Security of Payment Project: using machine learning models to help adjudication applicants be better prepared

Shengyang Sun 1418346

July 24, 2023

Contents

1 Introduction

1.1 Background

SOP, or security of payment, is an Act to provide entitlements to progress payments for persons who carry out construction work or who supply related goods and services under construction contracts.

1.2 Dataset

The dataset contains the application information for period between May 2021 to June 2022. We categorized the attributes based on the chronological order, i.e. whether they happen before, at, or after the determination of adjudication. 'Before' attributes contains key date, description of work, procedural documents potentially required during the application, monetary amount, and company information pertaining to both the claimant and the respondent and are basically used as candidate predictors for our main response variable used in regression model: Adjudicate ratio, 'At' attributes (or we call 'FACT') are information about the actual adjudication determination and variables we try to predict. 'After' attributes are mainly the breakdown of adjudicator fees between claimant and respondent, ANA and adjudicator and will not be of much interest to us with regard to the scope of our analysis.

1.3 Research Question and Target Audience

Our target audience would be the potential users of the SOP service, we strive to answer the following question for them:

> *For a potential user of the SOP service, given its before application condition, should it use the SOP service?*

We'll address this question by considering 3 criteria:

1. Is potential claimant's work provided covered under SOP?
2. What is the potential claimant's standing among all the previous applicants?
3. Given the potential claimant's prior-to-application situation, what is his expected 'Adjudicate ratio' defined as:

$$Adjudicate\ ratio = \frac{Adjudicated amount}{Claimed amount}$$

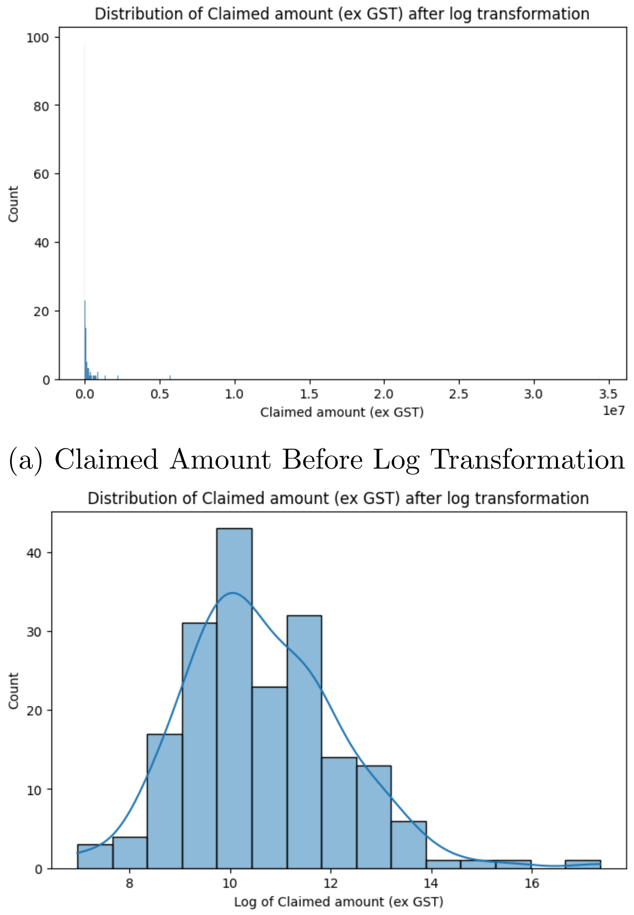
Each of the questions will be answered by applying knowledge of NLP, visualisation, and supervised machine learning model respectively.
Another potential audience of our research is Victorian government because we might expose some of SOP act's merit during our research which will encourage more people to use adjudication service instead of taking to the legal path, thus saving public resource.

2 Data Wrangling

2.1 Data Cleansing

Like every raw, unprocessed dataset, our adjudication dataset comes with numerous issues for us to address. Different categories of problematic data are briefly discussed below, and some representative cases are listed. For a full list of the preprocessing that we have done to the dataset to prepare it for a smooth data analysis, please refer to our source code.

