Table of Contents

Introduction

Research Question

Can we estimate what is the ***probability*** *to* ***gain a claim*** *regarding* ***building and construction contracts breaches*** *if the process is decided by an arbiter and goes all the way through a final determination?*

Target audience

This study has two target audiences. These are contract **claimants** and **respondents**. For this target audiences, the answer to this question can provide very important information associated to the **probability of winning the dispute**, the factors related to the increasing of this probability, and the results regarding to how much can the claimants get regarding what they initially claimed, and how much it would cost to them in average in case the clam is unsuccessful. The target audience may use this instrument as a reference model to assess the chances of getting a good outcome in a dispute and to increase the chances to get a good outcome using the most important features.

Dataset Introduction

Dataset Source and Description

This dataset is from
https://discover.data.vic.gov.au/dataset/building-and-construction-industry-security-of-payment-adjudication-activity-data

The VBA is a state regulatory entity created to control Victoria's building and plumbing industries. Hence, as sometimes there are disagreements arising over payment among constructors and contractors, the SOP Act has been designed to ease the process of due payment recovery under a construction contract in a fast and inexpensive manner without lawyers involved in the process.

Due to a lack of relevant expertise, we attempted to gain a preliminary understanding of the meaning of each feature in the dataset (see Appendix). During this process, we discovered many features that are irrelevant to the research question, as well as missing values and outliers. These issues will be addressed gradually in the subsequent steps

Dataset Attributes and Scope

In this context an open dataset has been provided by the government of Australia with information related to these payment disputes between 2021-05-11 and 2022-06-29. This dataset is composed of 45 fields and 324 samples of construction payments recovery processes. Within the document many of the fields are related to the SOP Act.

Data Wrangling and Preprocessing

Firstly, based on the research question, our target variable is to treat "**Adjudicated amount (ex GST)**" as a binary variable that is “0” standing for **successful claim** amount and “1” standing for **unsuccessful claim**. And the main features are related to claimants, responders, and information regarding the SOP Act in relation to the dispute process. Variables related to ANA and ANA’s Adjudicators (such as fee percentages, etc.) are not relevant to the research question as they are based on the arbitration outcome, making them less informative. Considering these factors, we decided to drop these columns.

Secondly, we employ the **Tukey** method to identify and mark outliers as missing values(NAN). Subsequently, we identify columns with high missing value rates and remove them to enhance the effectiveness of the dataset.

Thirdly, a **data quality** process was performed converting data types into a unified format. This resulted in three types of data: numeric (float), date/time (datetime), and categorical (object).

Fourthly, we proceed with transformations for certain columns:

The column "**Description of project and contract works**" contains a significant amount of text. We employ the Bag-of-Words (BOW) method combined with the K-Means algorithm to convert it into categorical variables. **Figure 1** demonstrates the application of the elbow method to determine the optimal number of clusters as 6.

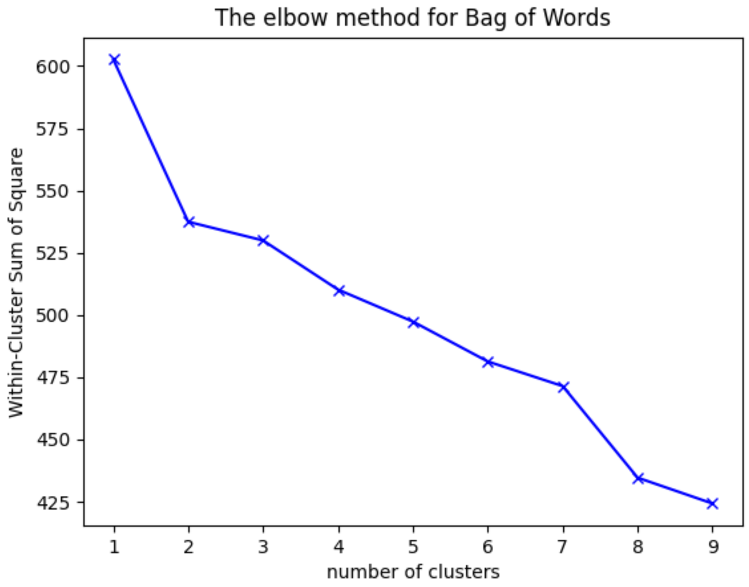


Figure 1: *Elbow method of Bag of Words*

The values in the four columns, "**Application Date**", "**Acceptance Date**", "**Determination Completion Date**" and "**Determination Released Date**" are in date format and are close in values, indicating that there may be a certain number of working days between them. Here, we plotted three graphs to investigate whether the year, month, and day of the acceptance date have a significant impact on the zero amounts of target variable. According to **figure 2** and **figure 3**, we ultimately select the month of acceptance and the arbitration duration as two features to replace the four previous date columns.

