University of
South Australia

Predictive & Descriptive Analytics 2024
Assignment 3

Classification

Students
Huyen Thi Thu Pham

Contents

1 Introduction and Recap

In previous assignments, we embarked on an exploration of data using R, focusing on the analysis of extracted features from the dataset. The assignment was initiated with a literature review of three articles, each concentrating on different aspects: healthcare fraud classifiers, the application of supervised machine learning on imbalanced datasets, and the significance of feature selection in healthcare fraud datasets.

Subsequent to the literature review, we delved into comprehensive data exploration and preprocessing. This involved addressing missing data, feature extraction, creation features, and elimination features, and the utilization of decision tree models to ascertain feature importance. We presented descriptive statistics via tables and figures, showcasing value distributions, histograms, bar plots, and box plots. Univariate analysis was employed to comprehend the distribution of individual features, while correlation analysis was used to investigate the relationships between crucial variables. For the construction of decision tree models, we employed R, with an emphasis on comprehending data partitioning and model parameterization. The data was divided into training and testing sets in a 70:30 ratio.

We explored various parameters like maximum depth, minimum node split size, and complexity parameter to understand their influence on model performance. Five decision tree models were constructed with varying parameter values, and their structures were visualized through plots. We also reported several performance indicators such as accuracy, precision, recall, F-score, kappa, and AUC for each model, derived from confusion matrices, to evaluate their predictive effectiveness.

Upon comparing the models, Model 4 emerged as the superior one, achieving the highest values across multiple performance metrics with kappa =0.54, accuracy =0.79, precision = 0.65, F1-score =0.71, recall =0.76. This model utilized only four features (State, Age, County, otherPhysician) and proved that these features were adequate for effective fraud detection within the dataset. The analysis underscored that fraud was more common in certain counties and states, thereby highlighting the importance of these features in our final model.

2 Data Exploration and Feature Selection Process

In the previous assignments, we focused on exploring and preprocessing the healthcare dataset to identify key variables that could effectively flag providers as potential fraud.

This assignment continues from where we left off, incorporating our findings and refining our approach. Here, we summarize the data exploration, preprocessing, and feature selection processes, highlighting key steps and decisions made, which pretty much the same in the previous assignment.

2.1 The Dataset

The first step is to understand what each of the variables represents in the dataset to gain domain knowledge and how they might relate to the potential fraud:

1. **Beneficiary Data**: Contains patient details such as age, ethnicity, gender, location, chronic health issues, deductible payments, and reimbursement amounts.
2. **Outpatient Data** : Contains claim details (with provider) for patients who received medical services without being admitted to the hospital.
3. **Inpatient Data** : Provides claim details (with provider) for patients hospitalized for medical services. It has 3 extra columns Admission date, Discharge date, and Diagnosis Group code.
4. **Provider Data** : Consists of the medicare providerID and their corresponding label of being potentially fraudulent or not.

2.2 Data Preprocessing

To prepare the datasets for analysis, we undertook several preprocessing steps:

Handling Missing Data: Missing information was imputed accordingly.

Data Merging

The **Inpatient** and **Outpatient** datasets were joined based on common columns and then merged with **Beneficiary** details using **'BeneID'**. Finally, potentially fraudulent information was merged using **'ProviderID'**.

Feature Creation: New features were added to enhance the dataset. In previous works, we generated mean_AttendingPhysician, mean_OperatingPhysician, mean_OtherPhysician which is group by attending physician, operating physician and other physician. In this assignment, we generate the mean aggregated features for every provider to represent the differentiation between the fraudulent and non-fraudulent cases; and count of different physicians Attended, clmDiagnosisCode and ClmProcedurecode, ChronicCond of a beneficiary.