1. Missing data: these are the data that are simply missing.
   - Cases that are 'not **yet** determined' contains no information about adjudication determination and are thus discarded.
   - 'Respondent advisers' has 42 cases of 'Not Known' which essentially are missing values. We imputed randomly chosen values based on the distribution of all the non-null values.
2. Logically or legally contradictory data:
   - Cases that are determined but has value in the column 'Reason not determined' are logically contradictory. So we set the reasons to null.
   - Cases that are applied under section 18(1)(b) (which is the case that the respondent does not provide payment schedule) but the 'Payment schedule provided' column has value 'Yes'. We set them to 'No'.



(a) Claimed Amount Before Log Transformation

(b) Claimed Amount After Log Transformation

Figure 1: Comparison of distribution before and after the log transformation

3. Highly skewed variable and outliers.
    - Outliers can have a huge leverage on the outcome of a regression model thus must be dealt with. In our dataset, 'Claimed amount' is highly skewed with an outlier of magnitude 34 million while majority of the data has value only under one million. This situation is ameliorated by log transformation as evidenced by the almost normal distribution after the transformation. (Figure 1)
4. Wrong data
    - 'Determination release date' has some rows with value 00-Jan-1900 which is the starting value of Excel's time which basically means null value.
5. Perfectly correlated or not informative attributes:
    - The information of whether the respondent has provided a payment schedule is implicitly embedded inside 'Section of Act application made under'. For example, if an application is made under s.18(1)(b), that means the respondent has not provided any payment schedule which means we could safely remove the column 'Payment schedule provided' without losing any information.

2.2 Extract meaning from text

After we finished all the cleansing, we assign each column to an appropriate data type before the last step of our data preprocessing: extracting meaning from the two string type variables:

- 'Description of project and contract works':
  This attribute is a small string of texts that is humanly readable but is very hard to incorporate into the data analysis, especially fitting into a regression model. What we do is applying some basic NLP techniques to transform them into a word count vector and performed a classification using K-mean with k=7. The reason we chose 7 is that there are 7 categories of construction that is covered under the SOP act. K-mean is a preferred classification method here because is an unsupervised ML technique and thus does not require a previously labeled data.
- 'Project postcode':
  Postcode has more than 130 categories while our data only contains more than 300 observations. We reduced the granularity of postcode by assigning postcode to LGA, which reduces the number of categories to 5.

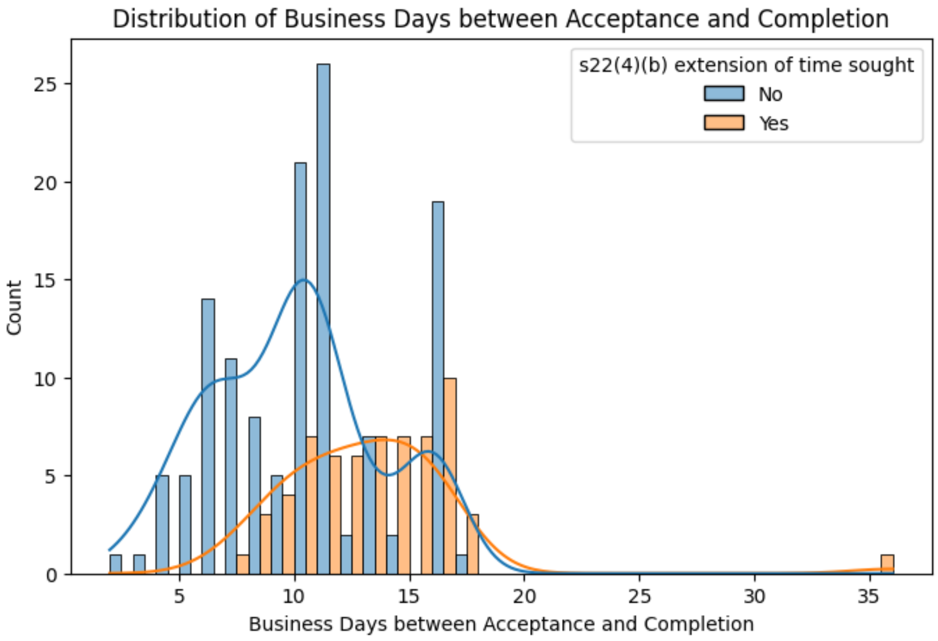


Figure 2: Distribution of Business Days between Acceptance and Completion

2.3 Exploratory Data analysis

In this section we explored the dataset with some EDA techniques to give the potential applicant a high-level picture of the SOP act.

