Fraud Detection In Home Insurance Claims Using Classification Techniques

Yihezhang Z5201685

Weijia Liu Z5163468

*Abstract* -- **In recent decades, the rapid development of social economics, the insurance industry has achieved huge progress especially home insurance. People regard it as a target of frauds since the large amount of money has been involved. The objective of this study is to identify suspicious fraud cases by different modelling techniques. The predictive models are capable to label high abnormal activities to minimise loss for insurance company.**

*Keywords* -- **Insurance claims, fraud detection, machine learning, data cleaning.**

Section 1. introduction and motivation (mention some methods name)

Fraud has been plaguing the insurance industry since the beginning of its existence. False claims incur substantial loss for the insurance providers, and they drive up the costs of insurance products for everyone. On the one hand, fraud detection itself incurs costs. If the company develops an accurate fraud detection system but the implementation is time-consuming, expansive or involving the breach of confidentiality of data, it is not of practical use. On the other hand, failure to recognize any frauds is clearly unadvisable. Therefore, investigating an effective model and fraud detection techniques is required to improve the quality of the service and minimize the unnecessary costs. This problem falls into the category of classification in the realm of machine learning.

This project helps investigators detect fraud cases from large databases by sophisticated model and techniques which are developed by following steps. Firstly, clean the data use shell command and python notebook. Secondly, by labelling the useful features we build the models based on training dataset. Methods of building models include decision tree, random forest and linear regression model. Then, we compare the performances of different approaches. The final predictive model provides the acceptable results and benefits others in the future.

This paper is constructed as follows: The next section depicts the overview of researches in related area, presents the problem and techniques that have been encountered and conducted. The third section describes the origin and structure of the dataset, including interpretation of the context that situated behind the data. The fourth section gives the process of data cleaning and selection of features to build different models. The fifth section displays the classification results and compare performances of each model using different metrices. The last section draws some conclusions and presents directions for future work. Additionally, references and appendices are also attached at the end of this paper.

Section 2. Literature review

Due to lack of experience and domain knowledge, we realize a comprehensive literature review on fraud detection projects is necessary. In order to get inspired on some similar ideas we read though a dozen of academic papers and we discover the followings are most related. Yi Peng etc. [1] introduced three predictive models: Naïve Bayes, decision tree and Multiple Criteria Linear Programming to be trained, they gave out the test results to compare the accuracy and also proposed some suggestions for future projects on frauds. Capelleveen etc [2] provided the outlier method of data mining technology for the health insurance fraud detection. This is also used for detecting the suspicious behavior of medical service providers. Zhenxing Hou [3] proposed a fraud risk analysis according to cluster analysis for isolation by distance clustering method. Clifton Phua etc [4] conducted a research survey which explored almost all published fraud detection studies and gave a comprehensive overview of different types of fraud, the methods and techniques people used and their limitations. They indicated unsupervised, semi-supervised and text mining from law enforcement approaches for different types of data.

Jing Li [7] conducted a detailed survey on statistical methods for fraud detection in health care data area. This survey classified the behaviors of fraud cases, identified the sources on which fraud detection has been conducted, provided crucial steps in data preprocessing, compared statistical methods that currently in use, and provided the advice on future directions. Thorton et [8] indicated a multidimensional data model and analyzed important approaches to help predict fraudulent cases. In addition, Weizong Zhang [9] has conducted the single-factor and two-factor analysis by SAS software according to the application of Logist regression model to classify the considerable factors in fraudulent claims. These techniques that have been mentioned above are quite effective and are good references for our own project.

Additionally, Rafael et [10] evaluated the behavior and influence of feature selection methods, performed undersampling strategy to improve the performance and used real data to check the results. The model achieved high efficiency by reducing the number of features. Qi Liu [5] introduced fraudulent behaviors in health care system and analyzed characteristics of dataset, compared and reviewed some existing fraud detection approaches. He also proposed a clustering model that contains information of geo-location to identify dubious claims. Furthermore, Ayhan Demiriz [6] evaluated the value of geographical information for deriving business rules to detect and prevent financial frauds and scams in his paper. These papers inspire us that location information could become one of the significant features.

