z5163468 Weijia Liu
z5201685 Yizhang He

1. Introduction

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. After thoroughly consideration and planning, we choose the challenging project of fraud detection.

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. Hence, these papers are great analysis and guide for us to do our own project.

3. Software and data description

Yuumi Insurance offered 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, time-stamped 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.

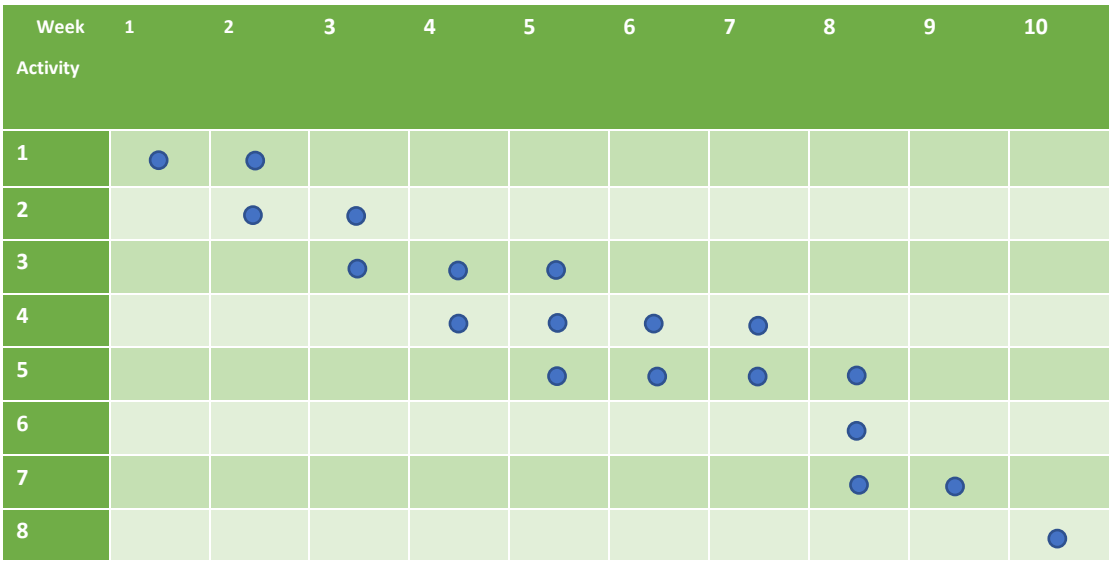
If the client chooses to accept the insurance policy and make a payment, this information is recorded as well, without the specific amount of money the client paid. When a claim is made, the company will look through the client's claim history

and examine all the information related to the client and decided if they will accept or deny the claim. Scripting languages like Python or Shell will be used for data cleaning purposes. Since the data is in the form of log, one may need to extract features of each piece of data and integrate them into a more tabular fashion. The proposed methods have well-written packages in programming languages like R and Python, both of which are popular tool for data mining and analytics. We might also need to write our self-defined code as helper functions.

4. Activities and schedule

The following activities should be conducted throughout the ten weeks.

1. Literature review and compare possible approaches.
2. Clearly define the problem and deeply understand the dataset.
3. Data pre-processing, cleaning and preparation.
4. Design, develop and test models and interpret the output
5. Compare model performance using pre-defined performance metrics.
6. Choose the model with the best performance.
7. Write the report
8. Presentation



5. Reference

[1] Peng Y., Kou G., Sabatka A., Matza J., Chen Z, Khazanchi D., Shi Y. peer_reviewed Lecture Notes in Computer Science,2007, Vol.4489(3), pp.852-858

[2] Capelleveen, Guido Cornelis van, “Outlier based predictors for health insurance fraud detection within U.S . Medicaid,” [D]. the requirements for the degree of Master of Science in Business Information Technology at the University of Twente, 2013.

[3] Hou Z. , “Application of Fraud identification clustering algorithm in the CRM of auto insurance,”Journal of Changchun Institute of Technology. Vol.1,2009, pp.97-99.

[4] Phua C., Lee V., Smith K., & Gayler R. (2012). A Comprehensive Survey of Data Mining-based Fraud Detection Research. Computers in Human Behavior, 28(3), 1002-1013. http://doi.org/10.1016/j.chb.2012.01.002