2.3.1 Descriptive statistics

Some high-level overview of the SOP act.

- Out of all the applications during the dataset period, 61.89% of them are determined. Reasons for cases not determined are withdrawal, settlement between parties and time-out.
- Although the SOP act requires the adjudicator to determine an application within 10 business days (s.22(4)(a) in SOP act) or 15 extra days if the respondent seeks to extend it under s.22(4)(b), you should not be too surprised if that takes longer than 10 days because that happens more than half of the time. If the respondent has applied for extension, you should expect the determination to arrive around 15 days with rare extreme case where it took more than 30 days to arrive at the determination. (Figure 2)

2.3.2 Pairwise Relationship

We plotted a heatmap for all the continuous variables in the dataset (Figure 3). Here are just a few of the interesting findings:

- Adjudicator fee is proportional to claimed amount but not in a strictly linear fashion that means it has some sort of soft cap which is evidenced by a higher correlation of 0.73 between 'adjudicator fee' and 'log claimed amount' than that of 0.4 between 'adjudicator fee' and 'claimed amount'.
- The more justified your application is (as shown by a high adjudicate ratio), the less you need to pay for the adjudicator fee. This is evidenced by a -0.76 correlation between 'Adjudicate ratio' and 'Adjudicator's fee payable by claimant'.
- The longer it takes to determine the adjudication, the higher adjudicator fee one need to pay. This is evidenced by a 0.39 correlation between 'Business day between application and acceptance' and 'Adjudicator fee'.

Another thing a potential claimant might worry about is that is he going to be discriminated because of the magnitude of his claimed amount. The answer is a clear no as evidenced by Figure 4 where we can see the there is no distinguishable pattern with the change of claimed amount.

2.3.3 Using OpenAI API to inform claimant whether his construction work is covered under SOP act

The legal document dictates seven categories of construction works that are covered under SOP act and three that are not. We prompted the user to input the type of construction work that he has provided and feed this information along with the covered/not covered categories into a OpenAI language model to ask it to classify whether the work