- Aged
- isDeath
- ClaimPeriod
- AdmissionPeriod
- Group by providers, we create

  - mean_TotalReimbursementAmt
  - mean_TotalDeductibleAmt
  - mean_InsClaimAmtReimbursed
- count_BeneID, ClmCount_Provider
- Count of a beneficiary
  - Total_physician_attend
  - Total_diagnoses
  - Total_procedure
  - Total_ChronicCond

Feature Removal:

- ID Features
- Features with all NA values
- Features from which other features were created
- Feature have no variability

Summary Statistics

Table 1: Summary Statistics for Selected Variables

Variable	Min	1st Qu	Median	Mean	3rd Qu	Max
mean_TotalReimbursementAmt	10	6240	6936	7506	8116	107090
mean_TotalDeductibleAmt	0	1063	1167	1218	1303	13885
mean_InscClaimAmtReimbursed	0	281.8	334.4	997.0	971.1	57000
Total_physician_attend	0	1	1	1.56	2	3
Total_diagnoses	1	2	3	4	5	12
count_BeneID	1	3	6	6.617	9	29
Age	40	82	89	87.84	97	115
AdmissionPeriod	0	0	0	0.4108	0	35
ClaimPeriod	0	0	0	1.728	0	36

Key Insights from Summary Statistics:

- The median age of the patients is 89 years old, indicating that a significant portion of the patients are elderly.
- Right-skewed distribution with high maximum values (107,090 for total reimbursement, 13,885 for total deductible, and 57,000 for insured claim amount

reimbursed). This skewness indicates that while most claims and reimbursements are relatively low, there are a few very high-value claims that significantly affect the mean values.

- The maximum time in admission for patients is 35 days, while it is 36 days to the claim period.
- The maximum mean of claim amount reimbursed group by provider is 107,090 highlighting the presence of some very high-value claims.
- The mean annual deductible amount group by provider is $997.0, indicating significant out-of-pocket expenses for patients.

Visualization

In previous assignment, we used histograms, bar chart, bar plot, pie chart, Scatter plot to visualise

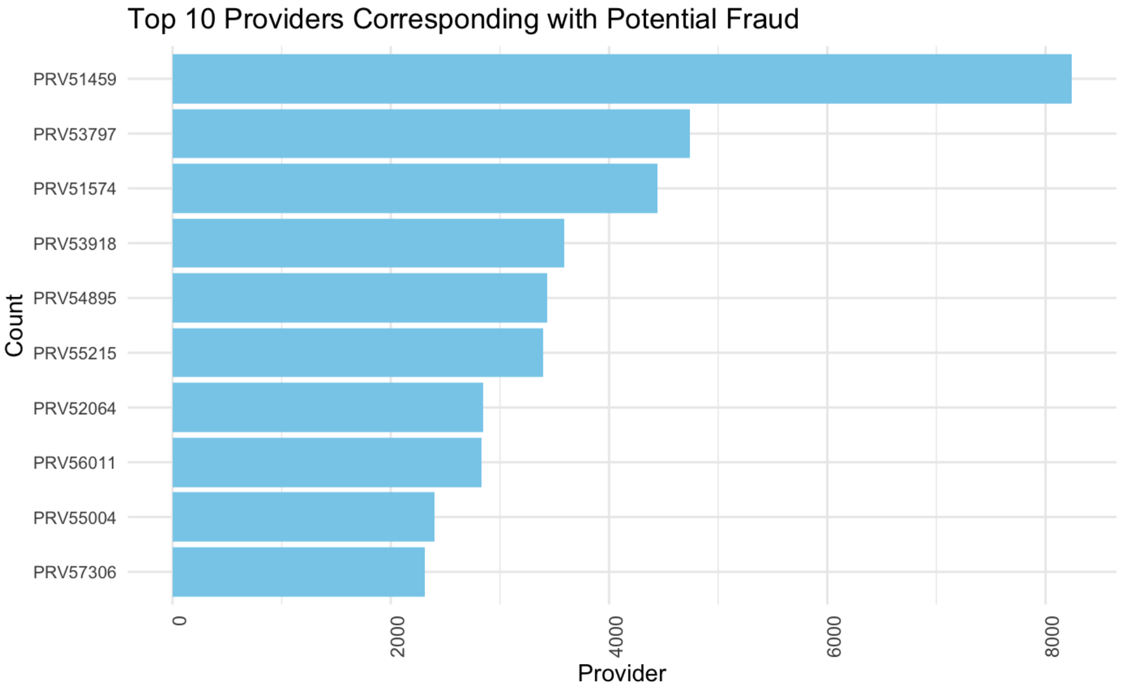
- Race
- Top 10 States by Potential Fraud
- Top 10 Counties with Potential Fraud
- Top 10 Attending Physicians with Potential Fraud
- Top 10 Providers with Potential Fraud
- ClaimPeriod vs. AdmissionPeriod
- Total Reimbursement Amount by Potential Fraud
- Distribution of Fraud & Non-Fraud Providers
- Categorical Features
- Numerical Features

