Predictive Analytics

UniSA

Name- Ankur Srivastava
Username-sRiay007

Table of Contents

Introduction *4*

Data Exploration and Feature Selection *5*

Building Classification Model *9*

Decision Tree: - **9**

KNN **11**

Naïve Bayes Algorithm **12**

Neural Networks **13**

Model Comparison and Conclusion *14*

Figure 1 Frequency of Claims 5
Figure 2 Claim Approval Rate 6
Figure 3 Correlation Matrix 7
Figure 4 Decision Tree 9
Figure 5 ROC Auc Graph 10
Figure 6 KNN Results 11
Figure 7 Naïve Bayes Result 12
Figure 8 Neural Networks Results 13

Introduction

In the first assignment, we embarked on a critical step towards developing a robust model to identify fraudulent behaviour in the healthcare insurance industry. The foundation of this Endeavor was laid by acquiring and meticulously examining four distinct training datasets. These datasets contained a wealth of information that was pivotal to understanding the patterns and characteristics of insurance claims. Our task was to sift through the abundant data and distil it down to the most salient features that could provide deeper insights into the behaviours of interest. To this end, we successfully merged the datasets and extracted key features that are often indicative of fraudulent activity. These features included the duration of insurance claims, the rate of claim approvals, the average amount claimed, and the frequency of claims—each a potential marker of anomalous behaviour.

In the second assignment, we progressed from data preparation to predictive modelling. Utilizing the features extracted previously, we constructed a decision tree model—a popular machine learning algorithm renowned for its interpretability and efficacy in classification tasks. The decision tree was trained to discern patterns that differentiate fraudulent claims from non-fraudulent ones, leveraging the inherent structure of the data. This step was critical in setting the groundwork for an analytical framework capable of identifying suspicious activities within healthcare insurance claims. The model's reliance on clear decision rules allowed us to make transparent predictions, an essential factor in the sensitive domain of fraud detection where accountability and explain ability are paramount.

Throughout these assignments, our journey has been marked by a methodical approach to tackling the complex issue of insurance fraud. By carefully selecting features that capture the essence of fraudulent behaviour and employing a decision tree model, we have made significant strides in creating a tool that can potentially save the healthcare insurance industry substantial financial resources. Our efforts have not only focused on achieving high accuracy in fraud detection but also on ensuring that our methods can withstand scrutiny and provide actionable insights. The culmination of this process equips us with a data-driven solution designed to enhance the integrity of insurance claims processes.

Data Exploration and Feature Selection

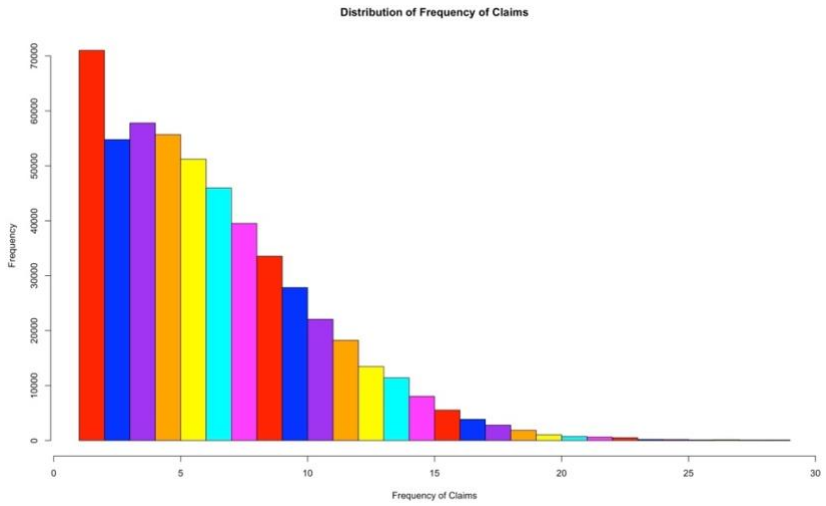

Figure 1 Frequency of Claims

The histogram illustrates the distribution of insurance claim frequencies, indicating how often claims are made. It reveals that most policyholders file a low number of claims, with the most common frequency being close to zero. The distribution is right-skewed, with a long tail that suggests a few policyholders file a much higher number of claims. This tail might be of particular interest in fraud detection efforts, as it could signify unusual activity. The decreasing heights of the bars as claim frequency increases confirm that higher frequencies of claims are less common. The varied colours of the bars do not seem to represent different categories but serve to differentiate between the frequencies visually. Overall, the graph points to a typical claim pattern with a small number of potential outliers.

