Fraud Detection in Forensic Data Analytics

BACKGROUND

Forensic data analytics is a rapidly growing field in the forensic industry that seeks to use data driven analytics to provide solutions to fraudulent accounting practices. A forensic accountant will use statistical approaches to determine the likelihood of fraudulent activity. For example, a traditional statistical approach such as Benford's law is used to detect fraud in accounting for expenses. This is because the leading digits in numbers follow a distribution of log(n+1) - log(n), where n is the number of the leading digit. However, as fraud adapts and becomes more advanced in avoiding being traced, it becomes increasingly important for firms to adapt by adopting new practices to detect the occurrence of fraudulent activities. By using supervised learning techniques such as regression, random forest and neural networks, historical transactions fraudulent and non-fraudulent can be used to classify fraudulent activities as high risk or low risk (Deloitte 2018).

AIMS

The aim of this project is to develop an approach to predicting fraudulent transactions by analysing previous fraudulent transactions to predict and detect future instances of fraudulent behaviour. To do so data associated with previous fraudulent transactions will be collected and cleaned for use. Different machine learning classifiers will be evaluated and compared and the most important attributes for each classifier chosen. Once the correct data, model and parameters are chosen, they will be deployed to analyse new situations.

RESEARCH PROJECT

Significance and Importance

Fraudulent transactions have always been an issue in the financial world and as a result it is important to monitor fraud and hold parties accountable when it is committed. Globally due to the incidence of fraudulent transactions in banking, the losses to banks is estimated to exceed 31 billion dollars globally. The other issue is that from a period of 2013 to 2018, the number of fraudulent transactions has increased by 34% per annum, whilst the number of transactions that were successful had increased by 31% per annum. As a result, it is important to switch to an approach that is proactive rather than reactive. By flagging potentially fraudulent transactions before they are processed, extra scrutiny can be shown to prevent these transactions from occurring (McKinsey 2018).

Innovation of Project

Because typical techniques in the forensic industry are reactive to crime, adopting a data driven approach that uses statistical analysis combined with machine learning is likely to proactively reduce the incidence of fraud. This will not only reduce the amount of successful fraudulent transactions significantly but discourage parties from engaging in fraudulent transactions, thus lowering the total amount of fraudulent transactions. The attempt

to be proactive rather than reactive is a major innovation in this field, and by employing machine learning techniques on the correct attributes, it can be very accurate at predicting fraud and much more accurate than other techniques.

Solving the fraud problem

Data containing multiple transactions will be combined from various data sources to build a table for data analysis. To process this data, frameworks such as spark and Hadoop will be used to process date before it is used for analysis. The attributes will include things such as the average transaction amount, yearly income, the average time between each transaction, and the age of account. The final table will contain up to 200 different attributes that will serve as features. The target variables will rank the risk of a transaction. An example of this can be seen in the table below. This is simply a demonstration, as more attributes will be considered when developing a dataset that can be used for analysis. It is important to consider as many attributes that are related to an account when compiling data at the preliminary stage of the investigation. See table1 for an example.

Once data is collected from banks on different accounts and whether there was any level of fraud the data can be cleaned so that a machine learning classifier can make better use of the data. Cleaning will involve splitting columns with dates, removal or filling missing values, removing characters such as strings from data, one hot encoding or regular encoding of categorical data. Once the data is cleaned some of the data will require pre-processing such as normalisation to finally have the data in a format that a machine learning classifier can easily understand and process information with. Most of this can be done using the pandas and sklearn libraries on python. Once that is done, the data will be split into training and test data, and a validation set included. a classifier such as random forest or XGBoost will be used to fit the dataset and determine the most important features. Features can also be determined by using Pearson correlation between the target variable. Once the correct features are selected the models can be tuned for proper use. This can be done with the use of gridsearchCV or manually. A classifier can be looped with multiple parameter setting and then compared by comparing accuracy, precision and recall scores to determine which parameters work best for the model. The model can then be compared with others to determine the best model. It is important to recognise that when classifying fraud that minimising false positive and maximising true positives is essential as cases must be followed up on once flagged which takes time and money. Thus, models with a high recall are preferred as they can maximise the number of true positives and minimise the number of false negatives. The model can then be evaluated on test data and a model then built to predict new transaction data.