Key insights from Data Visualisation:

- Certain beneficiaries might be experiencing or susceptible to fraud, particularly those with high reimbursements and high deductible payments.
- Patients with multiple chronic conditions were also notable.
- A small percentage of provider are involved in fraudulent could be responsible for a high number of fraudulent claims.
- Possibly fraudulent providers have high average claim reimbursement amounts and the highest reimbursement amounts in the datasets, along with a high average number of patient insurance claims.
- Possibly fraudulent providers may be more active in certain states and counties. Factors like a patient's age, location, total claim amount, and primary

doctor can indicate vulnerability to fraud and help differentiate between fraud and non-fraud providers.

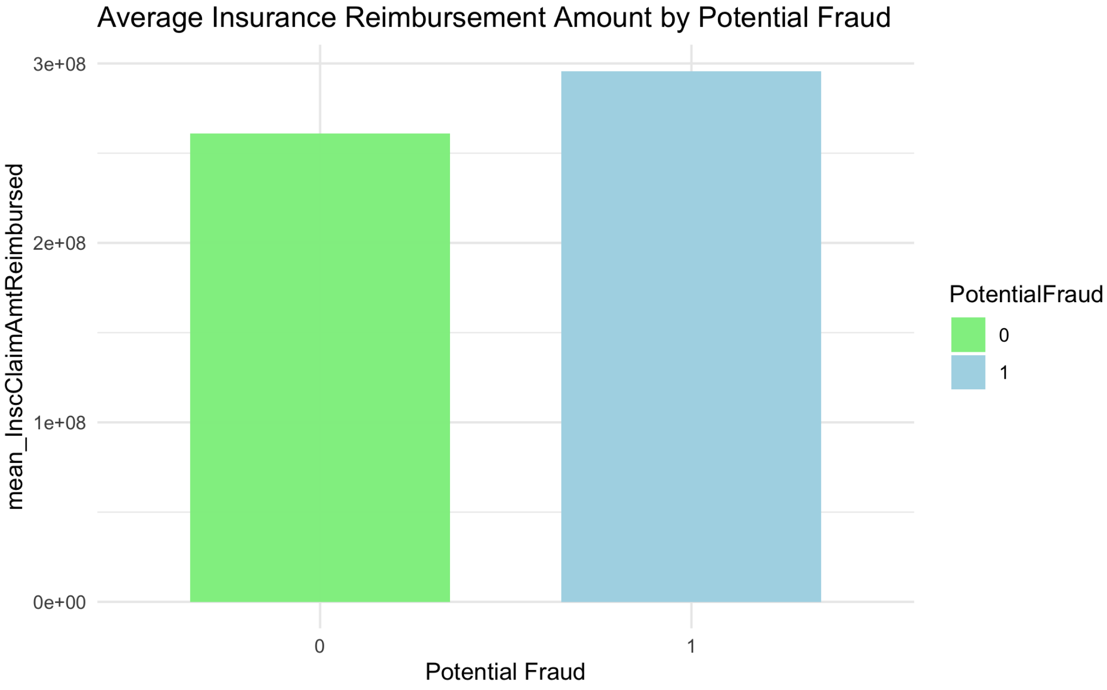
2.3 Chosen Plot and Explanation



In assignment 2, we employed a plot to presents the top 10 provider with the highest cases of fraudulent submissions.

The plot shows that PRV51459 leads with the highest share (about 8100 cases), while the differences among the other are relatively small.