In general, the papers suggest and prove that the machine learning techniques are effective and beneficial in detecting fraud activities. They are informative guide and inspiration for our following research.

Section 3. Preparation and interpretation of the dataset

Yuumi Insurance provided the log data that records the interaction between the customers and the company. The dataset contains more than 3 million records which enhances the level of difficulty. Each piece of information is structured, timestamped and describes a specific activity of a customer, including quotes, claims and payment status. When a customer enters one of the platforms (mobile app, website or phone calls), a quote will be generated based on the client's information.

If the client chooses to accept the insurance policy and make a payment, this information is recorded as well, without the specific amount of money the client paid. When a claim is made, the company will look through the client's claim history and examine all the information related to the client and decided if they will accept or deny the claim.

Section 4 Material and methods

The log data we are offered have two main characteristics. Firstly, it contains a large amount attributes to explain the personal circumstances of each client such as name, gender…. Secondly, each client is randomly assigned a customer id. By using this information to link different cases, we are capable to gain a global view of a series activities have been conducted by each client over time. The history of a series of actions is playing an important role in detecting fraudulent behaviors.

4.1 Goal setting

The goal of this project is to investigate an effective model and fraud detection techniques to improve the quality of the insurance service and minimize the unnecessary costs. This problem falls into the category of classification in the realm of machine learning.

4.2 Data cleaning

The log data should take an acceptance form that agrees with the constrains of the statistical model. It is a huge and challenging task that approximately takes us 70% time to complete. We try to find an appropriate way to address this complicated procedure. The relevant diagram and coding of the whole data cleaning process has been appended at the end of the paper. Details for each step are presented as follows.

First of all, we split the huge raw dataset into different files according to the status (highlighted in yellow in picture 1) of each case by shell command. Then, since the variables such as customer id, case id, platform… are all concatenated inside “Message” column, we need to separate them to become single predictors so that it is easier for us to build the model later.

|  |  |
| --- | --- |
| Message | Timestamp |
| 8f70c7577be8483 - mobile\_browser - Quote Started for customer: 99ccf1,1483192800.0 | 1483192800.0 |
| "1368d40a4f6e455 - mobile\_browser - Quote Completed for customer: 99ccf1 with json payload {'name': 'Nicole Berry', 'email': 'Nicole Berry@hotmail.com', 'gender': 'male', 'age': 29, 'home': {'type': 1, 'square\_footage': 311.80361967382737, 'number\_of\_bedrooms': 2, 'number\_of\_floors': 1}, 'household': [{'name': 'Oscar Berry', 'age': 25, 'gender': 'female'}, {'name': 'Mark Berry', 'age': 10, 'gender': 'female'}, {'name': 'Jacqueline Berry', 'age': 14, 'gender': 'male'}], 'address': '66 Lake Jamieview,PSC '}" | 1483193676.507667 |
| 90527688b31d445 - mobile\_browser - Claim Started for customer: 99ccf1, | 1483193794.689323 |
| c4013f44ea6d40c - mobile\_browser - Payment Completed for customer: 99ccf1, | 1483193794.689323 |
| 4c9ab2942b484f2 - pc\_browser - Claim Started for customer: 9bae09, | 1483197184.513103 |
| e67b69c9b4554c0 - pc\_browser - Claim Denied for customer: b7aab4 - reason : fraud, | 1483204320.4773 |
| … | … |

picture 1. Example of raw data

picture 2. Split raw data into different files

|  |  |
| --- | --- |
| Table Name | Attributes |
| payment\_completed | Case, name, email, gender, age, home\_type, square\_footage, number\_of\_bedrooms, number\_of\_floors, household, platform, customer\_id, timestamp, tag |
| Claim\_started | case, platform, customer\_id, timestamp, tag |
| Claim\_denied | case, platform, tag, timestamp, customer\_id |
| Claim\_Accepted | case, paid\_amount, platform, customer\_id, timestamp, tag |

no missing value need to be handled…

4.3 feature selection

The purpose of feature selection is to choose the more powerful predictors and create new variables which maximize the discrimination power in our model. In this case, we consider the following influential aspects.