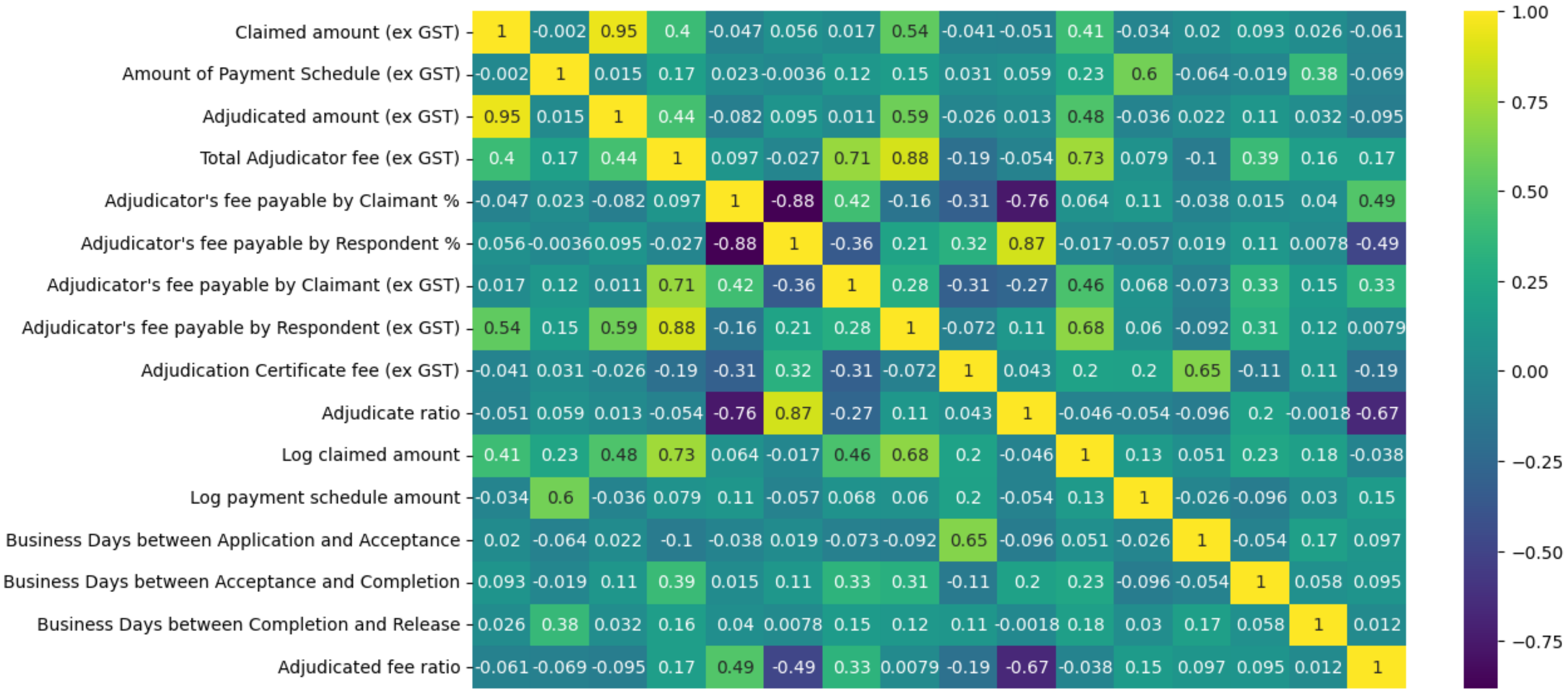


Figure 3: Pairwise Relationship within continuous variables

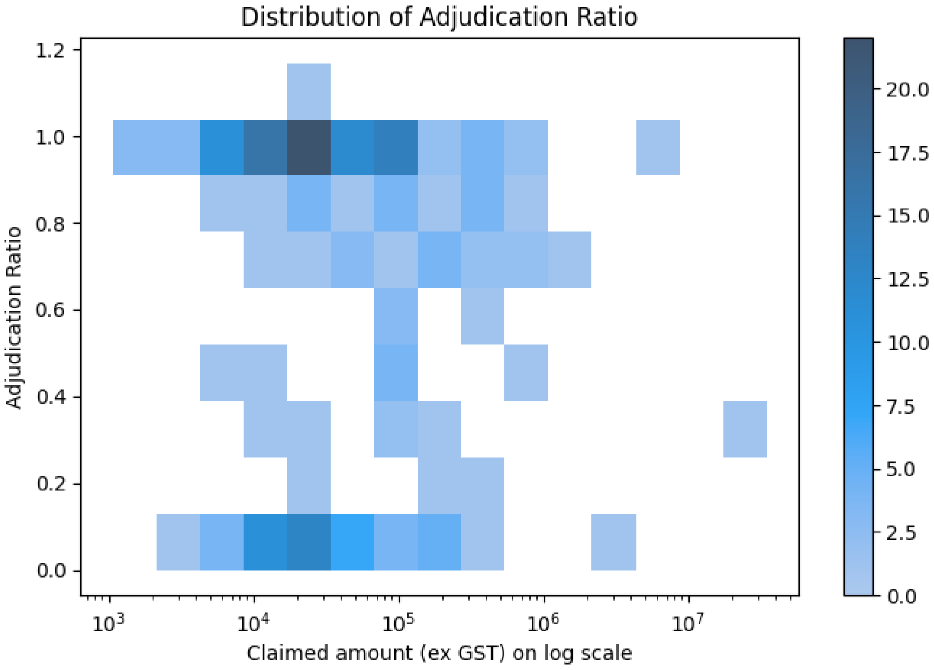
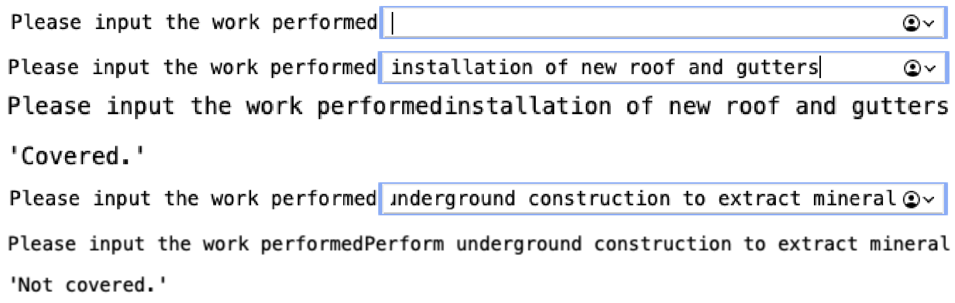
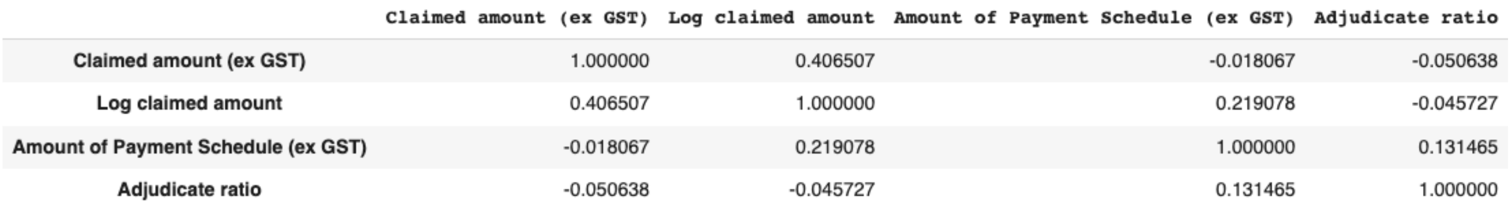


