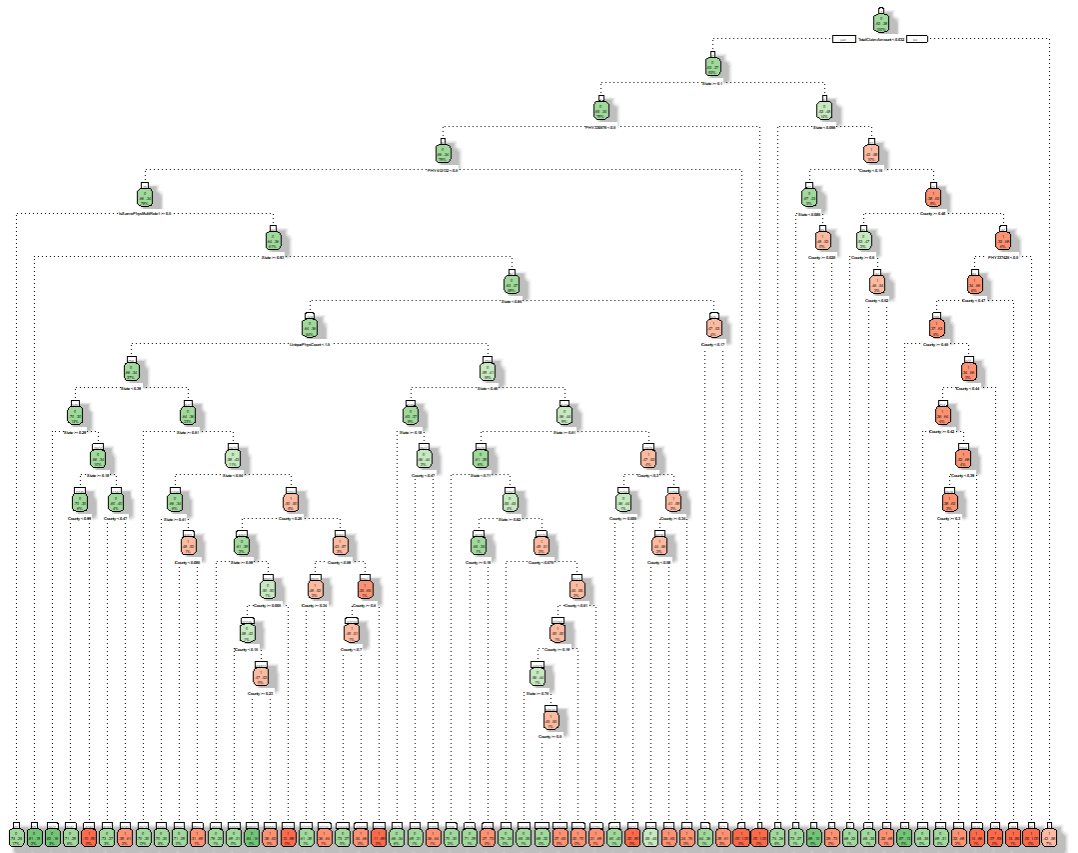


Figure 4.2: Fit.full model AUC

As shown in Figure 4.2, a full-grow model with an AUC of 0.829 suggests that the model performs well in distinguishing between classes. At the threshold of 0.352, the model achieves a good balance with 82.5% specificity (correctly identifying negatives) and 69.9% sensitivity (correctly detecting positives).

 Figure 4.3: Fit1 decision tree model

As shown in Figure 4.3, even though the Fit1 model still exhibits considerable depth in the tree when viewed on the graph, the model remains relatively complex, making it evident that it has performed multiple intricate data partitions to arrive at the final classification decisions. However, in comparison to the fit.full model, the tree's overall structure and individual nodes can now be roughly discerned.

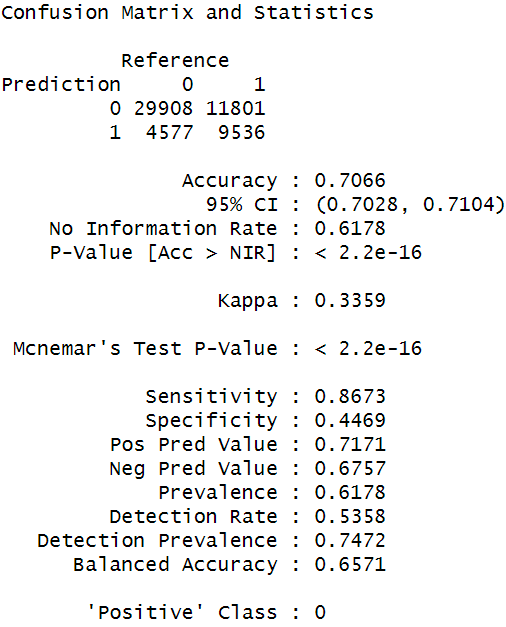


Table 4.4: Fit1 model Confusion Matrix and Statistics

The table 4.4 clearly reflects the Fit1 model’s performance, shows an accuracy of 70.66%, which indicates the proportion of correctly classified instances. The precision is approximately 71.71%, demonstrating the proportion of true positive predictions among all positive predictions. The recall is about 86.73%, representing the model's ability to correctly identify the positive class. The Kappa statistic is 0.3359, suggesting a fair level of agreement between predicted and actual classifications. The F1-Score is approximately 0.7854 suggests a reasonably good balance between precision and recall in the classification model.