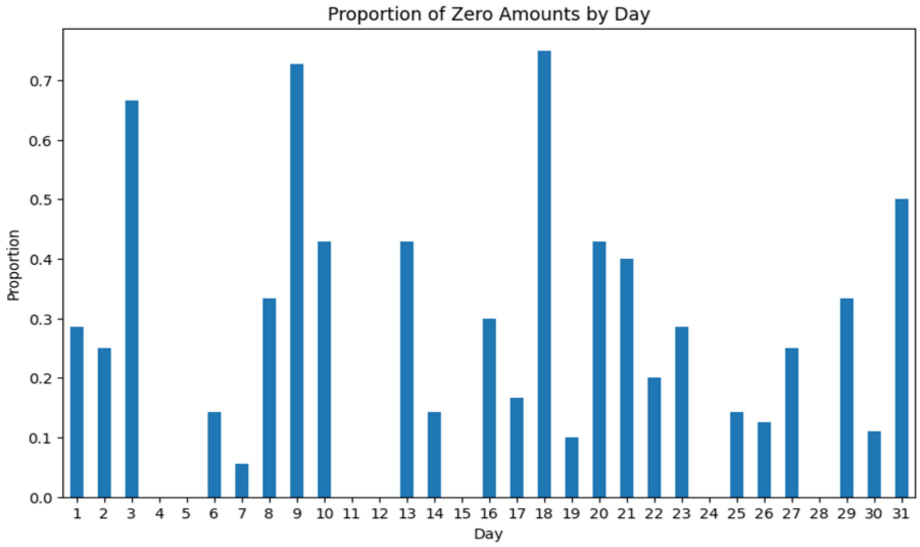


Figure 2: *Proportion of zero amounts by day*

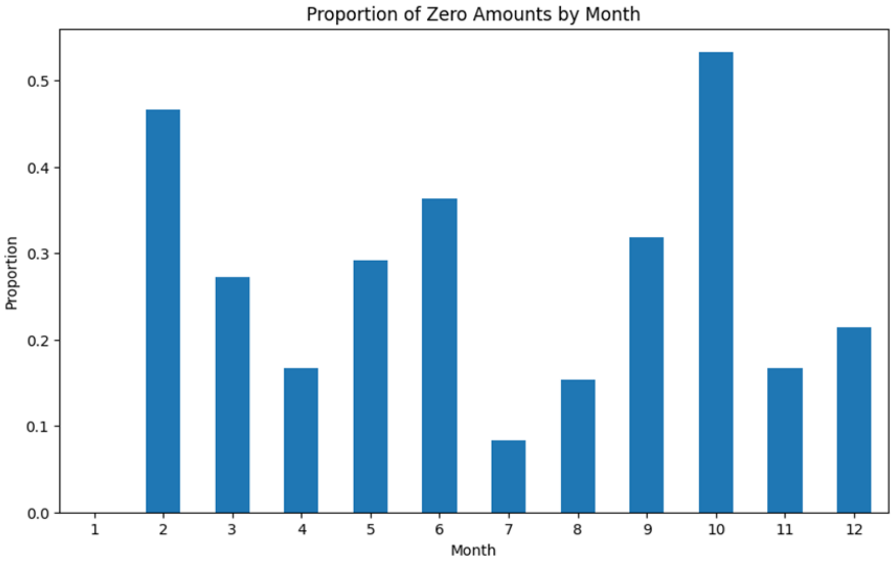


Figure 3: *Proportion of zero amounts by month*

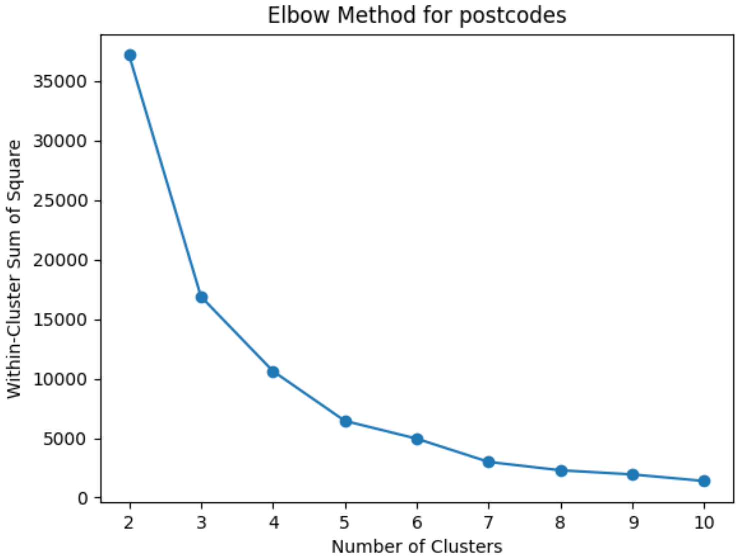


Figure 4: *Elbow method for Postcodes*

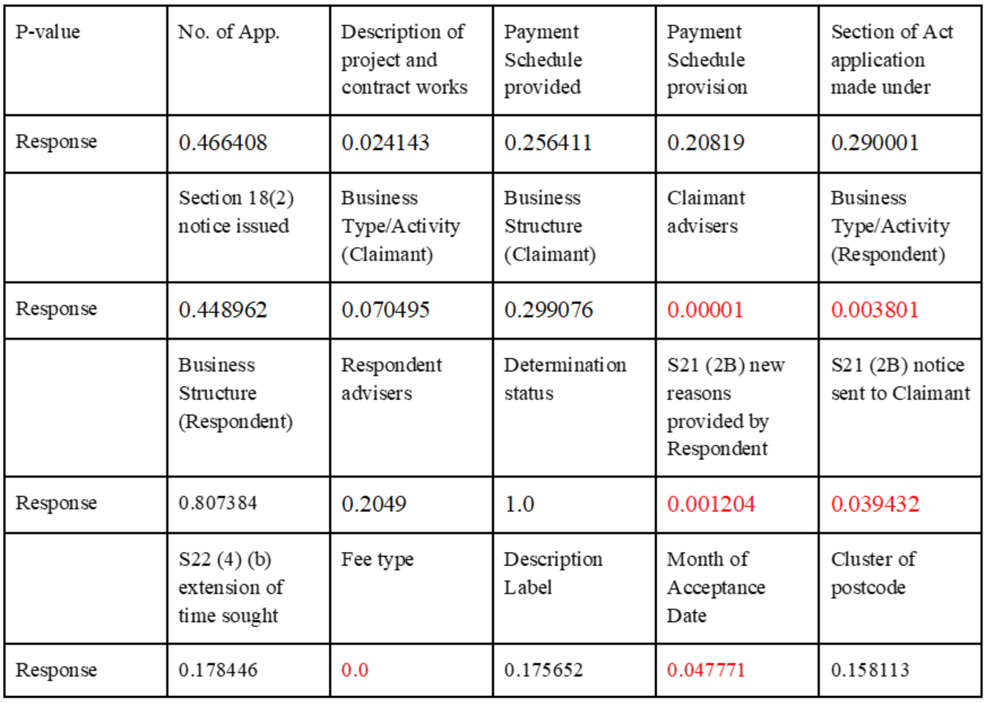
For the "Project postcode" column, since there are a bunch of different postcodes, directly converting it into a categorical variable we used the K-Means method to divide it into several groups and then convert the groups into categorical variables. We also used the elbow method provided by **Figure 4** to determine the optimal number of groups as 5.

Fifthly, for the remaining categorical variables, we group "empty" and "NAN" into a new category “**no_info**” within each respective column because there are many grouping variables containing missing values and imputation of all grouping variables would increase imputation error or generate bias and overfitting. Afterwards, we utilize **OneHotEncoder** to convert the modified column into dummy variables. As for the numerical variables, we employ **Multiple Imputation by Chained Equations (MICE)** imputation to handle missing values. Based on our understanding of the dataset, we applied data imputation over it as a prerequisite to run the machine learning models. The main reason to do this in a specific section was to guarantee the dataset quality previous to the modeling part.

Exploratory Data Analysis

Categorical Variables

The basic concept to identify the relationship between categorical covariates and response is use Chi-square test to find the significant variables and draw heatmaps to illustrate the information between them.



