HENRY WANDERA 17253129

COS720: COMPUTER AND INFORMATION SECURITY

INSURANCE PRACTICAL ASSIGNMENT

MY TASKS AND CHOICES

DATA GENERATION

I used Microsoft SQL Server 2016 to create my database, designed the table fields and populated my database with 100,000 fabricated insurance claims using the SQL Data Generator 3.

The following are fraudulent claims identified by the fraud claim reason and fraudulent claim indicator columns;

1. **Did not pay premium:** where a client who has never paid premium claims for compensation,
2. **Has no policy:** A fraudster without an insurance cover/account claims for money.
3. **Inflated figure:** Some insurance company employees connive with clients and give them more money than what they claimed for,
4. **Claimed before date loss:** Insurance companies discover that some clients ask for compensation even before making losses,
5. **Expired policy:** A fraudster/ client forget that his insurance cover expired and requests for compensation.

The claims were made basing on the following losses; fire, collision accident, theft, illness, death, vandalism, flood, marine transit, loss liability.

DATA CLEANING

I used the classification algorithm by classifying the existing data into several classes based on criteria that are pre-defined that is difference between sum insured and claim amounts, differences in various dates. This helped me to specify samples of incorrect data from the correct data by using SQL scripts and excel to clean my data.

EXPLORATORY DATA ANALYSIS

I used anaconda tool to employ pandas in python. My data was clean and reliable for this stage but I continued modify it to fit the machine learning stage. The source codes and visualizations like histograms, bar chats, scatter diagrams and box plots are attached for more information concerning the distributions of different variables.

DATA SCALING

Data scaling is important when the range of values of raw data varies widely. Some objective functions cannot work properly in machine learning algorithms without normalizing the dataset. For example, most classifiers calculate the distance between two points by Euclidean distance. Since I was using amounts to train my model, I did not rescale my figures but I calculated the difference between dates for my model to work well.

I used decision tree algorithm for machine learning which requires little effort from users for data preparation; it does not require rescaling or normalising variables because the tree structure will remain the same with or without the transformation.

PRIVACY PRESERVING DATA MINING

I used SQL scripts to implement the suppression and generalisation methods of K-anonymity algorithm. The names, streets, Date of birth of all clients were anonymized by replacing their values with asterisk '*'. The date of birth was fixed 1900/01/01 to all clients therefore their personal information cannot be compromised by any outside attacker or employee of the company except some administrators who have the privileges of viewing all the clients' information in my database.

MACHINE LEARNING (DECISION TREE ALGORITHM)

I used decision tree algorithm to train my model for any predictive analytics. My points of interest were amounts, dates and the fraud claim indicator.

The dependent variable fraud claim indicator **(Y)** was trained on **X** (both amounts and dates separately) that is **.fit(X, Y)** and predictions were made using various inputs as shown in the tables below.

PREDICTIONS BY DATES

Policy start date	**Policy end date**	**Date of loss**	**Claim date**	**Result**	**Reason**
2010-12-21	2016-12-21	2013-12-21	2014-05-16	0 (not fraud)	Correct dates
1999-12-21	2000-12-21	2012-12-10	2012-12-16	1 (fraud)	Expired policy
2002-01-01	2004-12-31	2003-06-10	2003-06-25	0 (not fraud)	Correct dates
2004-01-01	2008-12-31	2006-09-13	2006-05-05	1 (fraud)	Claiming before loss

PREDICTIONS BY AMOUNTS

Claim amount	Total premium	Sum insured	Result	Reason
200,000	70,000	100,000	0 (not fraud)	Correct amounts
200,000	0	100,000	1 (fraud)	No premium
700,000	700,000	700,000	0 (not fraud)	Correct amounts
555,000	15,000	700,000	1 (fraud)	More sum insured than claim

Decision trees implicitly perform variable screening, the top nodes on which the tree is split are essentially the most important variables within the dataset and the feature selection is completed automatically.

It requires little effort from users for data preparation it does not require rescaling or normalising variables because the tree structure will remain the same with or without the transformation and the nonlinear relationships between parameters do not affect tree performance. What I liked most about decision tree is that it is easy to interpret, understand and implement

APPLYING DATA SCIENCE TECHNIQUES TO CYBER SECURITY

Big data and data science are good technology weapons to impasse terrorism, fraud, cyber criminals and many other attackers (Khushbu Shah, Dec, 2015). They now ease intelligence led investigation processes through data analysis so that agencies can detect security threat. In this assignment, I have discovered and used the following techniques;

Collection/generation of data: The most important thing is collection and management of security data or information about all transactions by the company as this greatly ease investigations and exploit vulnerabilities for example huge data on potential insurance claims, terrorist behaviour from various data sources like online conversation. If this data is categorised and cleaned, data analysis can reveal the patterns to threats that are about to happen. I found vandalism cases contributing much of the claims and men on average were paying fewer premiums but claiming for more money than women. Therefore From time to time, we need to review our data, understand it, separate correct data from incorrect data to see if there are no data misplaced.

Visualization: I realised that data visualization is a powerful technique in understanding the patterns of any dataset therefore behaviour patterns of fraudsters and terrorists can be traced in order to detect of predict their actions. We can visualize data instantly answer questions that we never thought to ask before.

Machine learning: After training my model I was impressed by the results of the predictions, therefore my code can be able to detect fraud and notify the management about the fraudulent behaviour in the information provided by the client. For example the cyber security arm RSA of the US big data company EMC uses machine learning and advanced big data analytics methodologies to prevent online fraud. They have detected approximately 500, 000 attacks in 8 years – half of which were identified in 2012 alone (Khushbu Shah, Dec, 2015).

Anonymization: This technique is able to protect personal information about clients and even when there is a malicious attack to the company's database by any fraudster with intensions of modifying or identifying clients' information, he will fail to identify the customers' most sensitive information. For example the K-anonymity algorithm (Xu, 2016) I used is has been recognised effective against neighbourhood attacks in social networks.

SQL commands: My database is also protected against SQL injections as clients have limited access privileges unless granted chance.

Bibliography

Khushbu Shah, Dec, 2015. *Big Data and Data Science for Security and Fraud Detection.* [Online] Available at: http://www.kdnuggets.com/2015/12/big-data-science-security-fraud-detection.html

Xu, L. ,. J. C. C. Y. W. J. a. R. Y., 2016. A framework for categorizing and applying privacy-preservation techniques in big data mining.. *Computer,,* Volume 49(2), , pp. pp.54-62..