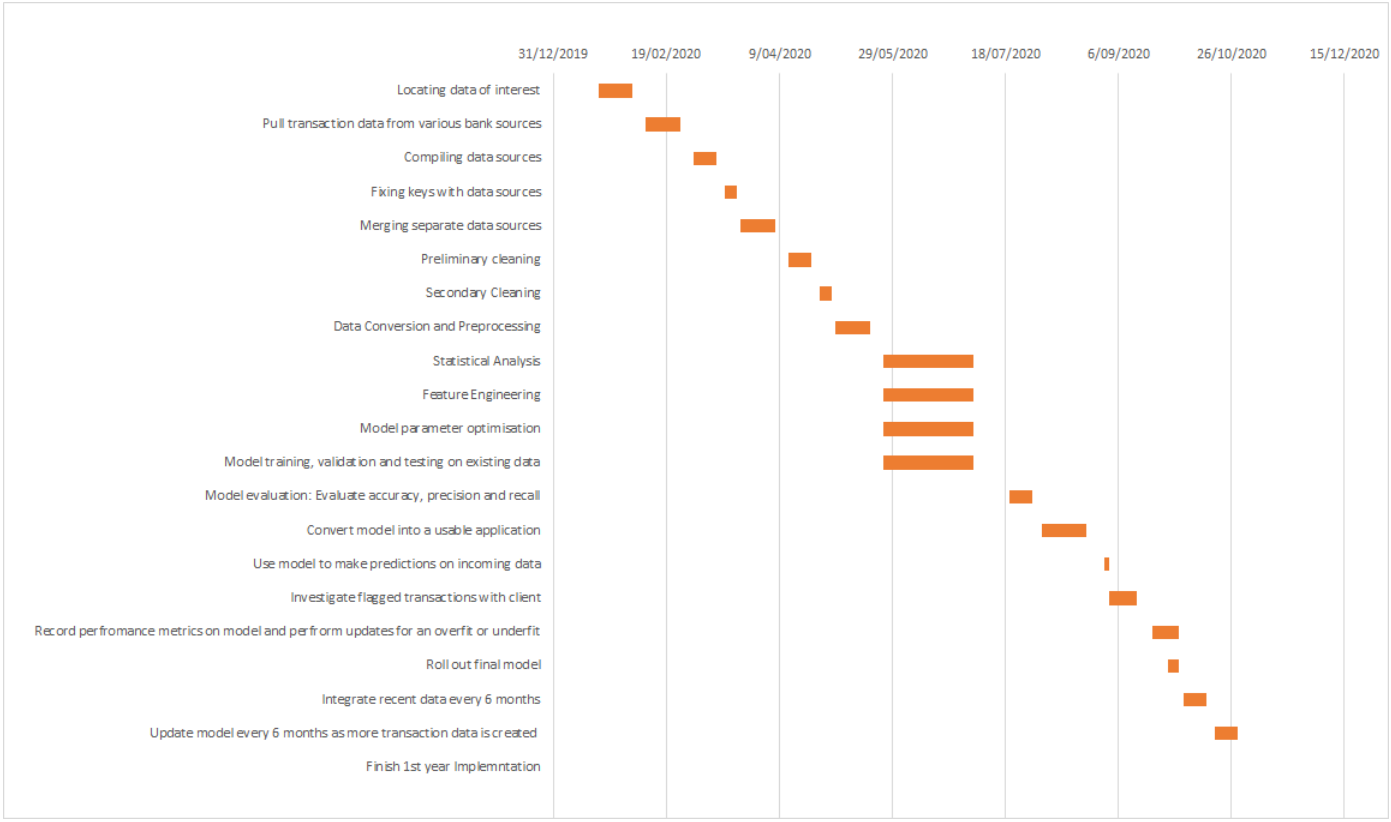
In [2]:

```
#Table 1: Example attributes for analysis

import pandas as pd

df = pd.DataFrame(data = {'Average transaction amount (per day)':[100, 300,5000],'Income':
                          'Average time between each transaction':['04:30:02','04:00:02',
                          'Age of sending account (years/month/day)':['4/3/6','3/4/12','8/
                          'Age of receiving account (years/month/day)':['5/2/1','2/1/30','
                          'Fraud Risk':['Low','Low','High']
                         })
df.head(3)
```

