Online Fraud Detection

Lokesh Punati 11514358

Rahul Mandaloju 11514828

Github link : https://github.com/Lokeshpunati/onlinefrauddetection.git

Goals and Objectives:

Machine learning involves artificial intelligence, and it is used in solving many problems in data science. One common application of machine learning is the prediction of an outcome based upon existing data. The machine learns patterns from the existing dataset, and then applies them to an unknown data set in order to predict the outcome. Classification is a powerful machine learning technique that is commonly used for prediction. Some classification algorithms predict with satisfactory accuracy, whereas others exhibit a limited accuracy.

Prediction of frauds happening online is regarded as one of the most important subjects in the section of clinical machine learning. It is difficult to identify fraud because of several contributory risk factors such as enormous data, Imbalanced data, Data availability, Misclassified data and many other factors. Due to such constraints, scientists have turned towards modern approaches like Machine Learning for predicting the frauds.

**Challenges Faced:**

1. Enormous Data is processed every day and the model build must be fast enough to respond to the scam in time.
2. Imbalanced Data i.e most of the transactions *(99.8%)* are not fraudulent which makes it really hard for detecting the fraudulent ones
3. Data availability as the data is mostly private.
4. Misclassified Data can be another major issue, as not every fraudulent transaction is caught and reported.
5. Adaptive techniques used against the model by the scammers.

The topic of fraud detection is so large that entire textbooks, training programs, and even companies are devoted to it exclusively. In addition to the complexity associated with this pattern of offending, there are many different “flavors” of fraud to include phishing, [spear phishing](https://www.sciencedirect.com/topics/computer-science/spear-phishing), breakout fraud, and ATM skimming just to name a few. Specific patterns of fraud and associated consequences also may differ based on the specific domain. For example, in addition to the financial consequences associated with medical identity theft, material changes to the medical record also can have significant consequences for the patient.

**METHODOLOGY:**

**1.Fraud scenarios and their detection:**

**Insurance claims analysis for fraud detection:**

Insurance companies spend several days to weeks assessing a claim, but the insurance business is still affected by scams. The most common issues are property damage, car insurance scams, and fake unemployment claims. The ticket to successful detection is a good dataset and carefully selected models.

**Fake claims.** Semantic analysis is a machine learning task that allows for analyzing bothstructured, table-type data, and unstructured texts. The feature helps detect fake and falsified claims in the insurance industry. For example, it improves car insurance claims processing. Machine learning algorithms analyze files written by insurance agents, police, and clients, searching for inconsistencies in provided evidence. There are many hidden clues in these textual datasets. The rule-based engines don’t catch the suspicious correlations in textual data, and fraud analysts can easily miss important evidence in boring investigation files. That’s why analyzing claims is one of the most promising spheres for machine learning applications.

**Duplicate claims and overstating repair cost.** Smart ML-backed algorithms are alsoefficient in duplicate claims detection or inconsistencies in car repair cost. Classifying data in repair claims solves the problem by uncovering hidden correlations in claim records or even behaviors of insurance agents, repair services, and clients. For example, the repair service company may provide higher pricing for the customers of a specific agent.

Let’s have a look at the results of the AI-based research of insurance vehicles claims conducted by Wipro. The company explored four datasets with such features as a vehicle style, client gender, marital status, license type, injury type, loss date, claim date, police notification date, repair amount, sum insured, market value, etc.

1. A.pre-research analysis disclosed:
   * Fraudulent claims are more likely not reported to police.
   * Old vehicles are more likely to be involved in fraud.
   * Eighty percent of accidents that happen during holidays involve fraud.
   * Scams are more likely to involve third parties than legitimate claims.

Then the data was processed using five different machine learning algorithms: Logistic Regression, Modified Multi-Variate Gaussian, Modified Randomized Undersampling, Adjusted Minority Oversampling, and Adjusted Random Forest. Eventually, the best results were achieved by the Modified Randomized Undersampling model that showed 79 percent accuracy.

**REFERENCES:**

1. Credit Card Fraud Detection Based on Transaction Behavior -by John Richard D. Kho, Larry A. Vea published by Proc. of the 2017 IEEE Region 10 Conference (TENCON), Malaysia, November 5-8, 2017
2. L.J.P. van der Maaten and G.E. Hinton, [Visualizing High-Dimensional Data Using t-SNE](https://lvdmaaten.github.io/publications/papers/JMLR_2014.pdf) (2014), Journal of Machine Learning Research
3. Machine Learning Group — ULB, [Credit Card Fraud Detection](https://www.kaggle.com/mlg-ulb/creditcardfraud) (2018), Kaggle
4. Nathalie Japkowicz, [Learning from Imbalanced Data Sets: A Comparison of Various Strategies](http://www.aaai.org/Papers/Workshops/2000/WS-00-05/WS00-05-003.pdf) (2000), AAAI Technical Report WS-00–05