Medical insurance claim approval prediction using statistical models

Abstract

This project centers on predicting insurance claim approvals, a vital aspect impacting both patients and healthcare providers. Utilizing data from 1,338 policyholders and examining attributes such as age, gender, BMI, daily activity, location, insurance charges, and smoking habits, our aim is to discern connections among these factors. The primary objective is to forecast whether insurance claims will be approved. Leveraging statistical methodologies like LM, Random Forest, and Gradient Boosting, we seek to predict patient behavior concerning claim approvals, empowering more informed decision-making in the healthcare domain. The analysis revealed that among the models evaluated, Random Forest emerged as the most robust, demonstrating superior performance in Precision, Recall, Accuracy, and F1 Score. Following closely, XGBoost exhibited competitive performance across all metrics, highlighting its efficacy. However, in this specific evaluation, Linear Regression (LM) showcased comparatively lower values across the assessed metrics.

Introduction

In the dynamic landscape of healthcare, effective management of medical insurance claims stands as a pivotal axis, impacting both patients and providers. The ability to forecast insurance claim approvals holds paramount significance, optimizing resource allocation, refining administrative workflows, and ensuring seamless healthcare delivery. This study delves deep into predictive analytics, employing statistical tools like LM (Linear regression Models), Random Forest, and Gradient Boosting techniques. Our goal is to predict insurance claim approvals based on policyholder behavior, empowering smarter decision-making processes within the industry.

By leveraging methodologies like LM, Random Forest, and Gradient Boosting, this study navigates the complex terrain of predictive analytics, specifically addressing the challenge of forecasting patient behavior in insurance claim approvals. This pursuit not only fortifies the financial resilience of healthcare providers but also catalyzes an evolution within the industry, fostering proactive strategies and optimized resource utilization.

Linear regression, a foundational statistical technique, serves as a fundamental method for modeling the relationship between a dependent variable and one or more independent variables. This approach aims to predict a continuous outcome by fitting a linear equation to the observed data points. The model estimates the effect of each predictor variable on the target variable, allowing for interpretations of the relationship's strength and direction. Linear regression assumes a linear relationship between the variables and presupposes that the errors are normally distributed with constant variance. It is widely employed in various fields such as economics, social sciences, and natural sciences to understand and predict numerical outcomes based on given explanatory variables.

Random Forest, an ensemble learning technique renowned for handling large, complex datasets, demonstrated exceptional prowess in predicting medical outcomes, as highlighted by Khalilia, M., Chakraborty, S., and Popescu, M.(Khalilia, M., Chakraborty, S., and Popescu, M., 2011). Their study utilized the Healthcare Cost and Utilization Project (HCUP) dataset to predict disease risk based on individuals' medical histories. Employing the National Inpatient Sample (NIS) data and employed Random Forest within an ensemble learning framework to address the data's high imbalance. Comparing methodologies like support vector machines, bagging, boosting, and Random Forest, they found that Random Forest excelled in managing imbalanced data and determining variable importance, achieving an average AUC of 88.79% in predicting eight disease categories. This study underscores Random Forest's potential in handling intricate medical predictions, suggesting its viability for tasks like medical insurance claim prediction, showcasing superior performance compared to other models.

Similarly, Gradient Boost, an influential ensemble method, has showcased remarkable predictive prowess through iterative improvements in model performance, as evidenced by Akbar, N. A. et al. (Akbar, N. A., Sunyoto, A., Arief, M. R., & Caesarendra, W., 2020). By employing random forest regression and Extreme Gradient Boosting (XGB), the study delves into

state-of-the-art techniques to detect and prevent fraud. Notably, the XGB Tree method, with random sub-sampling, achieves an outstanding 86% overall accuracy and an impressive 87% in identifying illegitimate providers. Comparatively, the XGB method demonstrates higher accuracy, especially in meticulously tuned clean data, surpassing the Random Forest Method's 81% accuracy and highlighting its potential in combating healthcare fraud.

In this project, we aim to contribute to the existing body of knowledge by conducting a comparative analysis of linear regression, Random Forest, and Gradient Boost methods in the context of predicting medical insurance claim approval. Through empirical evaluation and comparison of these methodologies, we seek to provide insights into their applicability and performance within this specific domain.