In this assignment, I grouped providers to gain deeper insights and used a bar chart to illustrate the relationship between mean_InscClaimAmtReimbursed and PotentialFraud. The analysis revealed that suspicious providers tend to have higher average insurance claim reimbursement amounts compared to genuine providers.



2.4 Feature Selection Process

The feature selection process aimed to identify the most relevant variables for predicting potential fraud, and determining which features to incorporate into our final dataset. Throughout all three assignments, we consistently adhered to the following key steps:

1. **Exploratory Data Analysis:** Initial exploration to understand variable distributions and relationships and gain valuable insights, using statistical summaries and visualizations (histograms, box plots, bar charts, correlation analysis, etc.)

2. **Outlier detection:** Outliers were identified through box plots and statistical distributions. They were retained as they could provide key fraud indicator information.

3. **Scaling features:** Scaling was employed in model K-Nearest neighbors and Naive Bayes model. Scaling was deemed necessary for these models due to their sensitivity to the magnitude of feature values, ensuring that all features contributed equally to the model's predictions and preventing bias towards features with larger scales. However, for the Decision Tree and Support Vector Machine (SVM) models, feature scaling was not applied as they are not sensitive to feature magnitude.

4. **Feature Importance from Models:** Decision Tree algorithm were employed to rank features based on their importance, details in section 4.2.

3 Building Classification Models

In this assignment, I used four models to find the best one for detecting fraudulent interactions, employing the algorithms SVM, Naive Bayes, Decision Tree, and KNN classifier on the dataset.

Below is a comparison between the different models and the performance measures for the best one.

3.1 Performance Indicators

In this dataset, the number of frauds is very less which introduces the problem of class imbalance. Thus, 'accuracy' won't be the correct metric to measure the performance of machine learning models. The model performances were evaluated based on several indicators below:

1. Precision (Positive Predictive Value): Precision is crucial for minimizing false positives, ensuring that identified cases are indeed fraudulent.

2. Recall (Sensitivity): Recall is essential for minimizing false negatives, ensuring that all actual fraud cases are identified.

3. F1 Score: The F1 score provides a balanced assessment, considering both precision and recall. It's particularly useful in scenarios with imbalanced classes.

We also use Area Under the Receiver Operating Characteristic Curve (AUC-ROC) in the best model performance to provide an overall assessment. It offers a comprehensive evaluation by assessing the model's ability to discriminate between positive and negative instances and provides insights into the model's overall performance across various thresholds.

3.2 Decision Tree

Settings and Parameter Optimization tunning

- Algorithm: Decision Tree using the CART method.

- Parameter Tuning: Use caret to train the rpart decision tree using 10 fold cross; validation using method "cv" resampling method and use 15 values for parameter tuning.

  Despite varying minsplit values (20, 10, and 5) while keeping cp at 0.001 and maxdepth at 10, the confusion matrix results remained consistent. This suggests that further adjustments to minsplit do not yield additional predictive value due to the sufficient complexity of the decision tree model.

Table 2: Comparison Between Different parameters

cp	maxdepth	Recall	Precision	F1 Score
0.01	10	0.6767	0.7544	0.7134
0.001	10	0.8321	0.8665	0.8489
0	5	0.6851	0.4856	0.5683

3.3 Support Vector Machines (SVM)

Settings and Parameter Optimization tunning

- Algorithm: Support Vector Machine uses different kernel: linear kernel, polynomial kernel, radial kernel
- Parameter Tuning: First , I performed grid search over cost in model uses linear kernel. The best model was selected based on its performance on the validation set, and its effectiveness was evaluated using the test dataset.

  Then I tuned the SVM model with a polynomial kernel by conducting a grid search over various degrees (3, 4, 5) and coef0 values (0.001, 0.01, 0.1, 1, 5, 10) using cross-validation to identify the parameter combination that provides the best model.

  Finally The SVM model with a radial basis function (RBF) kernel was optimized through a grid search over different gamma values (0.001, 0.1, 0.5, 1, 5, 10), utilizing cross-validation to select the hyperparameter that achieves the highest predictive performance.”