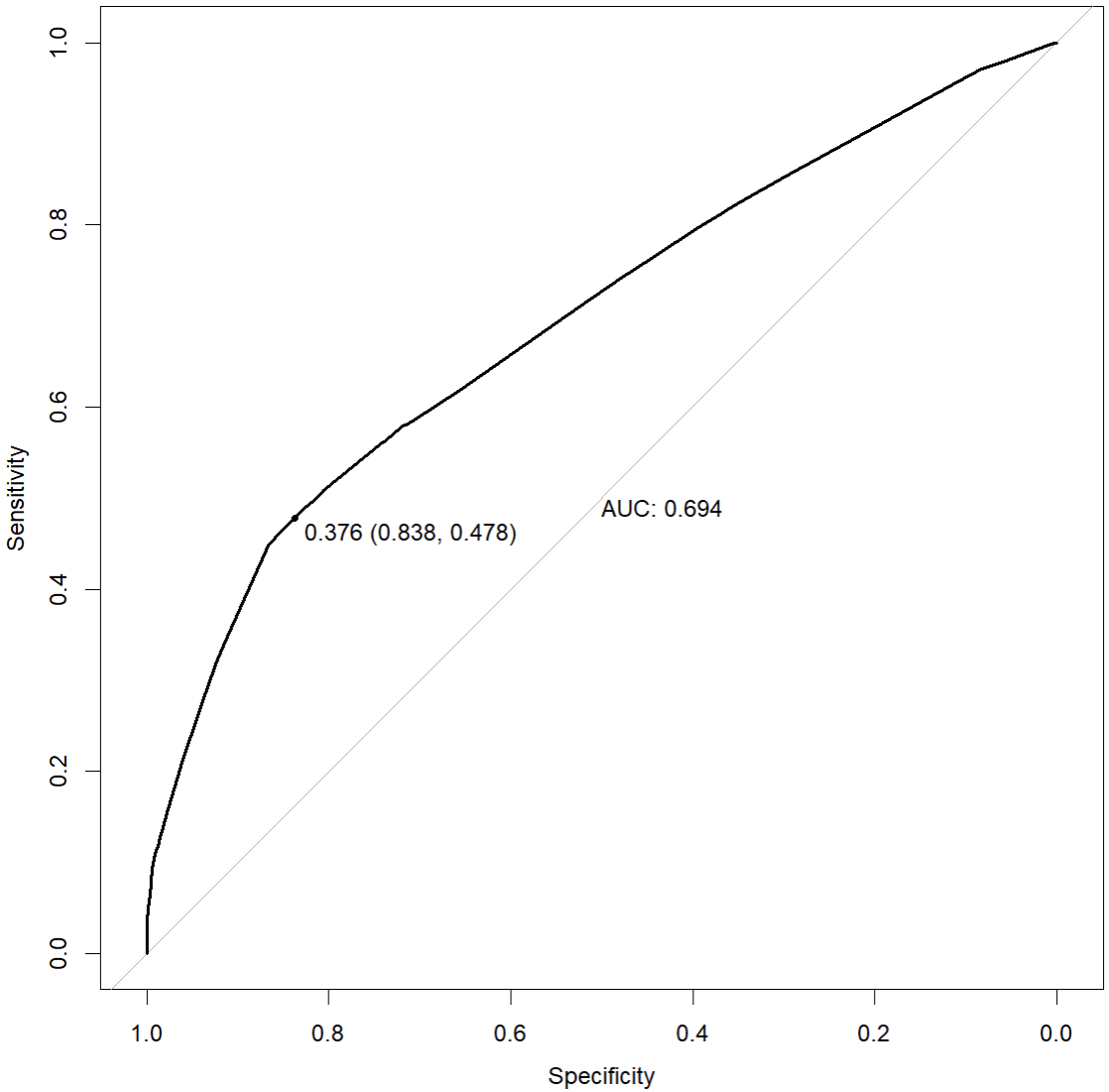
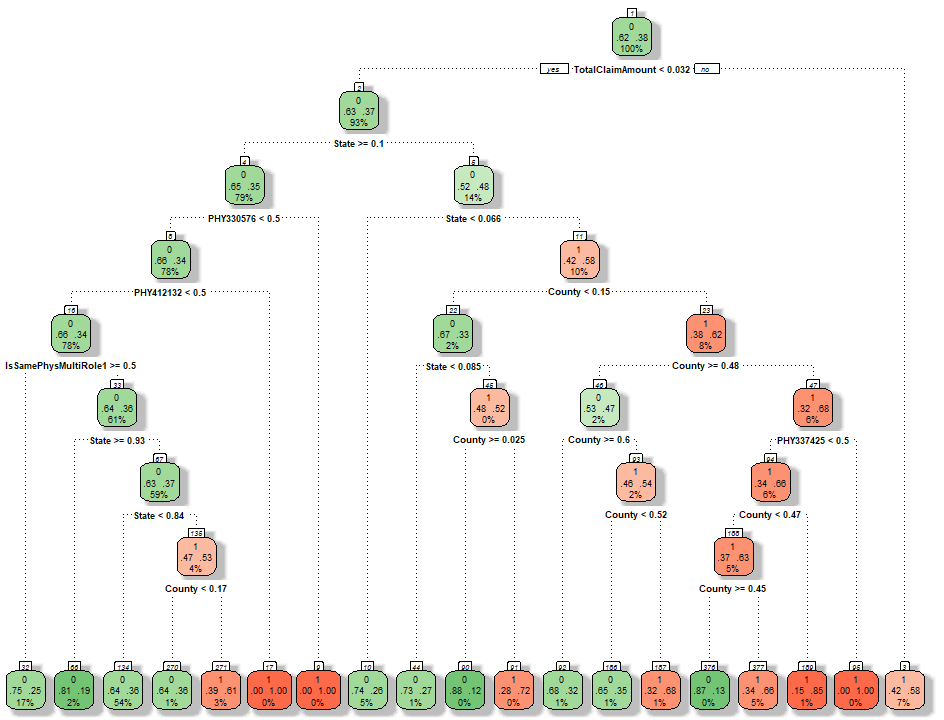