However, in the context of insurance claims, if a very small number of beneficiaries or providers are responsible for an unusually high frequency of claims, this could be a red flag for possible fraudulent activities. These outliers could potentially be exploiting the system, and such cases may warrant a closer examination. Insurance fraud analysts often look for patterns that deviate from the norm, and a statistical outlier in terms of claim frequency might be just such a pattern.

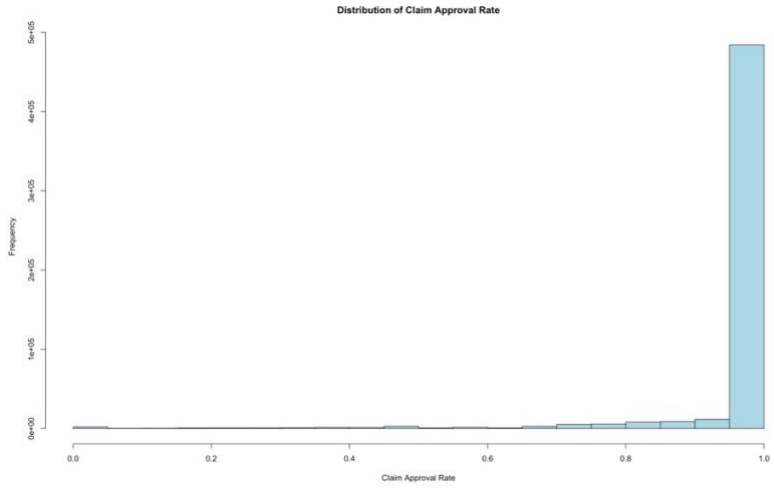

Figure 2 Claim Approval Rate

From the graph, we observe an overwhelming concentration of cases where the claim approval rate is at or near 1, which implies that most claims are approved. There is a very small number of cases with lower approval rates, as indicated by the negligible bars across the spectrum leading up to the peak at 1.

This distribution could suggest that there is a high propensity within the dataset for claims to be approved, or it could indicate that there is a default or systematic bias towards claim approval. However, without further context, it's difficult to ascertain whether this pattern is due to a genuinely high occurrence of legitimate claims, or if it potentially signals a lax approval process that could be exploited for fraudulent activities. It's also worth considering the role of data collection or entry processes, which might have influenced this distribution. For a more comprehensive analysis, it would be necessary to investigate the specifics of the claim assessment criteria, as well as to incorporate additional data regarding the claims themselves.

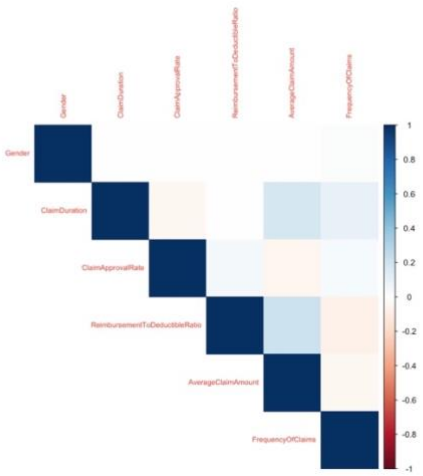

Figure 3 Correlation Matrix

The correlation analysis of your healthcare insurance dataset indicates that most variables do not have strong linear relationships with each other. Gender does not significantly correlate with any other variable, suggesting that it does not linearly influence the outcomes of interest. There is a moderate positive correlation between Claim Duration and Claim Approval Rate, hinting that longer claims might have a better chance of approval; however, the nature of this relationship is not clear-cut and requires further investigation. The Reimbursement to Deductible Ratio stands out as independent, lacking a linear relationship with the claim-related variables.