--frangment 0 (matched\_timestamp)

In our dataset, one of the most obvious patterns for fraud case is that the timestamp of completing a payment coincides with the timestamp of starting a claim with high precision. All of the fraudulent activities match this feature but not the other way around. We need to identify other predictors that eliminate normal people from the pattern of matched timestamp.

--fragment 1 (name and gender inconsistent)

Besides the most significant pattern, we catch that matched\_timestamp only provides us with 84% accuracy which implies that there is a small portion of people who are not frauds even though they conducted payment and claim at the same time. In order to dig out the hidden clue, we firstly go through the dataset manually to discover any useful information. To our surprise, one abnormal pattern appears. Among the fraud cases, the name and gender of the customer seems inconsistent. For investigating purpose, we extract the predictor of name and gender also the email as single variables in our modelling process, however, the results indicate that these added predictors enhance insignificantly in terms of performance.

evidence

--frangment2 (gender)

In the direction of revealing if gender has an influence on model prediction, we group the fraud cases according to customer’s gender and outcomes shown as following. There is approximately no difference between male and female in terms of frauds.

evidence

--fragment3 (suburb)

Several papers have mentioned that geographical information is helpful with fraud detection to some extent which also inspire us on our project since we are offered address for each customer. We calculate the fraudulent rate which is the ratio of number of fraud cases in total claims for each suburb and we obtain the following results.

evidence

Challenges

As our project progresses, we record anything significant in weekly checklist such as the difficulties we have encountered or the influential patterns we have noticed. This is a good approach to remind us not to make the same mistakes and to avoid doing the repeat work. Firstly, in this insurance dataset, there is no link between a customer id and case number. For example, as stated in our assumption above, if one customer id corresponds to more than one json payloads and he claimed several times, we couldn’t identify which case matches which payload. The inconsistences bring a lot of difficulties and confusion. The steps we take to overcome this obstacle are group the series of activities by platform, or if they are not using the same platform, then for that particular claim we map it with the nearest claim started activity.

Section 6. Conclusion and future directions

In recent decades, due to the rapid development of social economics, the insurance industry has achieved huge progress especially health care insurance. People regard it as a target of frauds since the large amount of money has been involved. Therefore, investigating an effective model and fraud detection techniques are required to improve the quality of the service and minimize the unnecessary costs. Along with this objective, efforts were made to clean the log data into tabular form, select key features in modelling to classify the abnormal behaviors, summarize and compare the performance of each model.

We were provided efficient techniques and methods by reading through the current documents and researches and based on the results of our project, there are some future directions we suggest. Firstly, since the pattern of fraud changes over time, it is necessary to periodically update the labeled data which is the key to determine the classification results. Secondly, all of the existing papers have provided techniques and approaches to identify fraudulent behaviors but none of these has reflected on the causes of fraud, however, to prevent fraud cases is our ultimate goal. Lastly, currently we are using artificially generated data to detect the insurance fraud, the future work will involve real data that consists more realistic situations.

Acknowledgement

This project was supported by …

* Analysis
* Goal of our statistical analysis: depending on the problem we decide to work on

For this statistical analysis, our goal for building this model is trying to detect the fraud cases efficiently and accurately.

* Data collection and exploratory analysis: What exploratory analyses did you do, graphical or otherwise. Which variables do we choose to use in your analysis and why?

To take a closer look at the data took help of “ .head()”function of pandas library which returns first five observations of the data set.

I found out the total number of rows and columns in the data set using “.shape”. Dataset comprises of \_\_\_ observations and \_\_ characteristics.