Data Description

The medical insurance information data contains 1339 rows * 9 columns, which is from Kaggle website: https://www.kaggle.com/datasets/easonlai/sample-insurance-claim-prediction-dataset/data .

A detailed data description is shown below:

- Age: The age of the policyholder, presented in years. This variable signifies the policyholder's age at the time of data collection.
- Sex: Gender classification of the policyholder. It's encoded as a binary variable, where 0 represents female policyholders and 1 represents male policyholders.
- BMI (Body Mass Index): BMI is a measure that indicates body weight relative to height, calculated as weight in kilograms divided by the square of height in meters (kg/m^2). It offers insight into whether an individual's weight is considered relatively high, low, or normal in relation to their height. The ideal BMI range is typically considered to be between 18.5 and 25.
- Steps: Average daily walking steps taken by the policyholder. The unit of measurement is steps per day, reflecting the average physical activity level of the individual.
- Children: The number of children or dependents of the policyholder. This variable quantifies the policyholder's familial responsibilities.
- Smoker: Indicates the smoking status of the policyholder. It's encoded as a binary variable, with 0 denoting non-smokers and 1 representing smokers.
- Region: Categorization of the policyholder's residential area within the United States. It's represented as a categorical variable with four levels: northeast (0), northwest (1), southeast (2), and southwest (3).
- Charges: Individual medical costs are billed by the health insurance company. The unit of measurement for this variable is typically in currency (e.g., USD), representing the amount billed for medical services or treatments.
- Insurance Claim: This variable determines whether an insurance claim was approved by the insurance company or not. It's encoded as a binary variable, with 1 indicating that an insurance claim was approved by the insurance company and 0 indicating not.

Goal

The primary aim of this project is to utilize a comprehensive dataset from a health insurance company, encompassing 1,338 policyholders and 6 critical attributes: age, gender, BMI, daily walking steps, residential location, insurance charges, and smoking status. The central objective revolves around predicting the approval status of insurance claims by the company. This predictive effort emphasizes honing accuracy by dissecting intricate relationships and interactions among these attributes. Through the construction of several predictive models and meticulous comparative analyses, we aim to discern robust patterns to anticipate insurance claim approvals efficiently.

Statistical Methods

Data Import and Preliminary Checks:

The initial phase involves importing the dataset into the statistical environment and conducting exploratory checks for missing values. Visual examination through histograms, boxplots, and density plots helps understand feature distributions and identify potential outliers.

Feature Distribution and Normality Check:

Numerical feature distributions are visually inspected through histograms or density plots to assess conformity to normality assumptions. Formal normality tests, like norm qq plots, are employed for statistical evaluation. When needed, transformations like log transformations are applied for improved distributions.

Conversion of Categorical Variables to Continuous:

Categorical variables are transformed into a numerical format suitable for analysis. Techniques like one-hot encoding or creation of dummy variables are employed for compatibility with analytical algorithms.

Data Splitting:

The dataset is split into training and testing sets, typically following an 80-20 ratio. The training set aids model development, while the testing set evaluates predictive performance. The ratios of 0 and 1 values in train test sets are also checked.

Variance Inflation Factor (VIF) Calculation:

VIF computation helps assess multicollinearity among numerical variables. High VIF values indicate problematic multicollinearity, prompting the elimination of correlated variables to enhance model stability and interpretability.

Model Methodologies:

Linear regression Models (LM):

Linear regression (LM) is a statistical method based on the principle of fitting a linear equation to observed data to model the relationship between a dependent variable and one or more independent variables. The basic mathematical concept behind LM involves constructing a linear equation that represents the relationship between variables. The objective of LM is to estimate the values of weight coefficient of factors that minimize the sum of squared differences between observed and predicted values (i.e., minimize the residuals or errors). This is typically achieved using the method of ordinary least squares (OLS), aiming to find the best-fitting line that describes the linear relationship between the variables. The coefficients are estimated to maximize the fit of the model to the observed data, allowing for predictions of the dependent variable based on known values of the independent variable(s).