On the financial side, a slight negative correlation is observed between the Average Claim Amount and the Frequency of Claims, indicating that as the number of claims increases, the average amount per claim tends to decrease slightly. However, the Frequency of Claims is not strongly influenced by how long a claim lasts or its approval likelihood, suggesting that the number of claims submitted by a beneficiary is an independent characteristic in this context. Overall, while the correlation matrix reveals some interesting patterns, it does not provide comprehensive insights into fraudulent behaviour, emphasizing the need for more nuanced analytical methods to detect fraud in healthcare insurance data.

- The features to be selected were based on the correlation matrix as to how much would they be correlated to be able to find the fraudulent behaviour based on the selected variables.

- No, Outlier detection wasn't used before as there seemed to be no use of it before building the model however outlier detection will be done via use of predictive modelling such as svm in this project.

- Yes, I scaled my variables to normalize them so that all the variables are equally treated for building the model. Scaling allows for a fair comparison by bringing all variables to the same scale without distorting differences in the ranges of values. Algorithms that rely on distance calculations, such as k-nearest neighbours (KNN) and k-means clustering, can be biased towards features with a wider range if the data is not scaled. Algorithms that involve computations with very high or very low values may be unstable and result in numerical precision issues. Scaling helps to mitigate this by maintaining values within a numerically stable range.

- Yes, Correlation analysis was done for investigation the relationship between the variables which we have covered in the practical work.

- Yes, Correlation analysis was done for investigation the relationship between the variables which we have covered in the practical work.

Building Classification Model

Decision Tree:-

In this model we keep the CP= 0, maxdepth = 6, minsplit=2.

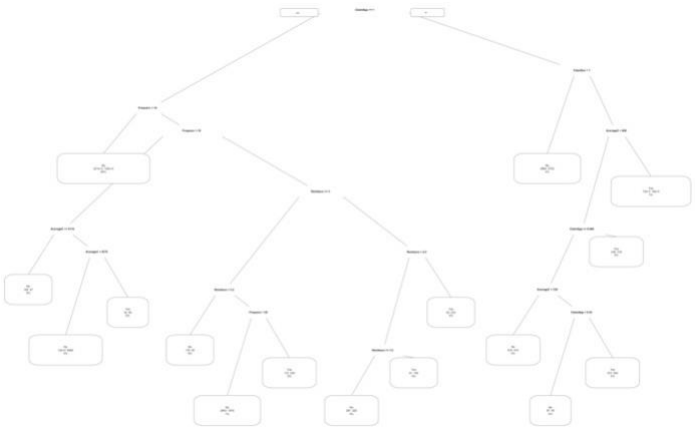

Figure 4 Decision Tree

This decision tree is built without pruning (CP = 0), resulting in a relatively complex tree. It prioritizes certain variables, particularly "ClaimApprovalRate," "ClaimDuration," and "AverageClaimAmount," as the most important features for making classification decisions. The summary provides a detailed view of the tree's structure and variable importance, which can be helpful for understanding how the model works and making predictions based on the tree.

- Accuracy: The model achieved an accuracy of approximately 62.94%, which means it correctly predicted the target variable (PotentialFraud) for about 62.94% of the cases in the test dataset.
- Precision: The precision of the model is around 63.36%. This metric measures the ability of the model to make accurate positive predictions (i.e., correctly identifying potential fraud cases).
- Recall (Sensitivity): The recall or sensitivity of the model is approximately 94.66%. This metric quantifies the ability of the model to correctly identify all actual positive cases (potential fraud), indicating that the model has a high true positive rate.
- F1-Score: The F1-score, which balances precision and recall, is around 0.76. It provides a single metric to assess the overall performance of the model, considering both false positives and false negatives.