It is also a good practice to know the columns and their corresponding data types, along with finding whether they contain null values or not.

The describe() function in pandas is very handy in getting various summary statistics. This function returns the count, mean, standard deviation, minimum and maximum values and the quantiles of the data.

To use linear regression for modelling, its necessary to remove correlated variables to improve your model. One can find correlations using pandas “.corr()” function and can visualize the correlation matrix.

* Model choice: what models did you fit?

GLM, decision tree??

* what assumptions did you make?

Each customer id corresponds to one customer.

Each customer may have different json payloads.

Each claim corresponds to the most recent json payload.

* Model fitting: what software did you use?

shell command to split the files into different sections

python to clean the data and plot the diagram

how is variable/model selection carried out. any relevant computing code to support your results.

we extract the variables from the dataset and based on our assumption we fit different models and by comparing the performance metrics such as roc curve to choose the best one.

* Diagnostics: What model checking diagnostics did you carry out if any? What are the limitations if any.
* Model assessment: How would you assess how useful/good your model is at prediction (if any)? what are the limitations of your model?

boxplot

kernel density (density plot)

variable matrix

pairwise plot??

how to check outlier (it that important for our case??)

definition of confusion matrix, type 1 and type 2 errors and introduce to roc and the reasons it suits this problem and how to measure the performance, calculate the area for model selection part

logistic regression

references :

1. Liu Qi, Miklos Vasarhelyi “Healthcare fraud detection: A survey and a clustering model incorporating Gro-location information.” 29th World Continuous Auditing and Reporting Symposium(29WCARS), Brisbane, Australia.2013
2. Ayhan Demiriz, Betül Ekizo÷lu. “Using Location Aware Business Rules for Preventing Retail Banking Frauds”
3. Li, Jing, Kuei-Ying Huang, Jionghua Jin, Jianjun Shi. “A survey on statistical methods for health care fraud detection”, Health Care Manage Sci DOI 10.1007/s10729-007-9045-4
4. Dallas Thornton, Roland M.Mueller, Paulus Schoutsen, Josvan Hillegersberg. “Predicting Healthcare Fraud in Medicaid: A Multidimensional Data Model and Analysis Techniques for Fraud Detection.”

Main Questions:

1. how to detect fraudulent cases accurately
   1. which predictors play important roles in our model
   2. have history
   3. new customer
2. Based on our prediction, how much money we can minimize for the company
   1. since we don’ have that information, can we predict how much money will be paid if a claim is accepted
   2. if we can’t predict based on given data can we use mean?

* case 与case之间也没有link, 不知道claim denied 对应的哪一个claim\_started.
* customer id and case number doesn’t have a link, so we can’t decide which case corresponds to which payload, then we can’t match the money to that case.

we can’t predict the money based on all this information

decide go with mean(paid\_amount)

* assumption :

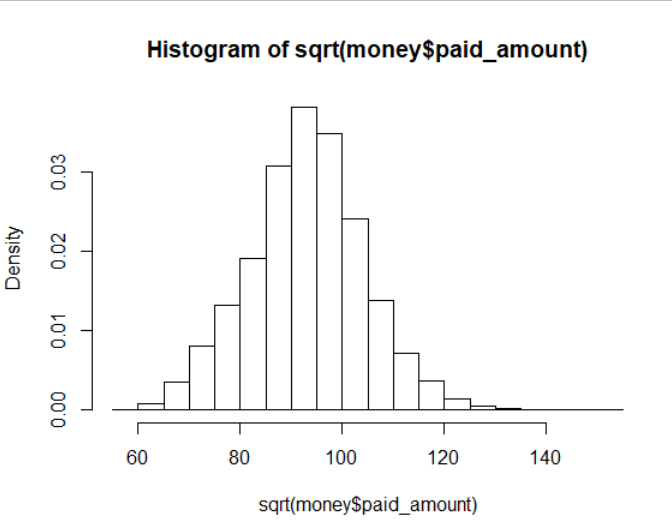
based on our model, if fraud, then investigate, otherwise pay.