Figure 4.4: Fit1 model AUC

Figure 4.4 displays the AUC of the Fit1 model's performance. The AUC of 0.694 indicates moderate model performance in distinguishing between classes. The optimal threshold of 0.376 balances sensitivity (47.8%) and specificity (83.8%), suggesting that the model can detect positives reasonably well while minimizing false positives.

Figure 4.5: Fit2 decision tree model

In Figure 4.5, you can clearly see the overall structure of the Fit2 model. By reducing the model's parameter 'maxdepth,' significant simplification of the model was achieved, making all the structures and nodes clearly visible. The model depth has been well constrained to 8. This demonstrates the substantial impact of 'maxdepth' when dealing with complex decision tree models.

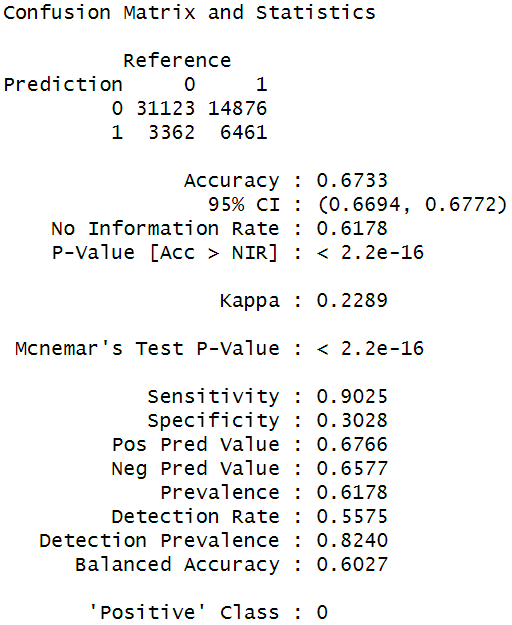


Table 4.5: Fit2 model Confusion Matrix and Statistics

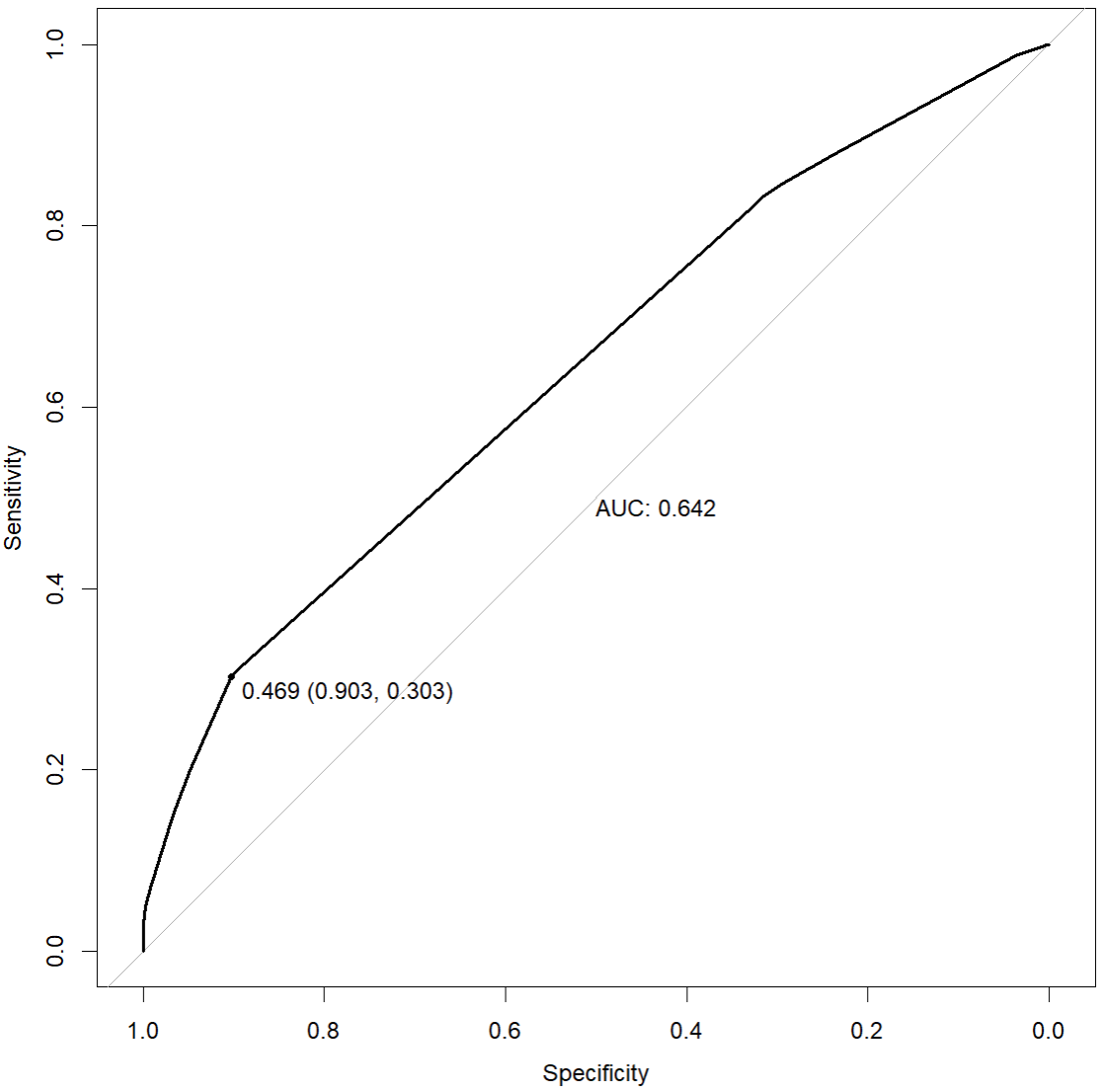
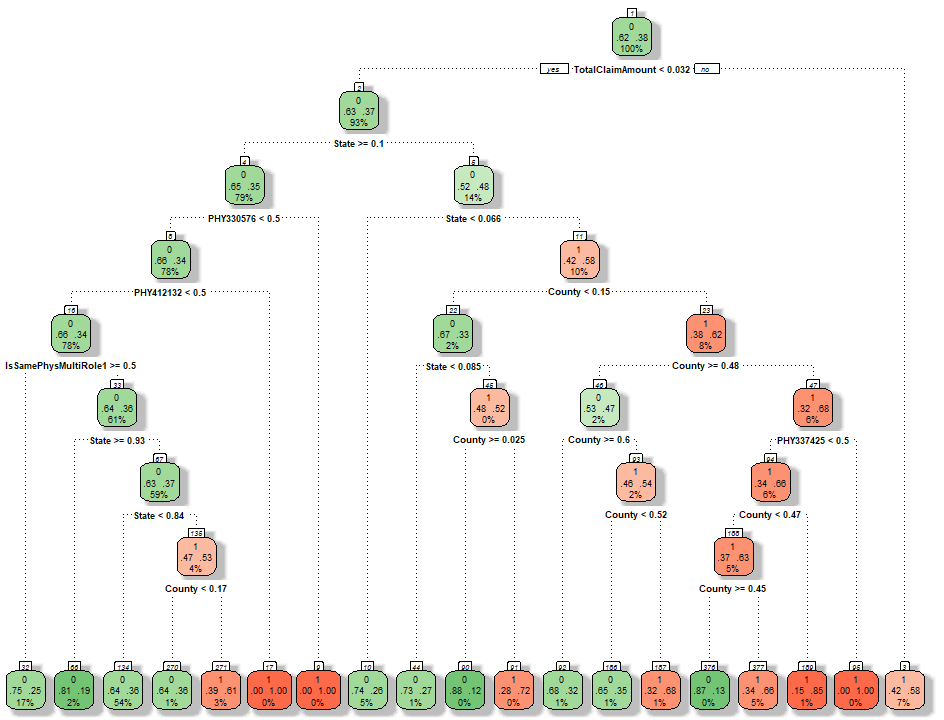
In Table 4.5 Fit2 model's performance, as indicated by the accuracy of 67.33%, demonstrates its ability to correctly classify instances. However, it's important to note that the model's kappa value of 0.2289 suggests only a fair level of agreement beyond what would be expected by chance. The recall is relatively high at 90.25%, highlighting the model's proficiency in correctly identifying the positive class. On the other hand, precision value of 0.6766, indicates a relatively good rate of accuracy for positive predictions. The F-Score is approximately 0.7746 indicates a good balance between precision and recall in Fit2 model's performance.

Figure 4.6: Fit2 model AUC

In Fit2 model’s AUC plot (Figure 4.6), the AUC value of 0.642 suggests that the model has moderate discriminatory power, but it may not be well-balanced in terms of sensitivity and specificity. The optimal threshold of 0.469 prioritizes specificity (minimizing false negatives (90%)) over sensitivity (correctly identifying positives (30%)).