Figure 4: Histogram of Adjudicate Ratio and Log Claimed amount



```
Please input the work performed |
Please input the work performed installation of new roof and gutters
Please input the work performedinstallation of new roof and gutters
'Covered.'
Please input the work performed underground construction to extract mineral
Please input the work performedPerform underground construction to extract mineral
'Not covered.'
```

Figure 5: Working example of work type filtering



	Claimed amount (ex GST)	Log claimed amount	Amount of Payment Schedule (ex GST)	Adjudicate ratio
Claimed amount (ex GST)	1.000000	0.406507	-0.018067	-0.050638
Log claimed amount	0.406507	1.000000	0.219078	-0.045727
Amount of Payment Schedule (ex GST)	-0.018067	0.219078	1.000000	0.131465
Adjudicate ratio	-0.050638	-0.045727	0.131465	1.000000

Figure 6: Correlation matrix between response and continuous variables

performed is covered or not. You can find a working example of it in Figure 5. Although for SOP, the definition of covered work is quite simple and people can just refer to it without any legal counsel. This technique can definitely be generalized to more complex legal definitions.

3 Feature Selection

For feature selection, we performed Pearson's correlation between response and continuous variables and Mutual Information categorical ones. This is cross validated with the visualisation analysis. Because we don't have domain knowledge in either legal or construction area, the computed values are compared relative to each other instead of having an absolute threshold.

3.1 Pearson's Correlation

Claimed amount suffers from some serious outlier issues and that's why we log transformed it to bring it back to a close to normal distribution. The effect can be clearly seen in the correlation matrix. Without this adjustment, one would wrongly believe that there's almost no relationship between claimed amount and response when in fact they are highly correlated. (Figure 6)

3.2 Mutual Information

Mutual information is calculated between response variable and 'BEFORE' categorical variables to discern those that possibly has a predicting power over response.

```
NMI between categorical variables and the response
variable                                            NMI
Section of Act application made under               0.09
Claimant advisers                                   0.08
s22(4)(b) extension of time sought                  0.04
Business Structure (Respondent)                     0.05
Business Type/Activity (Respondent)                 0.07
Business Type/Activity (Claimant)                   0.08
Respondent advisers                                 0.05
Business Structure (Claimant)                       0.08
Region name                                         0.06
s21(2B) new reasons provided by Respondent          0.03
```

We can see that relatively, 'Section under Act application made under', 'Claimant adviser', 'Business Type/Activity (Respondent)' have higher MI values. But in absolute term, their values are pretty low ranging from 0.07 to 0.09, and thus we'll not rule out any candidate variables for now and do some significance testing during model fitting stage and hope that the results will corroborate.