Random Forest:

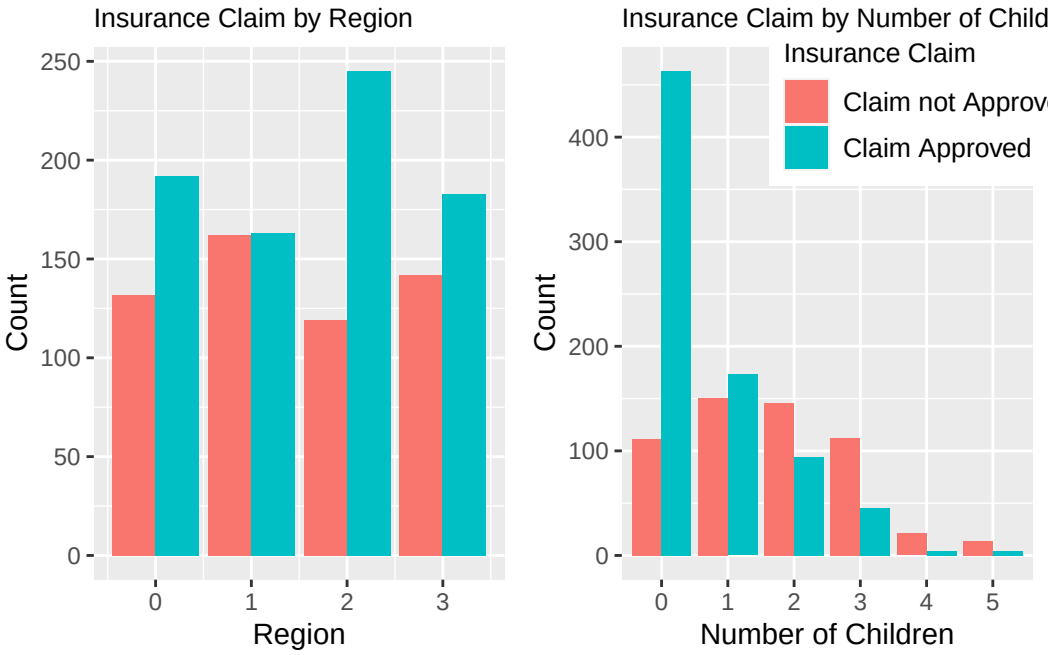
Random Forest is an ensemble learning method that builds a multitude of decision trees and merges their predictions to enhance overall accuracy and reduce overfitting. It excels in capturing complex relationships within data and is less prone to overfitting than individual decision trees. The algorithm introduces randomness by selecting random subsets of features for each tree, contributing to improved generalization. Random Forest aggregates predictions through a majority vote, making it robust and resilient to noise. The classification process involves constructing a forest of decision trees, with each tree casting a vote for the most likely class. The final prediction is determined by the majority of votes. Due to its robustness, capability to handle complex interactions, and resilience to overfitting, it's selected to capture non-linear relationships in the data.

Gradient Boosting:

Gradient Boosting is another ensemble method that iteratively builds decision trees, each focusing on the errors of its predecessors. It sequentially improves the model by minimizing the residual errors, resulting in a strong predictive algorithm. In our analysis, we will use the XGBoost (Extreme Gradient Boosting) variant, known for its efficiency and speed. The algorithm minimizes a loss function by adding weak learners (trees) in a sequential manner. Mathematically, the XGBoost algorithm minimizes the loss function. This method is employed for its capability to improve predictive accuracy through iterative learning, making it suitable for refining the model's performance.

By employing these diverse methods, the aim is to explore their strengths and weaknesses in predicting insurance claim filings. LM provides interpretability, while Random Forest and Gradient Boosting offer more complex modeling techniques that might capture intricate relationships within the data. The comparison among these methods will provide insights into their performance and suitability for the specific classification task at hand.

Results




```
TableGrob (1 x 2) "arrange": 2 grobs
  z     cells    name           grob
1 1 (1-1,1-1) arrange gtable[layout]
2 2 (1-1,2-2) arrange gtable[layout]
```

Figure 1. Insurance claim approval by the region of the policyholder and the number of children the policyholders have.