Table 3: Comparison Between Different parameters

kernel	Recall	Precision	F1 Score
Linear	1	0	0
Polynomial	0.6667	0.4286	0.5217
radial	0.9231	0.3428	0.5

3.4 K-Nearest Neighbors

Settings and Parameter Optimization tunning

- Scale variables before splitting the dataset into training and testing sets.
- The algorithm: The model uses the mlr3 package and the classif.kknn algorithm for classification.
- Parameter Tuning: The kNN model was tuned by iterating over different values of k (1, 25, 50, 75, 100) and evaluating its performance using 5-fold cross-validation with ”cv” resampling method.

Table 4: Comparison Between Different parameters

k	Recall	Precision	F1 Score
1	0.7434	0.8009	0.7711
25	0.7794	0.7908	0.7851
50	0.7935	0.7761	0.7847
75	0.8035	0.7650	0.7838
100	0.8105	0.7557	0.7821

3.5 Naïve Bayes

Settings and Parameter Optimization tunning

- Scale variables before splitting the dataset into training and testing sets.
- The algorithm: The model uses the mlr3 package and fitting a Naïve Bayes classifier using the classif.naive_bayes algorithm.
- Parameter Tuning: Grid search is employed to explore the parameter space and optimize the performance of the Naive Bayes classifier and 5 fold cross-validation using ”cv” resampling method to maximize classification F-beta score.

  I then define a tuning instance that combines the task, learner, resampling method, evaluation metric, and parameter space. The tuning will run for a maximum of 10 evaluations.

  After tuning, the best parameter values are extracted and used to train the Naive Bayes model on the full dataset.

  The model is then evaluated on a separate test dataset to assess its performance.

After tuning the Naive Bayes model with the Laplace smoothing parameter (laplace) ranging from 0 to 2, it was found that all values of laplace resulted in consistent performance metrics:

F-beta Score: 0.755

Accuracy: 0.6370

This indicates that the model’s performance stabilized across the parameter range, suggesting robustness to variations in the Laplace parameter.

- Precision = 0.2023
- Recall = 0.56887
- F1-score = 0.2985

4 Model Comparison

4.1 Selecting the Best Performing Model

Table 5 represent the performance measures from the optimal model for each method mentioned above.

Model	Recall	Precision	F1-score
SVM	0.9231	0.3428	0.5000
k-NN	0.8105	0.7557	0.7821
Naïve Bayes	0.5689	0.2023	0.2985
Decision Tree	0.8321	0.8665	0.8489

Table 5: Performance Metrics for Various Models

SVM Kernel Radial: This model has a high recall (0.9231) but low precision (0.3428), which means it is good at identifying positive cases but also produces a lot of false positives. The F1-score is 0.5, indicating a balance between precision and recall is not optimal.

KNN (k=100): This model has relatively high recall (0.8105), precision (0.7557), and F1-score (0.7821), indicating a better balance between identifying positive cases and limiting false positives.

Naïve Bayes: This model has the lowest recall (0.56887), precision (0.2023), and F1-score (0.2985) among the four models, indicating it is not the best choice for this task.

Decision Tree (cp=0.001, maxdepth=10): This model has the highest precision (0.8665) and F1-score (0.8489), and a fairly high recall (0.8321). This suggests it is the best model at both identifying positive cases and limiting false positives, and it achieves a good balance between precision and recall.

Confusion matrix

		Actual	
	fraud.predict	0	1
Predict	0	92568	8510
	1	11146	55240

Table 6: Confusion matrix for Decision Tree Model

From confusion matrix Table 6, we observe that the Decision Tree model effectively classifies non-fraudulent cases with high accuracy, as evidenced by the large number of true negatives (92,568) and true positives (55,240). The model's ability to correctly identify fraudulent cases is also noteworthy, as it identifies a significant

AUC Curve:

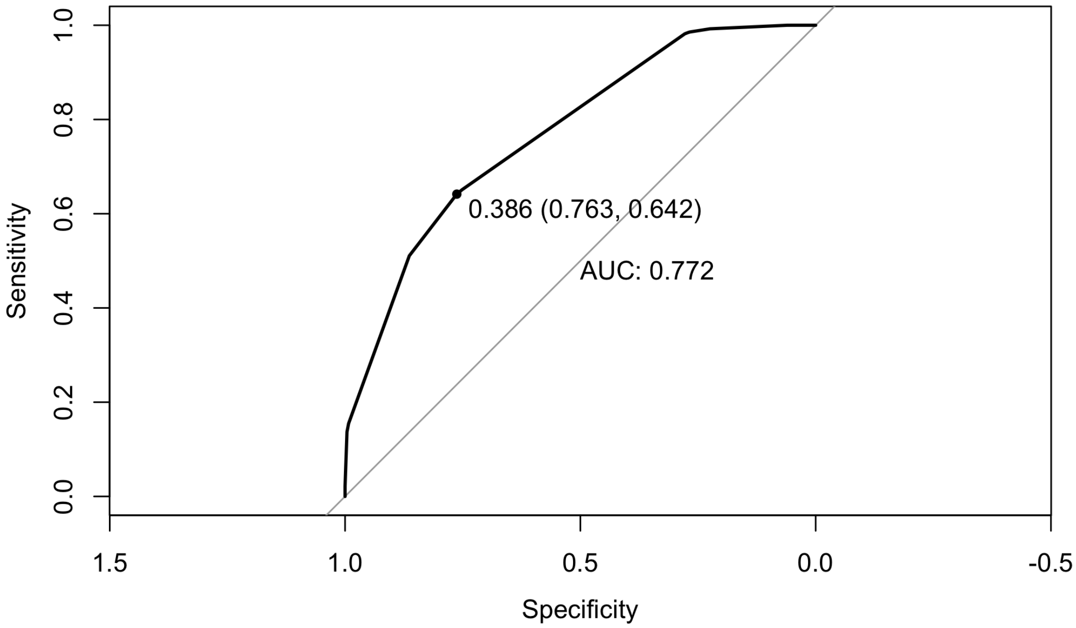


Figure 1: AUC Curve of Decision Tree Model

From Figure 1, we can see the Area Under the Curve (AUC) is 0.772, indicates that the model has a good ability to distinguish between fraudulent and non-fraudulent cases.

The point marked on the curve with coordinates (0.386, 0.763) suggests that the decision tree model has approximately 76.3% sensitivity (true positive rate) and 61.4% specificity (true negative rate) in predicting potential fraud.

4.2 Feature Importance Analysis

To understand the key factors distinguishing fraud and non-fraud providers, I examined model feature importances from the best-performing model.

Method varImp() is used to assess the importance of variables in a predictive model. Variable mean_InscClaimAmtReimbursed is the most contribute to the model's classification, and following is mean_TotalReimbursedAmt, details as below:

Variable	Value
AdmissionPeriod	3347.899936
Age	4.286926
count_BeneID	29.304428
County	24557.308788
mean_InscClaimAmtReimbursed	87485.533020
mean_TotalDeductibleAmt	77243.413961
mean_TotalReimbursementAmt	85612.372034
State	50906.274781
Race	0.000000
ClaimPeriod	0.000000

Table 7: Overall Summary

I also employed method filter.importance() to identify important features based on their importance scores. This method helps in ranking the features by their predictive power, confirming the findings from varImp().