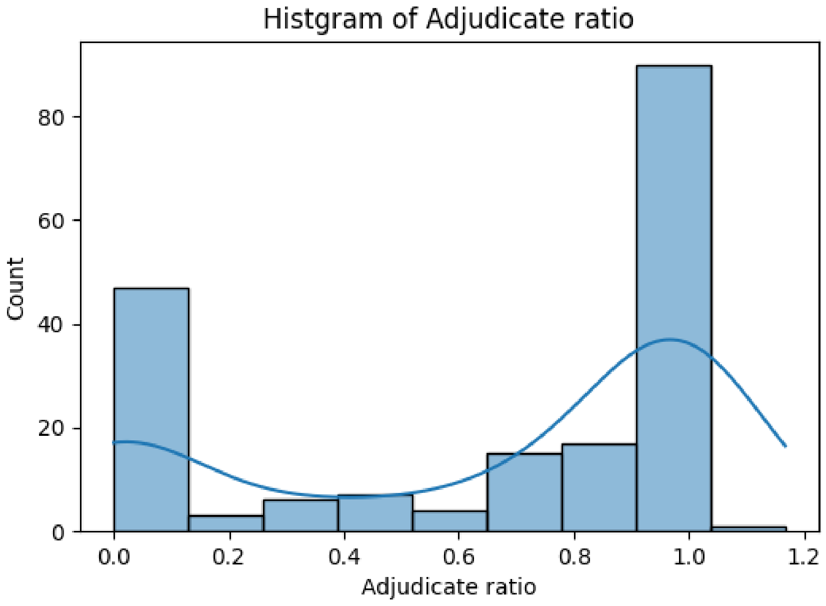


Figure 7: Histogram of Adjudicate Ratio

4 Model Fitting

4.1 The purpose of the model is two-fold

- To give potential claimant a general idea of how much he is expected to receive (as a percentage of his claimed amount) through the SOP service.
- To inform potential claimant which factors could have a significant influence on the outcome of the adjudication result so that claimant can be more prepared.

4.2 Justification of Model Choice

We choose to use Beta regression to fit our data because our response variable: the 'Adjudicate ratio', is a percentage value and is bounded by (0, 1). Also percentage value tends to have heteroskedasticity and skewness issues, which means their variance is higher in the middle and lower at the two extremes, and the distribution is not normally distributed. Both of these characteristics is clearly visible in the histogram of the response variable (Figure 7). Beta distribution is naturally heteroskedastic and can easily accommodate asymmetries.

4.3 The procedure of the model fitting

- randomly split dataset into train and test, the ratio is set at 90:10 because we want to have as much data in the train set as possible to avoid structural zero issue.
- Include every candidate predictors into the model.
- Filter out any insignificant predictors based on Wald test using 95% as critical value.
- Try to further remove variables and test whether they have as good fitness as the larger model using Likelihood Ratio Test to try to achieve an even more parsimonious model.
- Two link function, i.e. logit link and loglog link, are compared. Theoretically, loglog link should perform better with a dataset skewed toward the 1 side (large extreme value), but the two models have almost the same AIC. So we just arbitragely chose a more parsimonious model which uses loglog link.
- The model is used to predict the outcome of the test set.

4.4 Final model

Our final model is as follows (notice some categories are discarded due to structural zero issue) (Listing 1):

Listing 1: R output

```
Call:
betareg(formula = Adjudicate.ratio.trans ~ factor(Section.of.Act.application.made.under)
+ factor(Claimant.advisers), data = train,
    link = "loglog")

Deviance residuals:
    Min      1Q  Median      3Q     Max
-2.4495 -0.5442  1.0254  2.4974  2.6033

Coefficients (mean model with loglog link):
                                                       Estimate Std. Error z value Pr(>|z|)
(Intercept)                                              0.8012     0.2704   2.963 0.003047 **
factor(Section.of.Act.application.made.under)s.18(1)(b)  0.5318     0.1611   3.300 0.000966 ***
factor(Claimant.advisers)None                           -0.6286     0.2876  -2.185 0.028871 *
factor(Claimant.advisers)Solicitors                     -0.1613     0.2982  -0.541 0.588414

Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)  0.63058    0.05517   11.43   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Type of estimator: ML (maximum likelihood)
Log-likelihood: 229.2 on 5 Df
Pseudo R-squared: 0.1865
Number of iterations: 21 (BFGS) + 2 (Fisher scoring)
```

4.5 Interpretation of the Model

The final model turns out to be parsimonious, intuitive, informative all at the same time, which is super great!