From figure 1, it is evident that region 2 exhibits the highest count of claim non-approval among policyholders. Furthermore, policyholders without children tend to experience a higher rate of claim rejection by the insurance company. This graphical illustration underscores the potential influence of geographical factors, such as the policyholder's region, alongside family composition, specifically the absence of children, on the likelihood of insurance claim approvals.

```r
#Check for normality
par(mfrow = c(2, 3))
for (var in numeric_vars) {
  qqnorm(df[[var]], main = paste("Q-Q Plot for", var))
  qqline(df[[var]], col = 2)
}
```

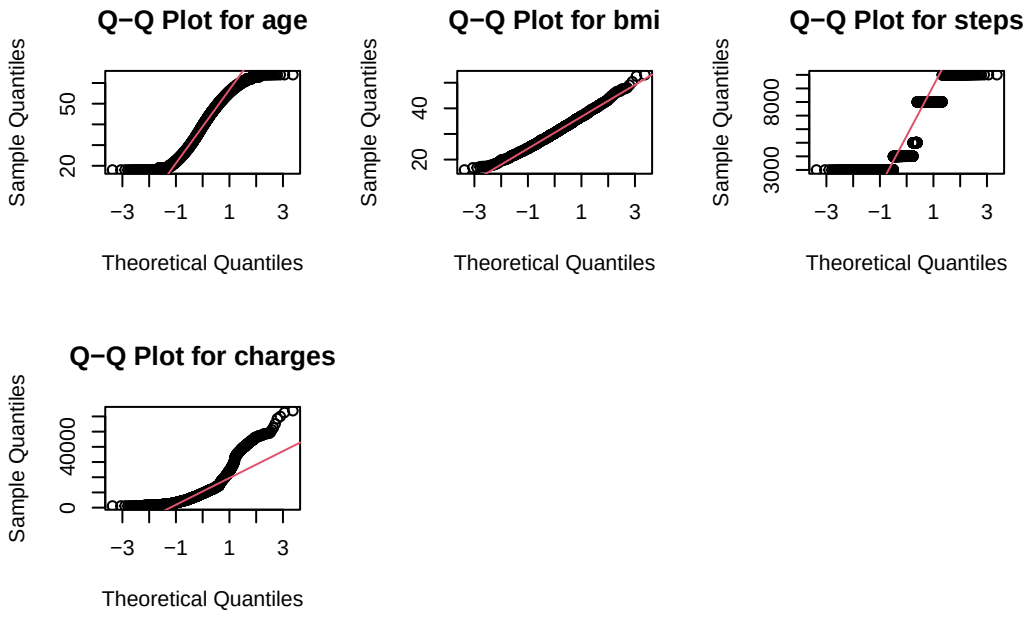


Figure 2. Normal QQplots for age, bmi, steps, children, charges to inspect their normality. The bmi and age are mostly following the line y=x which indicate the normal distribution of the data.

```
#Converting categorical variable to numerical
head(df1)
```

```
  age    bmi steps   charges insuranceclaim sex_0 sex_1 children_0 children_1
1  19 27.900  3009 16884.924              1     1     0          1          0
2  18 33.770  3008  1725.552              1     0     1          0          1
3  28 33.000  3009  4449.462              0     0     1          0          0
4  33 22.705 10009 21984.471              0     0     1          1          0
5  32 28.880  8010  3866.855              1     0     1          1          0
6  31 25.740  8005  3756.622              0     1     0          1          0
  children_2 children_3 children_4 children_5 smoker_0 smoker_1 region_0
1          0          0          0          0        0        1        0
2          0          0          0          0        1        0        0
3          0          1          0          0        1        0        0
4          0          0          0          0        1        0        0
5          0          0          0          0        1        0        0
6          0          0          0          0        1        0        0
  region_1 region_2 region_3
1        0        0        1
2        0        1        0
```

3	0	1	0
4	1	0	0
5	1	0	0
6	0	1	0

Table 1 shows the head of new data frame with all the categorical variable change into numerical.

```
Before removal of outliers, The dataset had 1338 samples.
```

```
After removal of outliers, The dataset now has 1191 samples.
```