 Figure 4.7: Fit3 decision tree model

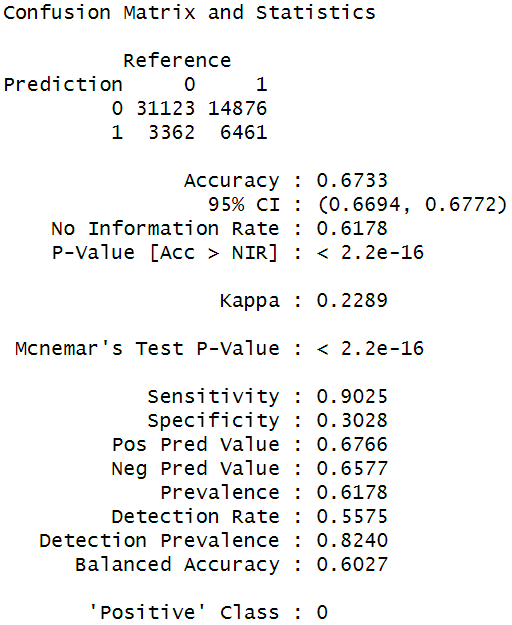
In the Fit3 model, the model's complexity was further reduced by lowering the minsplit parameter to 5, in addition to the previously reduced maxdepth. This was done to simplify the model further and observe the impact of minsplit. However, as shown in Figure 4.7, the decision tree model remained unchanged compared to the Fit2 model. This could indicate that the minsplit parameter has a limited influence on this dataset.

Table 4.6: Fit3 model Confusion Matrix and Statistics

From Table 4.6, it can be observed that the confusion matrix and statistical results of the Fit3 model are quite similar to the performance of the Fit2 model. Both models exhibit an accuracy 67.33%, a kappa value 0.2289, a recall of 90.25%, a precision value of 0.6766, and an F-score approximately 0.7746. This further confirms the correctness of Figure 4.7.

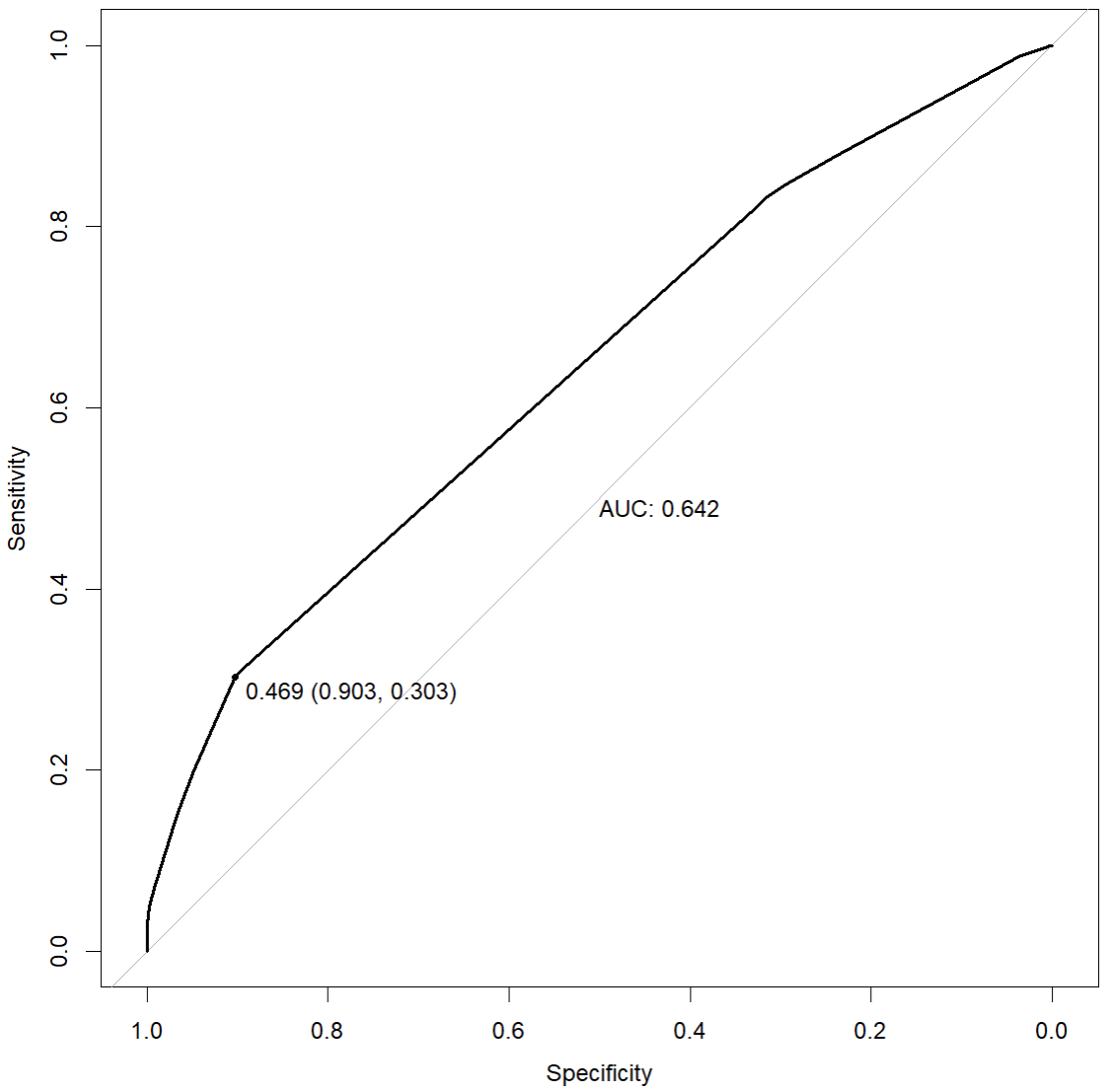


Figure 4.8: Fit3 model AUC

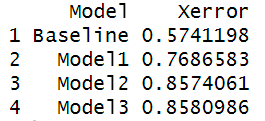
The AUC chart (Figure 4.8) for the Fit3 model clearly shows the same coordinates and performance as the Fit2 model.