- The coefficient of 0.5318 means comparing to applications applied under s.18(1)(a)(i) (where the respondent provided payment schedule), those applied under s.18(1)(b) (where the respondent failed to provide payment schedule) has $e^{0.5318} - 1 = 70.2\%$ higher log(Adjudicate ratio).

- The coefficient of -0.6286 means comparing to applicants who have a preparer to help with the application, those do not have any adviser has a $1 - e^{-0.6286} = 46.67\%$ lower log(Adjudicate ratio).

- The coefficient of -0.1613 and its insignificance means the effect of having a solicitor helping you with application is as good as having a preparer.

- In summary, a claimant is in a more advantageous position if the respondent does not provide a payment schedule, and he has some professional to help preparing with the application.

4.6 Performance

Overall, we don't expect our model to behave greatly in terms of predicting power as suggested by the low correlation and mutual information calculated in the feature selection section. This is evidenced by a low 'Pseudo R-squared' value of 0.1865. Besides, the final model only have 2 x 3 = 6 levels in total which means there can only be 6 possible predictions. Since most of the adjudicate ratio is 100%, we are bound to have larger over-estimation problems for those data points with a low adjudicate ratio and milder under-estimation problems for those data points with a adjudicate ratio close to 1. This is evidenced by the raw residual plot (Figure 8).

5 Limitation and Future Improvement

5.1 Preprocessing

The logical relationship between varialbes is very intricate and intertwined which makes the ordering of our preprocessing very vulnerable. A single switch of ordering between steps could render following steps unfunctional. This signals that our preprocessing is not very robust and is too specific to our dataset to generalize to other datasets. On the other hand, the nature of our preprocessing is highly accurate which satisfied both legal requirement and logical constraints instead of just dealing with missing values and mis-inputs. So it's really hard to strike a balance between robustness and accuracy

5.2 Feature Selection

Some domain knowledge in legal and construction will help us to better filter the features.

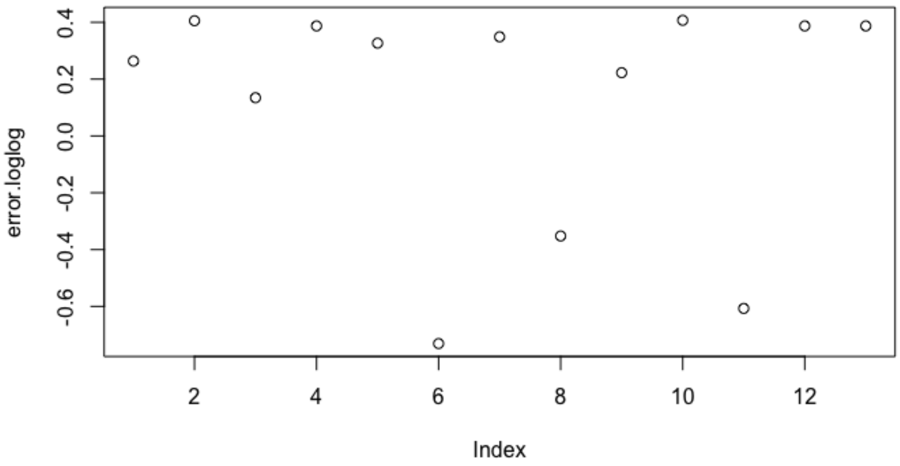


Figure 8: Raw residual of prediction on the test set

5.3 Model Fitting

If we have the time, we could incorporate adjudicator fee into our model to give the claimant a more accurate estimate of the amount he is going to receive. We could also reverse engineering to try to deduce how the adjudicator fee and certificate fee are calculated. Limited by the small data size, there would be a lot of structural zeroes if we try to include interaction terms in the model. Some of the variable pairs definitely have the potential of being predictive such as claimant's business activity:respondent's business activity, and region:type of construction work. With a larger dataset, model with higher order terms could be achievable.