P-value	No. of App.	Description of project and contract works	Payment Schedule provided	Payment Schedule provision	Section of Act application made under
Response	0.466408	0.024143	0.256411	0.20819	0.290001
	Section 18(2) notice issued	Business Type/Activity (Claimant)	Business Structure (Claimant)	Claimant advisers	Business Type/Activity (Respondent)
Response	0.448962	0.070495	0.299076	0.00001	0.003801
	Business Structure (Respondent)	Respondent advisers	Determination status	S21 (2B) new reasons provided by Respondent	S21 (2B) notice sent to Claimant
Response	0.807384	0.2049	1.0	0.001204	0.039432
	S22 (4) (b) extension of time sought	Fee type	Description Label	Month of Acceptance Date	Cluster of postcode
Response	0.178446	0.0	0.175652	0.047771	0.158113

Figure 5: *p-value of Chi-square test*

According to the table above, p-value less than 0.05 can indicate that the relationship between covariate and response is significant at 0.05 significance level. It is very intuitive to find that five variables such as "Claimant advisers", "Business type/Activity (Respondent)" are significant.

Figure 6 shows that if the claimant adviser is solicitors or the claim preparer is more likely to claim the amount back.

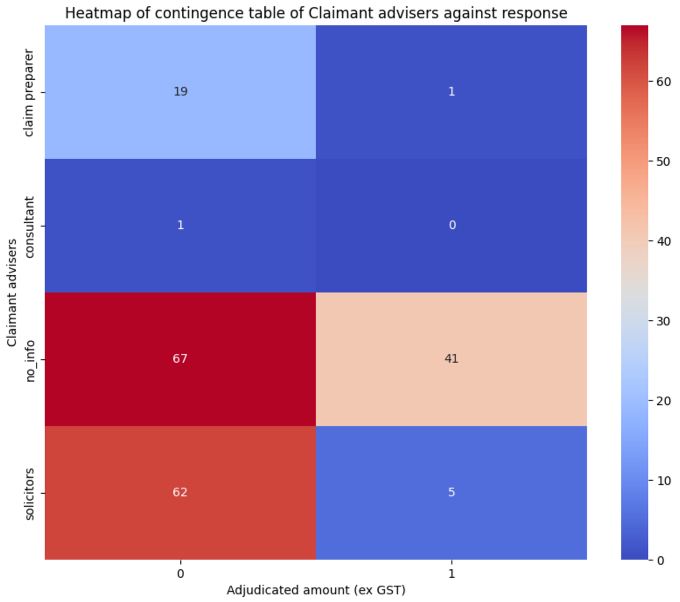


Figure 6: *Heatmap of Claimant Adviser against response*

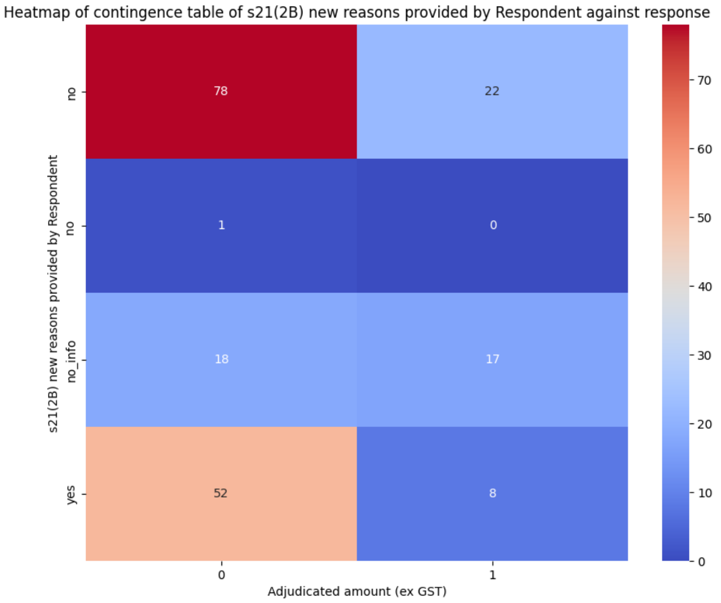


Figure 7: *heatmap of s21(2B) by respondent against response*

Figure 7 illustrates that if respondents provide new reasons based on s21(2B) it is more likely to claim the amount back.

Numerical Variables

In our dataset, we focus on two main numerical covariates: "Claimed Amount" and "Amount of Payment Schedule," which are essential for our analysis.

The observed plots indicate that the covariates roughly follow a normal distribution, but they are heavily skewed due to the presence of extreme values. This suggests the need for scaling or transformation before utilizing the covariates for further analysis.