Table 4.7：Cross-Validation Error

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Accuracy | Kappa | Recall | Precision | F-score | AUC | Xerror |
| Fit.full | 0.7815 | 0.5261 | 0.8617 | 0.8 | 0.8278 | 0.829 | 0.5741198 |
| Fit1 | 0.7066 | 0.3359 | 0.8673 | 0.7171 | 0.7854 | 0.694 | 0.7686583 |
| Fit2 | 0.6733 | 0.2289 | 0.9025 | 0.6766 | 0.7746 | 0.642 | 0.8574061 |
| Fit3 | 0.6733 | 0.2289 | 0.9025 | 0.6766 | 0.7746 | 0.642 | 0.8580986 |

Table 4.8：Model Performance Compare

#####################如果太长了可以用简化版 345words

In the comparison of these models, Fit.full and Fit1 perform relatively well, while Fit2 and Fit3 show slightly weaker performance.

The Fit.full model excels in various aspects. It achieves relatively high scores in terms of accuracy, Kappa statistics, F-score, and AUC. This indicates that the Fit.full model successfully classifies most of the samples, outperforming random classification. Moreover, it boasts a high recall rate, signifying its ability to effectively identify positive-class samples, while maintaining a high level of precision, striking a crucial balance. In cross-validation, its Xerror is relatively low, indicating that the model's performance remains relatively consistent across different data subsets.

The Fit1 model demonstrates exceptional recall performance, nearly on par with Fit.full, albeit with slightly lower accuracy and precision. Despite the lower accuracy, the F-score remains relatively high, signifying Fit1's strong overall performance. However, Fit1 exhibits a relatively higher Xerror in cross-validation, suggesting that its performance may vary significantly across different data subsets.

The only distinction between the Fit2 and Fit3 models lies in the Xerror metric. Xerror measures cross-validation error, assessing the performance variation of the model across different data subsets. In this case, Fit2 has a relatively high Xerror (85.74%), and Fit3's Xerror is slightly higher at 85.81%. This suggests that the performance of Fit2 and Fit3 exhibits significant instability across different data subsets, possibly due to inconsistent model behavior on various data subsets. Consequently, these two models demonstrate relatively unstable performance and are slightly inferior in overall performance compared to Fit.full and Fit1.

In conclusion, Fit.full and Fit1 demonstrate relatively stable and balanced performance. However, it's essential to consider some key factors since Fit.full is a fully grown decision tree model. Firstly, fully grown models are prone to overfitting, excelling in fitting the training data but lacking in generalization to new data. Secondly, such models are often very complex, challenging to interpret, and demand significant computational resources. Lastly, to address overfitting and complexity issues, pruning is typically required for fully grown trees. Therefore, after weighing performance, model complexity, and computational resources, we conclude that the Fit1 model is the preferred choice.

###########################

###########################简化版 208 words

When comparing these models, Fit.full and Fit1 stand out as strong performers, while Fit2 and Fit3 exhibit slightly weaker performance.

Fit.full excels in various aspects, demonstrating high accuracy, Kappa statistics, F-score, and AUC, indicating effective sample classification. It also maintains a high recall rate for positive-class samples while preserving precision. Cross-validation reveals relatively low Xerror, indicating consistent performance across data subsets.

Fit1, on the other hand, boasts exceptional recall performance, closely rivaling Fit.full, though it exhibits slightly lower accuracy and precision. The F-score remains high, highlighting its overall strength. However, Fit1 shows a higher Xerror in cross-validation, suggesting potential performance variation across different data subsets.

Fit2 and Fit3 are nearly identical, differing primarily in the Xerror metric. Both models display high Xerror values, indicating significant performance instability across various data subsets. This instability could be due to inconsistent model behavior on different data subsets, resulting in slightly inferior overall performance compared to Fit.full and Fit1.

In summary, Fit.full and Fit1 offer relatively stable and balanced performance. However, it's essential to consider that Fit.full is a fully grown decision tree model, which can lead to overfitting and complexity. Pruning may be necessary to address these issues. Considering performance, model complexity, and computational resources, Fit1 emerges as the preferred choice.

#######################################

By examining Figure 4.3 and the Fit1 model summary in the appendix, it becomes clear that the Fit1 model relies on several crucial attributes for making predictions. 'TotalClaimAmount' emerges as a central factor, acting as the initial decision point and underscoring its importance in detecting potential fraud. 'State' is consistently influential, indicating regional variations in potential fraud cases.

Additionally, 'PHY330576' and 'PHY412132' play pivotal roles, especially when combined with 'IsSamePhysMultiRole1,' emphasizing their predictive significance. 'UniquePhysCount' and 'County' are also notable, leading to distinct branches in the tree when specific thresholds are met. 'County' is used in multiple splits, highlighting its critical role in the model's decision-making process.