The outliers are removed and the following analysis is based on the cleaned data.

```
# Check for correlation
pairs(train_set[, c("age", "sex_0","sex_1" , "bmi", "steps", "children_0", "children_1","c
```

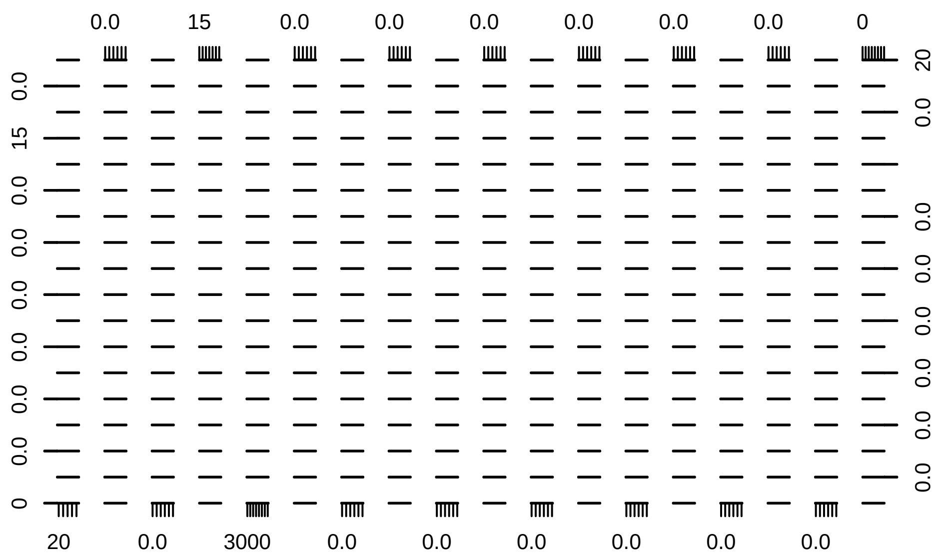


Figure 3 shows correlation between all features. There are high correlation between steps and bmi, smokers and charges and we applied LM with the interaction term to studied their prediction ability. The results are shown below:

```
there are higher-order terms (interactions) in this model
consider setting type = 'predictor'; see ?vif
```

```
          age            sex_0              bmi            steps
     1.619345         1.010154        10.090528        45.502533
   children_0       children_1       children_2       children_3
    16.443184        12.649407        10.146046         7.201628
   children_4         smoker_0         region_0         region_1
     2.123865        18.793166         1.564789         1.592682
     region_2          charges        bmi:steps smoker_0:charges
     1.591336        17.850770        25.751829        14.206959
```

After checking the variance inflation factors, the result shows that VIF 's corresponding to the interaction terms exceed 10 which indicate multicolineaity, thus we decided to use the model without interaction term.

```
summary(model.1)
```

```
Call:
lm(formula = insuranceclaim ~ age + sex_0 + bmi + steps + children_3 +
    children_4 + smoker_0 + region_0 + region_1 + region_2 +
    charges, data = train_set)

Residuals:
     Min       1Q   Median       3Q      Max
-0.98044 -0.34309  0.07143  0.31830  1.12172

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.772e-02  1.470e-01   0.665 0.506521
age          4.086e-03  1.210e-03   3.377 0.000762 ***
sex_0        3.551e-03  2.689e-02   0.132 0.894978
bmi          3.163e-02  3.775e-03   8.378  < 2e-16 ***
steps       -1.578e-05  8.513e-06  -1.854 0.064099 .
children_3  -4.387e-01  4.308e-02 -10.182  < 2e-16 ***
children_4  -5.030e-01  1.021e-01  -4.928 9.82e-07 ***
smoker_0    -5.434e-01  6.571e-02  -8.270 4.54e-16 ***
region_0     8.114e-02  3.918e-02   2.071 0.038636 *
region_1    -2.019e-02  3.833e-02  -0.527 0.598584
region_2    -7.415e-03  3.840e-02  -0.193 0.846919
charges     -5.814e-06  2.917e-06  -1.993 0.046533 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.414 on 941 degrees of freedom
Multiple R-squared:   0.32, Adjusted R-squared:  0.3121
F-statistic: 40.26 on 11 and 941 DF,  p-value: < 2.2e-16
```

```r
par(mfrow=c(2,2))
plot(model.1)
```