so improve true positive is fine

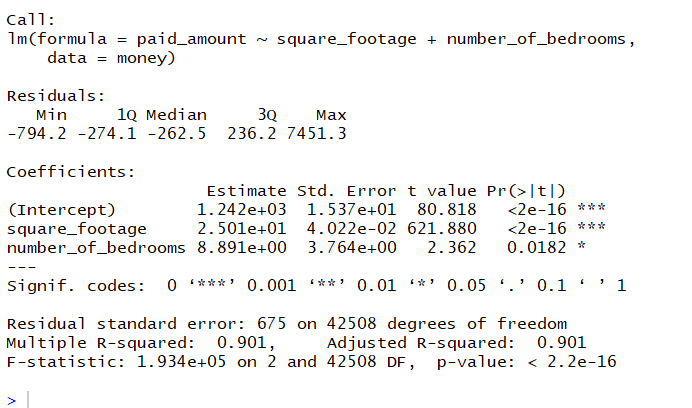
no need to care about false positive.

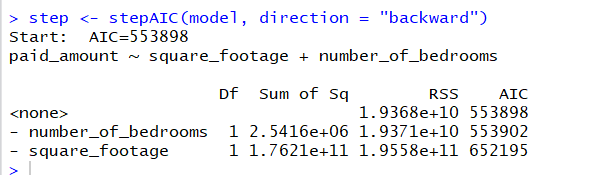
We attempt to estimate the potential loss generated by the false negatives. For each claim case denied, the coverage that could have been paid to the customer should the case wasn't denied is not available in this dataset. Fortunately, for each claim accepted, the coverage paid to the customer is accessible.

Using histogram, we were able to identify the distribution of coverage as normally distributed density and thus we choose a linear regression model as appropriate for our model to estimate the amount of money that could have been paid to a fraudulent customer, if he or she is misclassified as a normal customer.

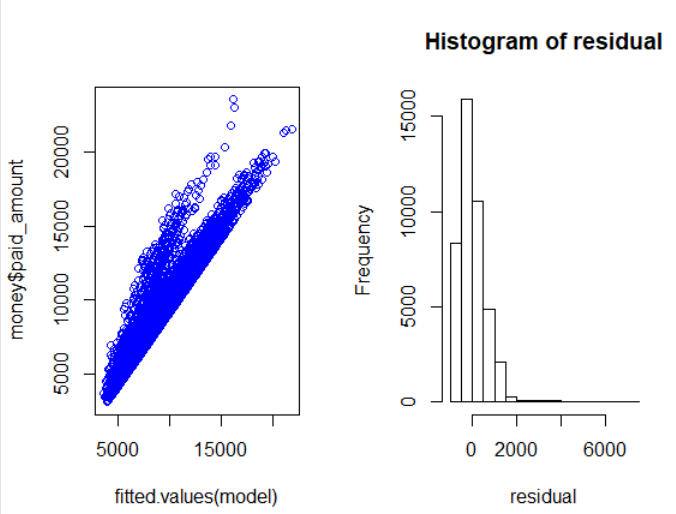


We obtained our best fitted variables by comparing summary tables of coverage against different variables and explorative analysis suggests that square footage of a home and numbers of bedrooms are the only significant predictors of coverage as per t-test statistics with p-values less than 0.05 also with the minimum value of AIC. Keeping other variables constant, we noticed that one unit increase in square footage of the house will result in an increase in coverage by 25.01.





There is a strong correlation (R square of 0.901) between fitted value (Yhat) and actual response (Y) as illustrated in the graph. Thus we would apply this model to predict coverage of a fraud case. The residuals are normally distributed despite slight right skewness.



The following boxplot indicates that the median of coverage for different numbers of bedrooms are pretty much the same, in particular there are lots of outliers above the fourth quantile which may represents that the coverage may depend on other factors that not given in the dataset, such as the reason why customer claims and so on…