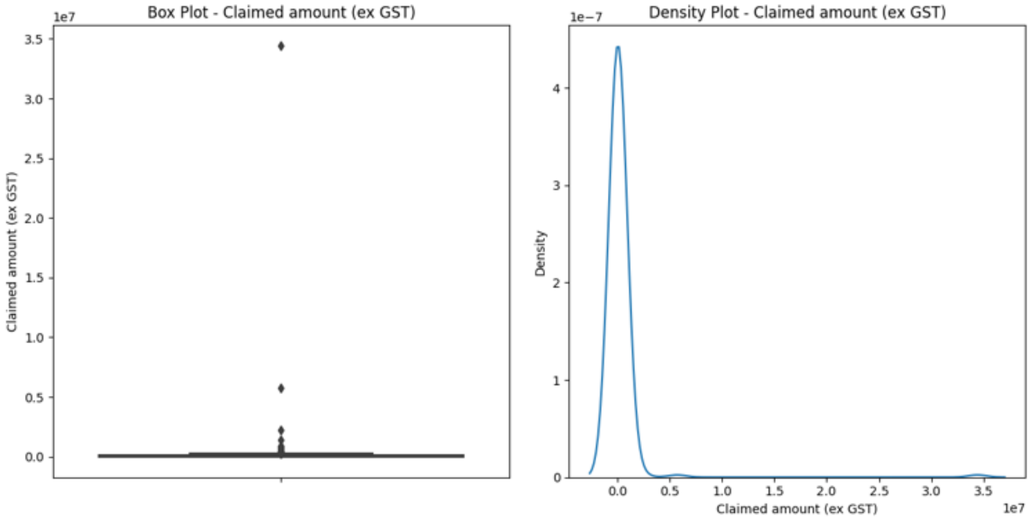


Figure 8: *Box-plot & Density plot for Claimed Amount column*

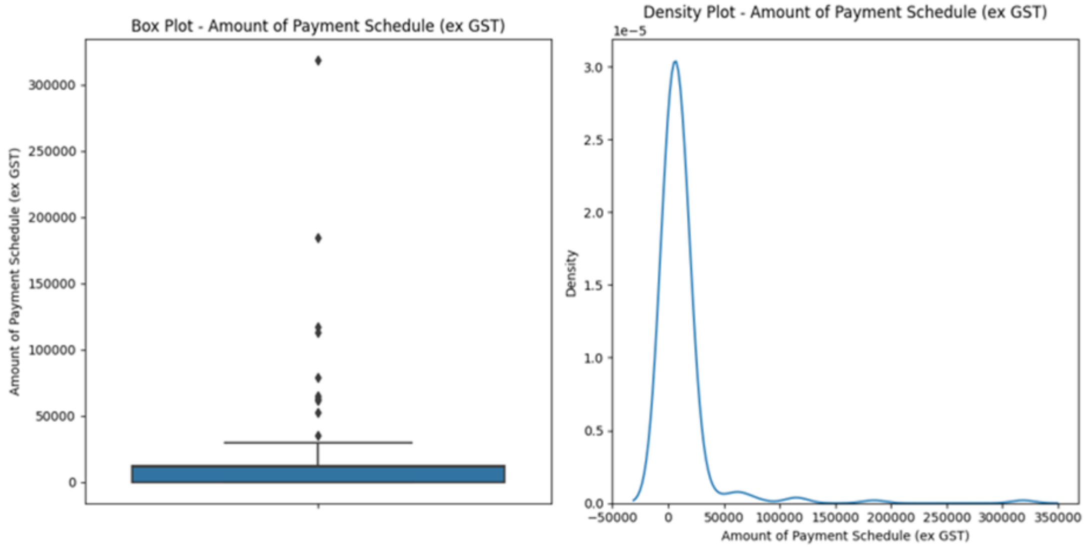
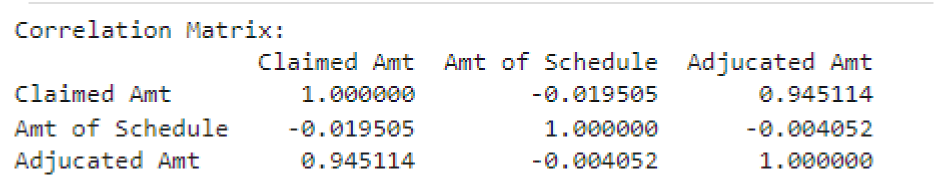


Figure 9: *Box-plot & Density plot for Amount of Payment Schedule column*

Additionally, we calculate the Pearson's correlation between these covariates and the response variable. The obtained correlation values show no extreme correlations that would pose any issues. The positive correlation between the Claimed Amount and the response variable is also expected, as it is evident that an increase in the claimed amount corresponds to an increase in the adjudicated amount.

