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. Then, 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.

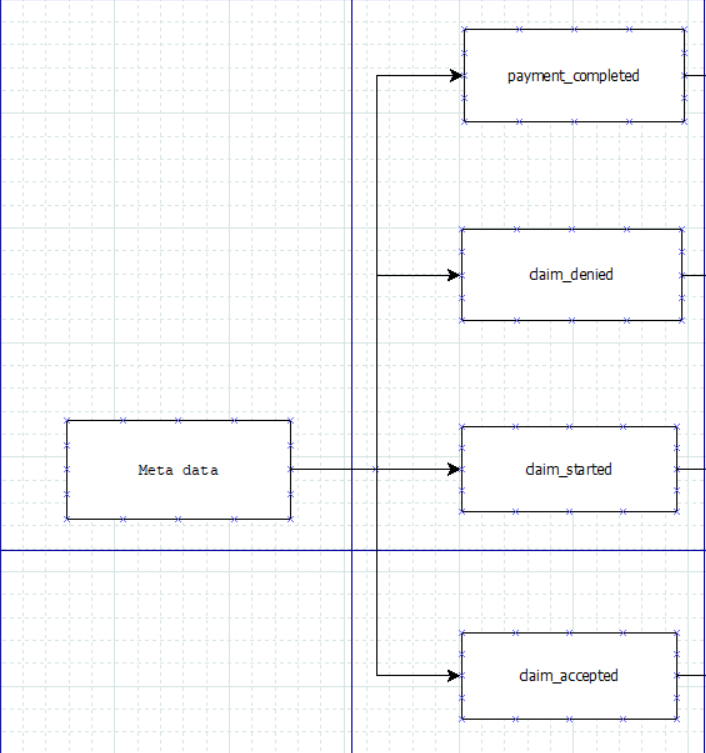
4.2 Data cleaning

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…

4.3 feature selection

data cleaning

1. split file based on case status
2. separate the variables out from one column for each file, such as name, gender, address
3. explain the structure for each file, create a table
4. for each file in 3, group by customer id, left join we then have a big table show

customer id no.of claims no. of denied no. of accepted

1. challenge: found more than one payload for each customer.

based on our assumption, one customer id corresponds to only one customer.

then how to explain the different personal information?

challenge: how to identity which case corresponds to which json payload?

by sorting the timestamp, it is indicated that a complete process is made up with a series of transactions.

For each claim we just group it with the most recent json payload.

1. start labelling the dataset

for each incoming case we label the important features such as if timestamps of claim started and payment completed are equal… these features are the keys to determine whether it is a fraud case or not

1. it is clearly shown that address becomes the second most important feature, and we just ignored that. So we started from extracting that piece of information from the dataset

在金融风控领域，基于位置信息的特点，通过位置信息可以很好地分析出借贷人真实的工作，生活等情况，从而在反欺诈方面有着得天独厚的优势. 在审批环节，需要核实借贷人填报的信息。利用位置数据分析服务，能够将借贷人填写的地址信息与采集分析后的位置信息进行比对，如果发现其提供了一个有较大偏差的地址，则该地址信息存在虚假的可能性。

* 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

Conclusion

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.”