In summary, 'TotalClaimAmount' and 'State' are primary drivers of predictions in the Fit1 model. However, 'PHY330576,' 'PHY412132,' 'IsSamePhysMultiRole1,' 'UniquePhysCount,' and 'County' are also essential variables for making accurate predictions regarding potential fraud, taking into account geographic and physician-related factors.

Appendix:

n= 446568

node), split, n, loss, yval, (yprob)

\* denotes terminal node

1) root 446568 170393 0 (0.61843885 0.38156115)

2) TotalClaimAmount< 0.03223657 416216 152855 0 (0.63275078 0.36724922)

4) State>=0.1037736 351577 122129 0 (0.65262517 0.34737483)

8) PHY330576< 0.5 349526 120078 0 (0.65645474 0.34354526)

16) PHY412132< 0.5 347805 118357 0 (0.65970299 0.34029701)

32) IsSamePhysMultiRole1>=0.5 74705 18879 0 (0.74728599 0.25271401) \*

33) IsSamePhysMultiRole1< 0.5 273100 99478 0 (0.63574515 0.36425485)

66) State>=0.9339623 11134 2083 0 (0.81291539 0.18708461) \*

67) State< 0.9339623 261966 97395 0 (0.62821511 0.37178489)

134) State< 0.8396226 242534 87068 0 (0.64100703 0.35899297)

268) UniquePhysCount< 1.5 163415 54822 0 (0.66452284 0.33547716)

536) State< 0.3867925 60126 17965 0 (0.70121079 0.29878921)

1072) State>=0.2924528 14820 2730 0 (0.81578947 0.18421053) \*

1073) State< 0.2924528 45306 15235 0 (0.66373107 0.33626893)

2146) State>=0.1792453 27310 8069 0 (0.70454046 0.29545954)

4292) County< 0.9914915 26986 7776 0 (0.71185059 0.28814941) \*

4293) County>=0.9914915 324 31 1 (0.09567901 0.90432099) \*

2147) State< 0.1792453 17996 7166 0 (0.60180040 0.39819960)

4294) County< 0.4654655 11239 3054 0 (0.72826764 0.27173236) \*

4295) County>=0.4654655 6757 2645 1 (0.39144591 0.60855409) \*

537) State>=0.3867925 103289 36857 0 (0.64316626 0.35683374)

1074) State>=0.6132075 54689 16498 0 (0.69833056 0.30166944) \*

1075) State< 0.6132075 48600 20359 0 (0.58109053 0.41890947)

2150) State< 0.5377358 25923 8931 0 (0.65547969 0.34452031)

4300) State>=0.4056604 21058 6400 0 (0.69607750 0.30392250) \*

4301) State< 0.4056604 4865 2334 1 (0.47975334 0.52024666)

8602) County< 0.08508509 2033 590 0 (0.70978849 0.29021151) \*

8603) County>=0.08508509 2832 891 1 (0.31461864 0.68538136) \*

2151) State>=0.5377358 22677 11249 1 (0.49605327 0.50394673)

4302) County< 0.2552553 8079 3172 0 (0.60737715 0.39262285)

8604) State>=0.5943396 3041 672 0 (0.77902006 0.22097994) \*

8605) State< 0.5943396 5038 2500 0 (0.50377134 0.49622866)

17210) County>=0.005005005 4223 1783 0 (0.57778830 0.42221170)

34420) County< 0.1451451 2078 638 0 (0.69297401 0.30702599) \*

34421) County>=0.1451451 2145 1000 1 (0.46620047 0.53379953)

68842) County>=0.2252252 391 62 0 (0.84143223 0.15856777) \*

68843) County< 0.2252252 1754 671 1 (0.38255416 0.61744584) \*

17211) County< 0.005005005 815 98 1 (0.12024540 0.87975460) \*

4303) County>=0.2552553 14598 6342 1 (0.43444307 0.56555693)

8606) County< 0.5855856 9550 4562 1 (0.47769634 0.52230366)

17212) County>=0.3358358 4477 1756 0 (0.60777306 0.39222694) \*

17213) County< 0.3358358 5073 1841 1 (0.36290164 0.63709836) \*

8607) County>=0.5855856 5048 1780 1 (0.35261490 0.64738510)

17214) County>=0.5955956 3244 1590 1 (0.49013564 0.50986436)

34428) County< 0.6956957 1268 347 0 (0.72634069 0.27365931) \*

34429) County>=0.6956957 1976 669 1 (0.33856275 0.66143725) \*

17215) County< 0.5955956 1804 190 1 (0.10532151 0.89467849) \*

269) UniquePhysCount>=1.5 79119 32246 0 (0.59243671 0.40756329)

538) State< 0.4622642 36921 13509 0 (0.63411067 0.36588933)

1076) State>=0.1792453 26853 9058 0 (0.66268201 0.33731799) \*

1077) State< 0.1792453 10068 4451 0 (0.55790624 0.44209376)

2154) County< 0.4654655 6058 1874 0 (0.69065698 0.30934302) \*

2155) County>=0.4654655 4010 1433 1 (0.35735661 0.64264339) \*

539) State>=0.4622642 42198 18737 0 (0.55597422 0.44402578)