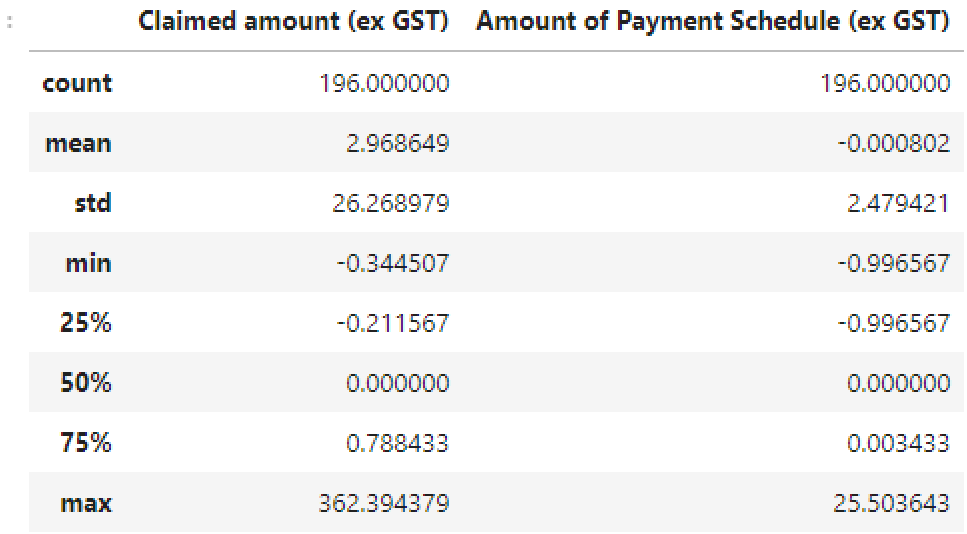


```
Correlation Matrix:
                  Claimed Amt  Amt of Schedule  Adjucated Amt
Claimed Amt          1.000000        -0.019505       0.945114
Amt of Schedule     -0.019505         1.000000      -0.004052
Adjucated Amt        0.945114        -0.004052       1.000000
```

Figure 10: *Correlation Matrix of Numerical Covariates w.r.t Response*

Preprocessing of Numerical Variables

For the processing of the numerical covariates we have used RobustScalar because it is least affected by outliers or extreme values. It uses the median and interquartile range (IQR) instead of the mean and standard deviation which does not influence the scaling process while handling for extreme values.



	Claimed amount (ex GST)	Amount of Payment Schedule (ex GST)
count	196.000000	196.000000
mean	2.968649	-0.000802
std	26.268979	2.479421
min	-0.344507	-0.996567
25%	-0.211567	-0.996567
50%	0.000000	0.000000
75%	0.788433	0.003433
max	362.394379	25.503643

Figure 11: *Rescaled Description of Numerical Covariates*

Machine Learning Methods

Supervised Learning Method(s)

Models to run:

- **Logistic regression:** Classification model to predict binary outcomes, it is fitted by applying a sigmoid function over a linear regression, which allows us to somehow interpret the results obtained by the model.
- **Random Forest** and **Xgboost**: Both are tree-like algorithms that work through

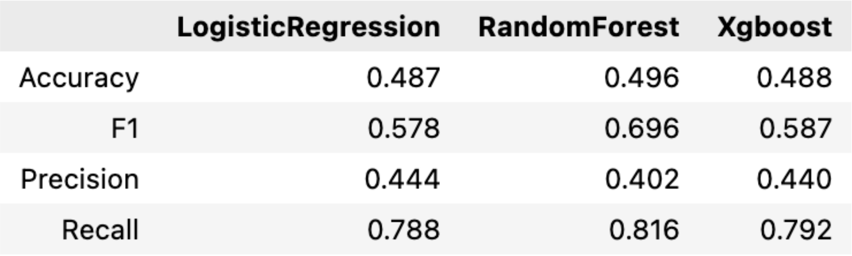
`decision Tree` ensembles with different methods to fit a regression variable. These algorithms by their decision tree origins allow us to know the feature importance of the independent variables when predicting the response.

Bootstrapping: It is a method to address the unbalanced dataset to avoid the negative effects of the unbalance. We did not use cross-validation or hold-out validation because the dataset is small. Using bootstrap as a replacement yielded better results.

Dataset Balancing: When we are facing unbalanced datasets, there are other methods to avoid the unwanted impact of them in the evaluation metrics. In this case we chose to use an **oversampling** method called **SMOTE**, this method works by creating synthetic copies of the minority class values in order to balance the dataset classes proportions.