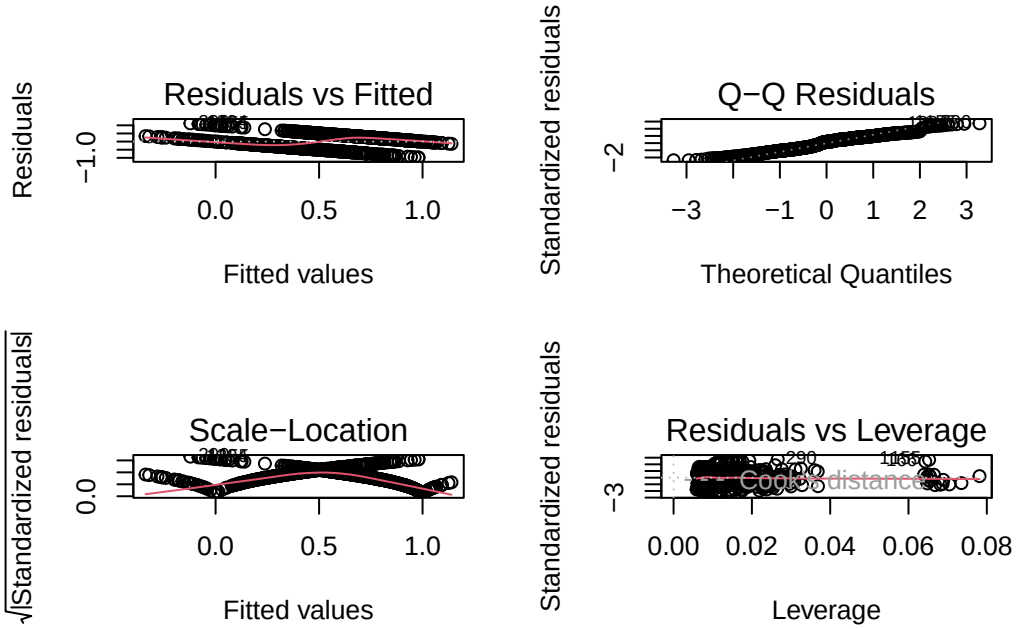


Figure 4 shows that the LM model with a F-stats=40.26 and R^2=0.32 suggests that this model is not a very good fit overall. Based on the p-values corresponding to t-statistics, some variables(age, bmi, children_0, region_0 and charges) are significant in explaining insurance-claim, while others seem ineffective. And the Residual vs Fitted plot indicate the no constant variance and the Normal Q-Q plot shows not normality. The Scale-Location plot shows that there is no outliers in this model, and the Residuals vs Leverage plot shows that there are some high leverage points.

```r
car::vif(model.1)
```

```
      age      sex_0        bmi      steps children_3 children_4   smoker_0
 1.601231   1.004639   2.747483   2.583871   1.012044   1.015153   2.393360
region_0   region_1   region_2    charges
 1.553953   1.573151   1.586961   2.406882
```

```r
#ROC curve
plot(roc_curve, main = "ROC Curve", col = "blue", lwd = 2)
abline(a = 0, b = 1, lty = 2, col = "red")  # Random classifier line
legend("bottomright", legend = paste("AUC =", round(auc_value, 2)), col = "blue", lwd = 2)
```