ROC and AUC

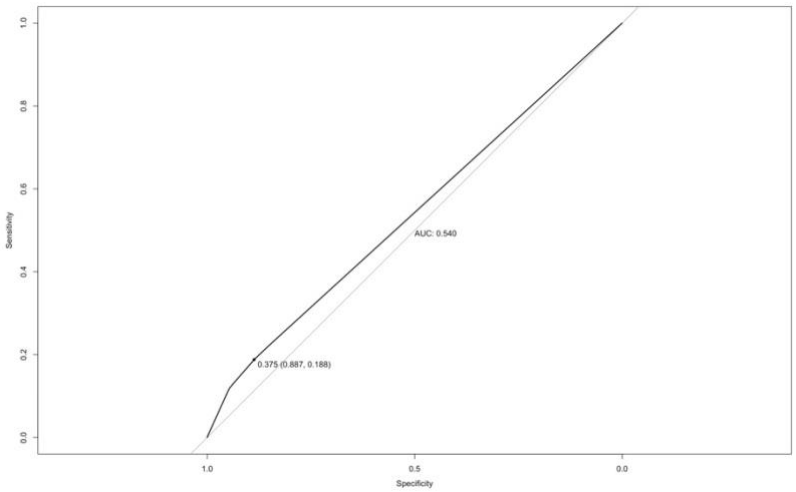

Figure 5 ROC Auc Graph

- ROC AUC (Area Under the Curve): The ROC AUC score is approximately 0.54. This metric is related to the model's ability to distinguish between positive and negative cases.
- Optimal Threshold: The optimal threshold for classification is approximately 0.36. This threshold is used to make binary predictions based on the predicted probabilities.
- Sensitivity (True Positive Rate): The sensitivity or true positive rate is approximately 22.80%. It quantifies the ability of the model to correctly identify positive cases.
- Specificity (True Negative Rate): The specificity or true negative rate is around 14.97%. It measures the ability of the model to correctly identify negative cases.

KNN

Where K=30

```
> # Print the confusion matrix
> print(conf_matrix_k30)
Confusion Matrix and Statistics

          Reference
Prediction    No   Yes
       No  58548 32892
       Yes  9881  9607

               Accuracy : 0.6144
                 95% CI : (0.6115, 0.6173)
    No Information Rate : 0.6169
    P-Value [Acc > NIR] : 0.955

                  Kappa : 0.091

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.8556
            Specificity : 0.2261
         Pos Pred Value : 0.6403
         Neg Pred Value : 0.4930
             Prevalence : 0.6169
         Detection Rate : 0.5278
   Detection Prevalence : 0.8243
      Balanced Accuracy : 0.5408

       'Positive' Class : No
```

Figure 6 KNN Results

The confusion matrix presented shows the performance of a binary classification model, presumably a K-Nearest Neighbours (KNN) classifier with k = 30. The matrix indicates the model has an overall accuracy of 61.44%, meaning it correctly predicts both fraudulent and non-fraudulent cases a little over half the time. The sensitivity, or true positive rate, is relatively high at 85.56%, indicating the model is good at identifying true non-fraudulent cases as such. However, the specificity, or true negative rate, is quite low at 22.61%, suggesting the model struggles to correctly identify true fraudulent cases.

The positive predictive value, or precision, is 64.03%, and the negative predictive value is 49.30%, which are moderate. The Kappa statistic of 0.091 is near zero, implying that there is little agreement between the predicted and actual values beyond what would be expected by chance. The balanced accuracy of 54.08% further demonstrates this disparity, reflecting the average of sensitivity and specificity and suggesting that the model does not perform well at classifying the positive class (fraudulent cases) compared to the negative class. Overall, while the model is relatively accurate, its utility may be limited by its low specificity and imbalanced predictive performance across classes.

- Accuracy (61.44%): - The KNN model with k = 30 neighbours correctly predicts the class label for approximately 61% of the cases in the test dataset.

- Misclassification Error (38.64%): - About 39% of the predictions made by the model are incorrect, which is a substantial proportion, indicating room for improvement.

- Precision (63.98%): - When the model predicts the positive class, it is correct approximately 64% of the time, suggesting that there might be a moderate number of false positives.

- Recall (85.49%): - The model successfully identifies about 85% of actual positive cases, indicating good sensitivity.
- Kappa Statistic (0.091): - A very low Kappa value indicates that the model's predictive ability is not much better than random chance.