Model Selection and Reason

Here, we tested the different proposed models with standard hyperparameters using **bootstrapping**. Then, we evaluated their predictive performance on the testing set using metrics such as 'Accuracy', 'F1', 'Precision', and 'Recall'. The results are as follows:



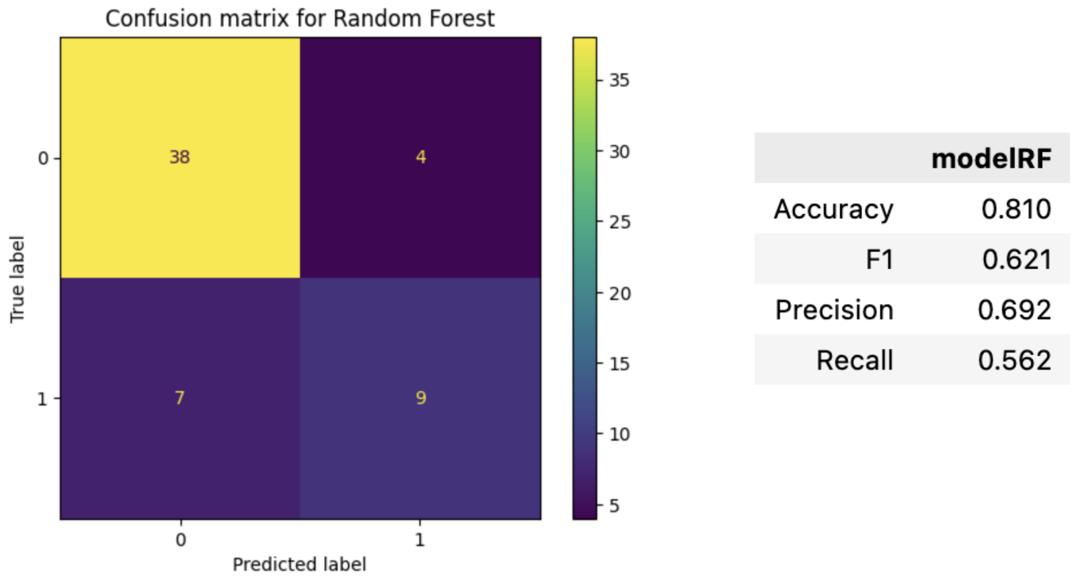
	LogisticRegression	RandomForest	Xgboost
Accuracy	0.487	0.496	0.488
F1	0.578	0.696	0.587
Precision	0.444	0.402	0.440
Recall	0.788	0.816	0.792

Figure 12. *Bootstrapping evaluation metrics*

The best average result in **bootstrapping** out of the three models is the `**RandomForest**` model. So we decided to use this model as our final model to predict our response. For this task we are going to train the same model using the dataset with the oversampling technique to get our final model.

Model Evaluation

We notice that the distribution of response variables is rather imbalanced (nearly 76% VS 24%), which may highly affect the performance of the model. Therefore we use **SMOTE** which is a kind of oversampling technique to balance response categories, followed by implementing the `**RandomForest**` model. The result is formed as confusion matrix as following:



	modelRF
Accuracy	0.810
F1	0.621
Precision	0.692
Recall	0.562

Figure 13. *Random Forest -Confusion matrix and evaluation metrics*

By comparing the confusion matrix with the previous results, it can be observed that the model's overall predictive ability has significantly improved. However, we can see that The models selected are underfitting the data. It is normal due to the size of the dataset and the naive approach for modeling.

Research Results

Based on the above results, we will answer the research question from two perspectives:

Firstly, the selected model could employ a feature sorting approach, where in the model identified 80 out of the 83 variables to be useful in predicting the final results. The graph presented below showcases the top 20 features ranked based on their impact on the model:

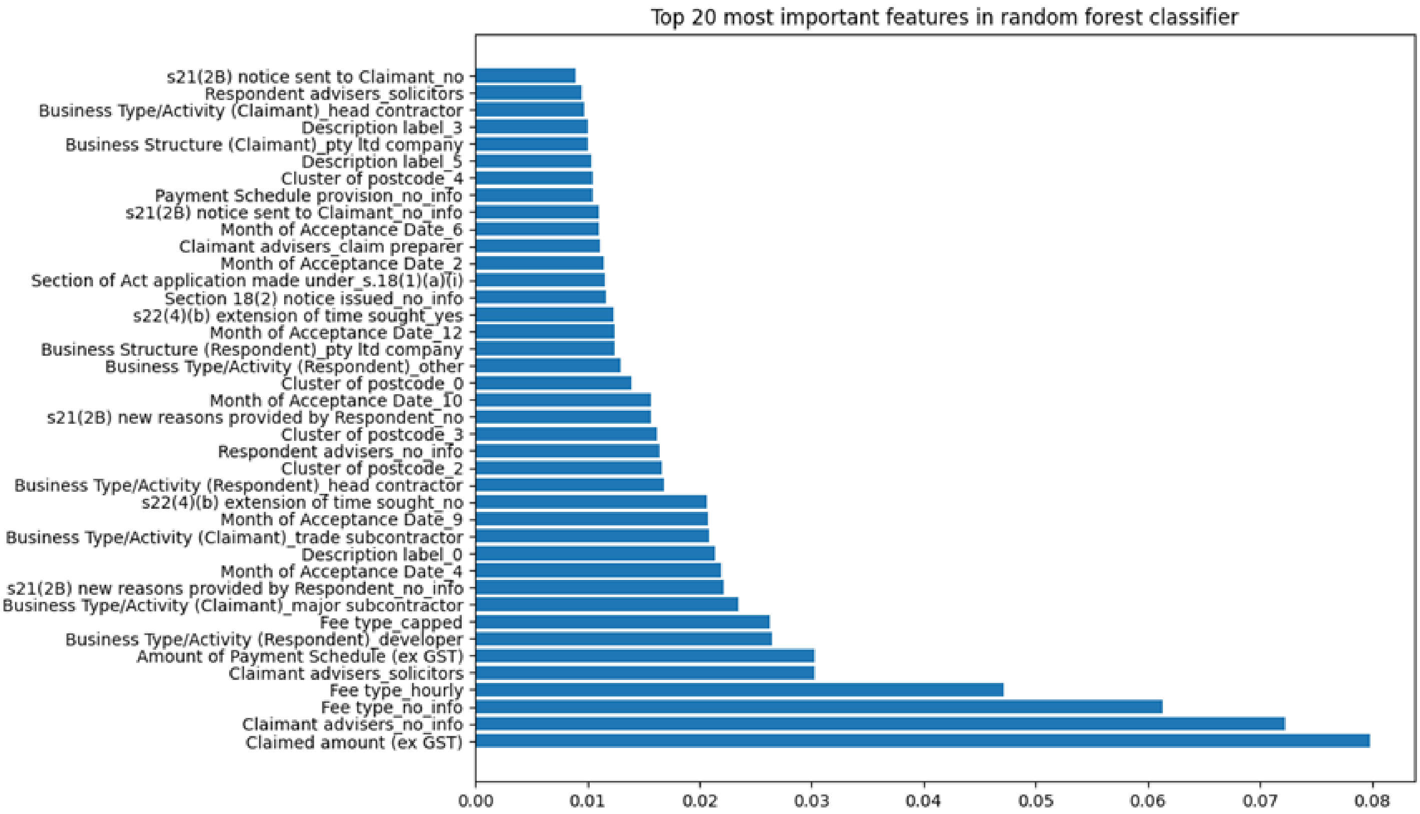


Figure 14. *Top 20 features selected with Random Forest*

Among the top features we can find some to be very relevant, including claimant advisers, claimed amount, postcode, and fee type. This information is crucial as it can indicate which features are most critical for predicting the target variable. Additionally, analyzing the predictive accuracy can provide insights into the data quality of these features within the dataset.

Secondly, referred to Figure 15, we can use metrics such as **Precision, F1, Recall, and Accuracy** to measure the prediction performance. For example, if we set a precision threshold of 0.65, we can consider the predictability of whether claimants will be compensated.

Limitations and Future Improvements

- The selected dataset has some major problems in terms of metadata, there are no data dictionaries nor field specifications, also the dataset present some inconsistencies regarding data schema, where fields that in theory are numerical or dates have string values, there is also a quality problem in relation to data completeness and data standardization, the features present high numbers of NaN values (in some cases over 50% of incompleteness) and there are no standards in the categorical values (Use of capitalized and lower characters indistinctly). Finally, one last problem found is that the data received does not have enough information to generalize the results obtained. Thus, these results can only be partially interpreted as a result for this exercise.

- Although the bootstrap approach is very useful to generate statistics over the evaluated metrics, the improvement of these models can go on two ways, the first one is the improving on the dataset, increasing the data set size as well as the data quality, the other one if centered in the modeling techniqˇues where it would be useful to apply hyperparameter tuning on the models in order to increase their predictive capacity.

Further information

For further information the full code and documents are available in the project repo which will become available after 19/05/2023 5:00pm:

https://github.com/Guticar94/COMP20008_Assignment_2