1078) State>=0.6132075 26474 10380 0 (0.60791720 0.39208280)

2156) State< 0.7075472 9998 2955 0 (0.70444089 0.29555911) \*

2157) State>=0.7075472 16476 7425 0 (0.54934450 0.45065550)

4314) State>=0.8207547 6090 2120 0 (0.65188834 0.34811166)

8628) County>=0.1751752 5303 1548 0 (0.70808976 0.29191024) \*

8629) County< 0.1751752 787 215 1 (0.27318933 0.72681067) \*

4315) State< 0.8207547 10386 5081 1 (0.48921625 0.51078375)

8630) County< 0.07507508 1396 336 0 (0.75931232 0.24068768) \*

8631) County>=0.07507508 8990 4021 1 (0.44727475 0.55272525)

17262) County< 0.6056056 6591 3286 1 (0.49855864 0.50144136)

34524) County>=0.1851852 5058 2232 0 (0.55871886 0.44128114)

69048) State>=0.745283 2742 955 0 (0.65171408 0.34828592) \*

69049) State< 0.745283 2316 1039 1 (0.44861831 0.55138169)

138098) County>=0.4954955 575 182 0 (0.68347826 0.31652174) \*

138099) County< 0.4954955 1741 646 1 (0.37105112 0.62894888) \*

34525) County< 0.1851852 1533 460 1 (0.30006523 0.69993477) \*

17263) County>=0.6056056 2399 735 1 (0.30637766 0.69362234) \*

1079) State< 0.6132075 15724 7367 1 (0.46851946 0.53148054)

2158) County< 0.2952953 6061 2654 0 (0.56211846 0.43788154)

4316) County>=0.005005005 5456 2169 0 (0.60245601 0.39754399) \*

4317) County< 0.005005005 605 120 1 (0.19834711 0.80165289) \*

2159) County>=0.2952953 9663 3960 1 (0.40981062 0.59018938)

4318) County>=0.3358358 8010 3556 1 (0.44394507 0.55605493)

8636) County< 0.5755756 3644 1624 0 (0.55433589 0.44566411) \*

8637) County>=0.5755756 4366 1536 1 (0.35180944 0.64819056) \*

4319) County< 0.3358358 1653 404 1 (0.24440411 0.75559589) \*

135) State>=0.8396226 19432 9105 1 (0.46855702 0.53144298)

270) County< 0.1651652 6140 2208 0 (0.64039088 0.35960912) \*

271) County>=0.1651652 13292 5173 1 (0.38918146 0.61081854) \*

17) PHY412132>=0.5 1721 0 1 (0.00000000 1.00000000) \*

9) PHY330576>=0.5 2051 0 1 (0.00000000 1.00000000) \*

5) State< 0.1037736 64639 30726 0 (0.52465230 0.47534770)

10) State< 0.06603774 20672 5293 0 (0.74395317 0.25604683) \*

11) State>=0.06603774 43967 18534 1 (0.42154343 0.57845657)

22) County< 0.1451451 6753 2222 0 (0.67096105 0.32903895)

44) State< 0.08490566 5187 1403 0 (0.72951610 0.27048390) \*

45) State>=0.08490566 1566 747 1 (0.47701149 0.52298851)

90) County>=0.02502503 522 65 0 (0.87547893 0.12452107) \*

91) County< 0.02502503 1044 290 1 (0.27777778 0.72222222) \*

23) County>=0.1451451 37214 14003 1 (0.37628312 0.62371688)

46) County>=0.4754755 10058 4765 0 (0.52624776 0.47375224)

92) County>=0.5955956 2834 894 0 (0.68454481 0.31545519) \*

93) County< 0.5955956 7224 3353 1 (0.46414729 0.53585271)

186) County< 0.5155155 3128 1084 0 (0.65345269 0.34654731) \*

187) County>=0.5155155 4096 1309 1 (0.31958008 0.68041992) \*

47) County< 0.4754755 27156 8710 1 (0.32073943 0.67926057)

94) PHY337425< 0.5 25745 8710 1 (0.33831812 0.66168188)

188) County< 0.4654655 21918 8133 1 (0.37106488 0.62893512)

376) County>=0.4454454 1371 172 0 (0.87454413 0.12545587) \*

377) County< 0.4454454 20547 6934 1 (0.33747019 0.66252981)

754) County< 0.4354354 18895 6822 1 (0.36104790 0.63895210)

1508) County>=0.4154154 2127 745 0 (0.64974142 0.35025858) \*

1509) County< 0.4154154 16768 5440 1 (0.32442748 0.67557252)

3018) County< 0.3853854 12855 4906 1 (0.38164138 0.61835862)

6036) County>=0.2952953 2014 622 0 (0.69116187 0.30883813) \*

6037) County< 0.2952953 10841 3514 1 (0.32413984 0.67586016) \*

3019) County>=0.3853854 3913 534 1 (0.13646818 0.86353182) \*

755) County>=0.4354354 1652 112 1 (0.06779661 0.93220339) \*

189) County>=0.4654655 3827 577 1 (0.15077084 0.84922916) \*

95) PHY337425>=0.5 1411 0 1 (0.00000000 1.00000000) \*

3) TotalClaimAmount>=0.03223657 30352 12814 1 (0.42217976 0.57782024) \*