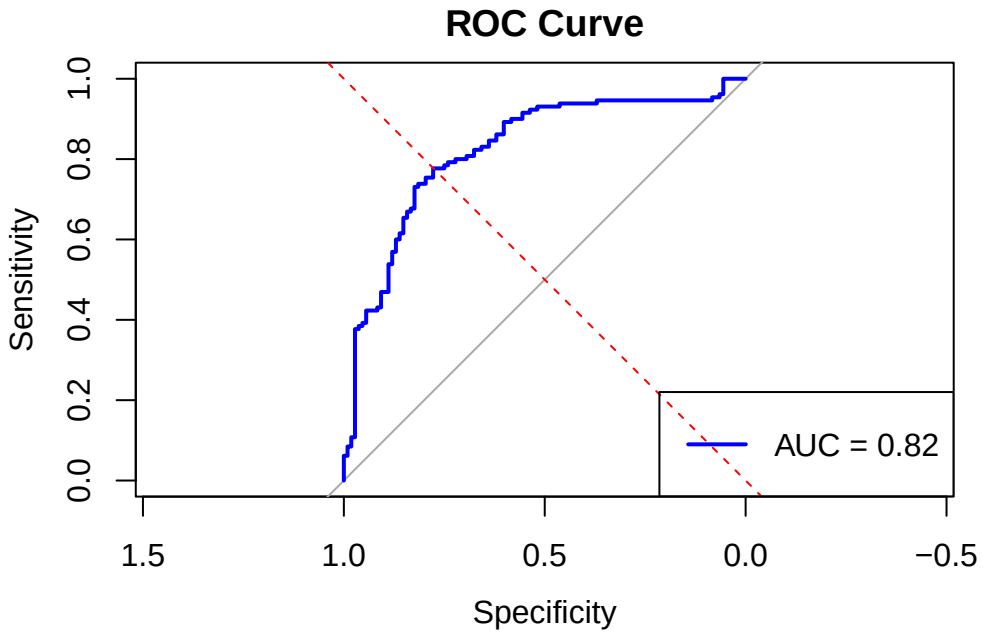


Figure 5 is the ROC curve for predicting test data under LM model.

```
[1] "LM Accuracy: 0.76890756302521"
```

```
[1] "LM Precision: 0.815126050420168"
```

```
[1] "LM Recall: 0.746153846153846"
```

```
[1] "LM F1 Score: 0.779116465863454"
```

```
          Model Precision    Recall  Accuracy   F1score
1            LM 0.8151261 0.7461538 0.7689076 0.7791165
2 Random Forest 0.8615385 0.8682171 0.8529412 0.8648649
3       XGBoost 0.8759124 0.8571429 0.8508065 0.8664260
```

Table 2. Model performance of the statistical models we used.

In Table 2, comparing the performance metrics—Precision, Recall, Accuracy, and F1 Score—of Linear regression Models (LM), Random Forest, and XGBoost reveals distinct patterns. Random Forest displayed the highest Precision (0.862), followed closely by XGBoost (0.876) and LM (0.815). Additionally, Random Forest and XGBoost exhibited slightly superior Recall values (0.868 and 0.857, respectively) compared to LM (0.746). In terms of Accuracy, Random Forest secured the highest value (0.853), succeeded by XGBoost (0.851) and LM (0.769). Both Random Forest and XGBoost demonstrated relatively higher F1 scores (0.865 and 0.866, respectively) compared to LM (0.779). Overall, the analysis highlights Random Forest as the most robust model, excelling in Precision, Recall, Accuracy, and F1 Score. XGBoost closely followed, demonstrating competitive performance across all metrics, while LM showcased relatively lower values in this specific evaluation.

Summary and conclusion

The Random Forest model surpasses the LM model in terms of accuracy, precision, recall, and F1 Score. The Random Forest model's high accuracy and balanced performance make it a robust choice for predicting medical insurance claim approvals. These results underscore the effectiveness of ensemble methods, such as Random Forest, in capturing complex relationships within the data, leading to improved predictive capabilities. Given the superior performance of the Random Forest model, it is recommended for deployment in real-world scenarios for predicting medical insurance claim approvals. Continuous monitoring and periodic model updates should be implemented to ensure optimal performance over time. Interpret ability of the Random Forest model may be enhanced through feature importance analysis to understand the key factors influencing predictions.

Reference

Akbar, N. A., Sunyoto, A., Arief, M. R., & Caesarendra, W., 2020. Improvement of decision tree classifier accuracy for healthcare insurance fraud prediction by using Extreme Gradient Boosting algorithm. In 2020 International conference on informatics, multimedia, cyber and information system (ICIMCIS) (pp. 110-114). IEEE.

Khalilia, M., Chakraborty, S., & Popescu, M., 2011. Predicting disease risks from highly imbalanced data using random forest. BMC medical informatics and decision making, 11, 1-13.