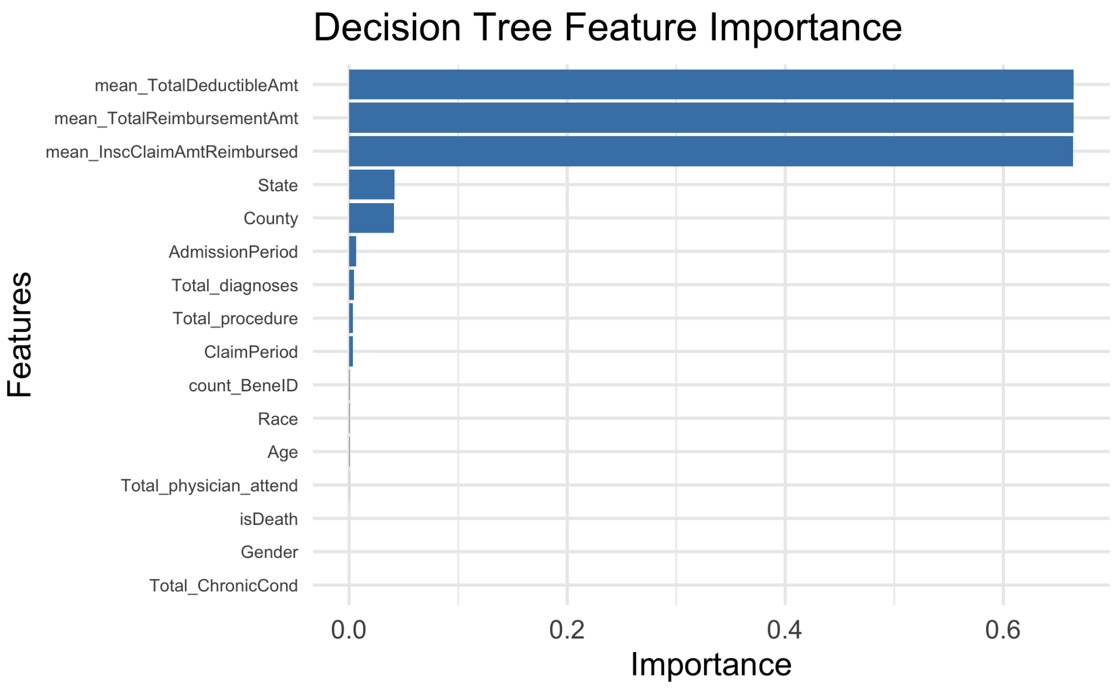


Figure 2: Feature Importance Decision Tree Model

Figure 2 presents the feature importance rankings from the Decision Tree models. The top three features are mean_InscClaimAmtReimbursed, mean_TotalDeductibleAmt, mean_TotalReimbursementAmt.

1. mean_InscClaimAmtReimbursed: This variable indicates the average amount reimbursed for insurance claims. High reimbursement amounts can be associated with fraudulent claims, especially if the amounts are unusually large compared to typical claims.
2. mean_TotalReimbursementAmt: This represents the average total amount reimbursed across all claims. Fraudulent activities often involve exaggerated or repeated claims to maximize reimbursements.
3. mean_TotalDeductibleAmt: This is the average deductible amount associated with claims. Fraudulent claims might exhibit patterns such as unusually high or low deductibles to avoid detection or to game the reimbursement process.

These variables are crucial for fraud detection, directly relating to financial transactions in healthcare services, which are frequent targets of fraud. Unusual patterns or outliers in these variables, such as high reimbursement amounts and specific geographic locations, often indicate suspicious behavior. By focusing on these key variables, we can significantly improve the accuracy and robustness of our fraud detection models.

5 Conclusion

At the outset of my analysis, I focused on the beneficiaries and patients in the dataset. Through data exploration, I uncovered several key insights:

There are certain beneficiaries, as listed below, who might be currently experiencing fraud or who appear more vulnerable to fraudulent activities.

- Patients associated with high reimbursement claims.
- Patients who have incurred substantial deductible payments.
- Some of these patients also exhibit a high number of chronic conditions.

Following this, I examined the characteristics of fraud and non-fraud providers in both the inpatient and outpatient datasets. This investigation revealed the following key distinctions between the two groups, as shown in Table 8:

Possibly Fraud Providers	Non-Fraud Providers
High average claim reimbursement amounts; some of these providers have the highest reimbursement amounts in the dataset	Low average claim reimbursement amounts
High average insurance reimbursement amount claims.	Low average insurance reimbursement amount claims.
A narrow range of patient age.	A wider range of patient age.

Table 8: Comparison of Possibly Fraud and Non-Fraud Providers

Another important observation is that potentially fraudulent providers might be more prevalent in specific states and counties. Factors such as a patient's age falling within a particular range, their geographic location (state/county), the total amount of their claims, and their primary healthcare provider can sometimes increase their susceptibility to fraud. These features can also assist investigators in distinguishing between fraudulent and non-fraudulent providers.