Naïve Bayes Algorithm

```
> print(nb_results)
Confusion Matrix and Statistics

          Reference
Prediction    No   Yes
       No  87436 52248
       Yes 10729  8056

               Accuracy : 0.6026
                 95% CI : (0.6002, 0.605)
    No Information Rate : 0.6195
    P-Value [Acc > NIR] : 1

                  Kappa : 0.028

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.8907
            Specificity : 0.1336
         Pos Pred Value : 0.6260
         Neg Pred Value : 0.4289
             Prevalence : 0.6195
         Detection Rate : 0.5518
   Detection Prevalence : 0.8815
      Balanced Accuracy : 0.5121

       'Positive' Class : No
```

Figure 7 Naïve Bayes Result

- Accuracy: The model has an accuracy of approximately 60.26%, meaning that out of all the predictions made, about 60.26% of them were correct.

- 95% CI (Confidence Interval): The confidence interval for accuracy is between 60.02% to 60.5%, indicating that if the model's performance were assessed on multiple samples, the accuracy would fall within this range 95% of the time.

- No Information Rate: This rate is 61.95% and represents the accuracy that could be achieved by always predicting the most frequent class. The model's accuracy is slightly below this rate, suggesting that it is not performing much better than a naive guess.

- P-Value [Acc > NIR]: The p-value for testing whether the model's accuracy is significantly greater than the no information rate is 1, which is not below the common significance level (e.g., 0.05), indicating that the model's accuracy is not statistically significantly better than the no information rate.

- Kappa: The Kappa statistic is 0.028, which is very low, indicating that there is little agreement between the model's predictions and the actual values after accounting for the agreement that would be expected by chance alone.

Neural Networks

```
> # Print the confusion matrix results
> print(nb_results)
Confusion Matrix and Statistics

          Reference
Prediction    No   Yes
       No  87436 52248
       Yes 10729  8056

               Accuracy : 0.6026
                 95% CI : (0.6002, 0.605)
    No Information Rate : 0.6195
    P-Value [Acc > NIR] : 1

                  Kappa : 0.028

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.8907
            Specificity : 0.1336
         Pos Pred Value : 0.6260
         Neg Pred Value : 0.4289
             Prevalence : 0.6195
         Detection Rate : 0.5518
   Detection Prevalence : 0.8815
      Balanced Accuracy : 0.5121

       'Positive' Class : No
```

Figure 8 Neural Networks Results

- The model's accuracy is 60.26%, which means it correctly predicts the class for about 60% of the instances.
- Sensitivity is high (0.8907), indicating that the model is good at correctly identifying the positive class.
- Specificity is low (0.1336), suggesting that the model struggles to correctly identify the negative class.
- Positive Predictive Value (Precision) is 62.60%, indicating that when the model predicts the positive class, it is correct 62.60% of the time.

In summary, the model performs reasonably well in identifying the positive class but struggles with specificity, and the overall accuracy is influenced by the class imbalance. Depending on the specific goals and context, you might want to consider adjusting the model or using additional evaluation metrics.

Model Comparison and Conclusion

The decision tree model shows slightly higher accuracy and precision but has a significantly higher recall compared to the KNN, Neural Networks and Naïve Bayes models. The Naïve Bayes model has the lowest accuracy and kappa statistic, indicating its predictive performance is quite limited. The recall rate is particularly high for the Decision Tree model, suggesting it is the most sensitive in detecting true positive cases of fraud.

The Decision Tree model outperforms the other two models in terms of recall (sensitivity), indicating it is best at identifying true positives, which in this context means correctly identifying fraudulent claims. It also has the highest F1-score, which balances precision and recall, making it a robust choice when both false positives and false negatives are important to minimize.

The K-Nearest Neighbours (KNN) model with K=30 shows a good balance between precision and recall but falls slightly short of the Decision Tree model in terms of overall accuracy and significantly in terms of recall. It does, however, have a slightly higher precision than the Decision Tree model, which means it is slightly better at identifying true positives while minimizing false positives.

The Naïve Bayes model has the lowest performance in accuracy and kappa statistic, suggesting it may not be as effective for this problem. The low kappa value indicates that there is minimal agreement between the model's predictions and the actual data, beyond what would be expected by chance. Moreover, the P-Value indicating that the accuracy is not significantly better than the No Information Rate suggests that this model may not be useful in practice for detecting fraud compared to random guessing.