Out[2]:

	Average transaction amount (per day)	Income	Average time between each transaction	Age of sending account (years/month/day)	Age of receiving account (years/month/day)	Fraud Risk
0	100	60000	04:30:02	4/3/6	5/2/1	Low
1	300	60000	04:00:02	3/4/12	2/1/30	Low
2	5000	20000	00:05:00	8/3/2	0/0/2	High

Project Timelines with Gantt Chart



Project outcomes and benefits

The final model should be able to flag 70-80% of fraudulent transactions that are found. This will save the businesses large amount of money over time. For one client they had saved 131000 USD in the first few weeks and over a year more than 2000000 USD (McKinsey 2018). These models can run for long timeframes before

being optimised again and thus can save large amounts of money. Costs for each client are usually front ended as later costs are to maintain models which is less labour intensive. It is also important to recognise that this work can be replicated for multiple banks and even insurance companies. Doing so would create more revenue for the project. Revenue from a client would be calculated on approximately a third of the projected cost savings over 7 years. The client can optionally choose to continue using the model after 7 years in which more revenue is possible. This would be approximately 4,550,000 USD of revenue on 14,000,000 USD of cost savings.

PERSONNEL

Most of the costs of a project are labour intensive as IT projects take the data from client sites and store it cheaply on the cloud. Fixed costs are divided through the years as these are shared amongst other projects that will be embarked on in following years. It is importnat that the project will be complete within a year. The othe years will be simply keeping in touch with the client.The table breaks down the cost for one project below:

Profit over 7 years = Revenue over 7 years (4,550,000) – Costs over 7 years (2,036,000) Profit over 7 years = 2,514,000 Average profit per year = 359,000 (Just one project!)

In [3]:

```
#Table 2: Projected cosrts over the span of the project
pd.read_csv('C:/Users/sashv/OneDrive/Documents/costs.csv')
```

Out[3]:

	Cost Type	Cost Y1	Cost Y2	Cost Y3	Cost Y4	Cost Y5	Cost Y6	Cost Y7
0	Project manager	200000	20000	20000	20000	20000	20000	20000
1	Data Engineering Team	450000	45000	45000	45000	45000	45000	45000
2	Data Science Team	550000	55000	55000	55000	55000	55000	55000
3	Various Software and Cloud	20000	3000	3000	3000	3000	3000	3000
4	Hardware	20000	0	0	4000	0	0	4000
5	Office Costs	50000	5000	5000	5000	5000	5000	5000
6	Total Cost	1290000	123000	123000	127000	123000	123000	127000
7	Revenue	650000	650000	650000	650000	650000	650000	650000
8	Profit	-657000	534000	534000	530000	534000	534000	530000

Video Pitch

Link: https://www.youtube.com/watch?v=0Q6FCy4jzXw&feature=youtu.be (https://www.youtube.com/watch?v=0Q6FCy4jzXw&feature=youtu.be) https://youtu.be/0Q6FCy4jzXw (https://youtu.be/0Q6FCy4jzXw)

References

Fancher. D, Lalchand. S, Rial. E, Balasubramaniam, 2018. S, Forensic analytics in fraud investigations (White paper), Deloitte, viewed 7th October 2019 https://www2.deloitte.com/us/en/pages/advisory/articles/forensic-

analytics-in-fraud-investigations.html (https://www2.deloitte.com/us/en/pages/advisory/articles/forensic-analytics-in-fraud-investigations.html)

Mckinsey Analytics, Risk, Mckinsey, viewed 7th October 2019. <https://www.mckinsey.com/industries/financial-services/our-insights/combating-payments-fraud- (https://www.mckinsey.com/industries/financial-services/our-insights/combating-payments-fraud-) and-enhancing-customer-experience>

Mckinsey Analytics 2018, Combating payments fraud and enhancing customer experience, Mckinsey, viewed 7th October 2019 https://www.mckinsey.com/business-functions/mckinsey-analytics/careers/risk (https://www.mckinsey.com/business-functions/mckinsey-analytics/careers/risk)