Data3001 - Assignment 1: Choice of a Problem
Insurance Kids
Serena Xu (z5164959) | Dean Hou (z5163159) | Neel Iyer (z5165452)

## Section 1 - Introduction and motivation

We are given log data for a hypothetical insurance company, Yuumi. Yuumi insurance covers house and personal possessions against loss or damage caused by the following insured events: storm, fire, lightning, theft and escape of liquid. The insurance policy is priced according to the risk of the customer based off an array of information, done by actuaries within the company. Irrespective of the price of the insurance all policy has a coverage $25,000.

We will be looking at the issue of fraud detection. Currently, suspicious claims are detected using basic computer algorithms, which are further investigated by humans who cross check the customer's historical asset information. We will make the assumption that Yuumi does not yet have a machine learning process to check fraud.

Insurance fraud is a huge issue in the industry costing the Australian economy $2.2 billion every year according to 2018 statistics by the Insurance Fraud Bureau Australia 2018. Any form of fraud is a cost for the insurance company as well. We have chosen two scenarios where fraud makes a negative impact on the business:

1. Company making unnecessary payments to dishonest customers
    a. A customer makes a false claim about an accident and the company makes a payment.
    b. The customer buys a premium after the accident has already happened, and makes a claim.
    c. The customer has an otherwise legitimate claim that is exaggerated or "built up".
    d. The customer makes a claim for an injury not related to the accident
2. Company acquires extra human resources to check for fraud

In both scenarios, Yuumi loses money due to customer fraud and dishonesty. Our mission is to detect such scenarios early so the company does not make any extra payments. We aim to use the log data provided to us by Yuumi to predict claims that are fraudulent, thus saving the company resources and money .

Some further considerations include the definition of accuracy. Are false negatives more costly to the company than false positives? This will involve researching the cost of human resources checking for fraud, and the actual cost of fraudulent payments.

## Section 2 - Brief literature review

Microsoft Azure Stream Analytics has a real-time fraud detection service, which is scalable to companies that purchase the Azure infrastructure. However this is often limited to its capabilities, especially transforming raw data. Furthermore, companies must subscribe to obtain access, and is very costly. Thus, we believe that our solution for Yuumi is more cost effective and tailored to suit the obtained quality of data.

Many machine learning models have been made for fraud detection, and also very similar models for spam and anomaly detection. However they do not utilise insurance data but rather other forms of data such as financial and credit card data, so the models will not be as useful. Thus, we have decided to continue with this project as it will be developed specifically for Yuumi and its data.

We have also discovered well developed NLP processes for text analysis, such as the word2vec package, TFIDF and bag of words models. We will be implementing these in our text cleaning process to tabularise the customer data using word similarities.

## Section 3 - Software and data description

We will use the Yuumi insurance log dataset provided by Jacky Koh, as well as finding other log datasets to clean and analyse using our model. The data we have been provided is a .csv file called data. The data contains two columns: message, which contains a column of strings that records the user's actions such as inputs, user device, user id, type of action etc.; and timestamp, which contains a column of UNIX epoch times as floats recording when the user accessed the online system. The message column will need to be tokenized to gain further insights. The data has a about 3 million rows, and has no missing values. However this does not mean that the data has erroneous values. The log data includes data on which user's claims were rejected and the reason they were rejected (including rejected due to fraud) which can be used to determine the validity of an insurance claim.

The software we intend on using is Jupyter Notebook, and we will use Github for collaboration. Our notes and reports will be written on Google docs, stored in a Google Drive for collaborative access. Jupyter Notebook is a web based interactive environment for creating notebooks. This notebook is a JSON document that contains a list of input and output cells as well as text cells. These cells can be used to produce graphs, plots and rich media using the Python programming language. Python also includes many packages to manipulate data such as Pandas, as well as packages on NLP and machine learning that can help us create models on the data.

Github is a web-based hosting service for code/document version control that uses the Git branch structure. It provides a platform for access control, bug tracking, feature requests as well as source code management. Projects on Github can be accessed through the terminal command line interface and all the standard Git commands work with it. We plan to upload out Jupyter Notebook onto our Github for shared access to the code, using branches to maintain stable builds of our code.

Section 4 - Activities and schedule

Activities are scheduled for week 4 to week 10 (7 weeks)

| Task | Description | Schedule |
|---|---|---|
| Clean/preprocess data | - Tokenize message column, One hot encode variables which can be discretised<br>- Convert timestamp into local datetime<br>- Get rid of unwanted/unrelated columns | 1 week (week 4) |
| Perform EDA | - Look for patterns in the data with graphs and find other log data that could help<br>- Contextualise and determine how accuracy can be measured | 1 week (week 5) |
| Research models and how to use them | - Research models related to this type of data and understand how they work<br>- Choose models that can be used on this project/understand how to build a model that can be used on this project | 1 week (week 6) |
| Create/train and test models | - Use fraud column (quote completed or uncompleted due to fraud)as response variable in binary format (0/1)<br>- Use other parameters as features for the model<br>- Test at least 3 different models<br>- Interpret/analyse model results | 2 weeks ( 7-8) |
| Create report | - Write report, add documentation to notebook explaining features, models (how they work) and interpretation of results | 1 week (week 9) |
| Create presentation | - Create presentation and write notes on contents of presentation based on report | 1 week (week 10) |

Section 5 - References


1. Silcox, A. (2019). *Insurance Fraud Costs Australians $2.2Billion Annually - SME Business Insurance Brokers*. [online] SME Business Insurance Brokers. Available at: https://www.pscconnect.com.au/insurance-fraud-costs-australians-2-2billion-annually/ [Accessed 23 Jun. 2019].
2. Docs.microsoft.com. (2019). *Anomaly detection in Azure Stream Analytics*. [online] Available at: https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-machine-learning-anomaly-detection [Accessed 23 Jun. 2019].
3. Docs.microsoft.com. (2019). *Real-time fraud detection using Azure Stream Analytics*. [online] Available at: https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-real-time-fraud-detection [Accessed 23 Jun. 2019].
4. Medium. (2019). *Natural Language Processing: Text Data Vectorization*. [online] Available at: https://medium.com/@paritosh_30025/natural-language-processing-text-data-vectorization-af2520529cf7 [Accessed 23 Jun. 2019].
5. Towards Data Science. (2019). *Detecting Credit Card Fraud Using Machine Learning*. [online] Available at: https://towardsdatascience.com/detecting-credit-card-fraud-using-machine-learning-a3d83423d3b8 [Accessed 23 Jun. 2019].
6. Towards Data Science. (2019). *Detecting Financial Fraud Using Machine Learning: Winning the War Against Imbalanced Data*. [online] Available at: https://towardsdatascience.com/detecting-financial-fraud-using-machine-learning-three-ways-of-winning-the-war-against-imbalanced-a03f8815cce9 [Accessed 23 Jun. 2019].
7. Effective Coverage. (2019). *How Do Insurance Companies Detect Fraud?*. [online] Available at: https://www.effectivecoverage.com/3684/insurance-companies-detect-fraud/ [Accessed 23 Jun. 2019].
8. NetMap Analytics. (2019). *fraudulent insurance claims exposed by data visulisation*. [online] Available at: https://netmap.com.au/insurers/ [Accessed 23 Jun. 2019].
9. GitHub. (2019). *GitHub features: the right tools for the job*. [online] Available at: https://github.com/features [Accessed 23 Jun. 2019].