|  |  |
| --- | --- |
|  | |
| Scientific Research + Literature –Assessment 2  TU060 : Individual Annotated Bibliography on Credit Card Fraud Topic | |
|  | |
| Ciaran Finnegan – Part Time – First Year 2021/2022  MSc in Computer Science (Data Science)  Student No : D21124026  27/3/2022 |  |
|  |  |

Table of Contents

1 Technical Topic – Data Analytics Specialism 1

1.1 Assignment Submission Topic 1

1.2 High Level Description of Journal / Articles in this Report 1

2 Bibliography of Sources 1

2.1 Anomaly Detection Algorithms for Credit Card Fraud Detection 1

2.2 Feature Engineering for Credit Card Fraud Datasets 1

2.3 Just Being Really Clever with Credit Card Fraud Detection 1

2.4 Just Being Clever with Credit Card Fraud Detection 1

2.5 I Mean Really Clever with Credit Card Fraud Detection 1

3 CA Guidance (remove before submission) 1

# Technical Topic – Data Analytics Specialism

## Assignment Submission Topic

The chosen topic for this research bibliography assignment is;

*Techniques to improve the Machine Learning workflow processes when attempting to detect credit card fraud with highly imbalanced datasets.*

## High Level Description of Journal / Articles in this Report

Section 2 of this assignment report provides the more detailed overview of the articles/journals selected and reviewed for this particular topic of Machine Learning techniques in Credit Card Fraud detection.

Below is a high-level list of the subjects covered in the selected articles/journals;

* ML algorithm selection for unbalanced credit card fraud datasets. (Section 2.1)
* Feature Engineering of credit card fraud datasets. (Section 2.2)
* The next smart idea in cc fraud detection. (Section 2.3)
* A really clever idea in cc fraud detection. (Section 2.4).
* Super smart ideas in cc fraud detection. (Section 2.5)

# Bibliography of Sources

## Anomaly Detection Algorithms for Credit Card Fraud Detection

***Reference***

Ceronmani Sharmila, V., R., K., R., S., D., S., & R., H. (2019). Credit Card Fraud Detection Using Anomaly Techniques. *2019 1St International Conference On Innovations In Information And Communication Technology (ICIICT)*, *1*(1), 1-4. doi: 10.1109/iciict1.2019.8741421

***Aim***

To demonstrate the efficiencies of Local Outlier Factor (LoF) and Isolation Forest (IF) algorithms in detecting credit card fraud. The paper describes that instances of credit card fraud ‘*take place every once in a while’* and that these anomaly detection algorithms can be shown to be less computationally expensive than more ‘generic’ algorithms.

***Methods***

A historical (Kaggle) dataset of 27K credit card transaction is loaded into an ML experiment run in Jupyter Notebooks, with all required processing libraries preloaded.

The data is split between test and training sets, in line with standard ML workflow processes, and applied in sequence through the unsupervised LoF and IF algorithms. These algorithms provide and aggregate framework to identify the ‘anomalies’, which are the transactions marked as fraudulent.

Data visualisation techniques are used to display the key features in the dataset, and their relevance to the anomaly scoring.

***Conclusions***

This conference article declares that the LoF and IF used by the authors is quicker and more efficient that general Decision Tree approaches and is also more understandable than Neural Network techniques.

***Gaps in the Research:***

The range of other algorithms against which LoF and IF are compared is limited to a brief discussion on Decision Trees and Neural Networks.

The paper does not contain any tabular data with which to strengthen the argument that LoF and IF algorithms are faster at detecting fraud.

***Critique***

There is important reference in the paper than speed can be important in credit card fraud detection, particularly with real-time requirements. The implementation of the aggregation LoF and IF algorithms is well thought out and the logic sound.

However, although the visualisations on feature importance are interesting, what is missing is tabular data on actual model accuracy, including the False Negative rate, and benchmark on speeds against other algorithmic approaches.

## Feature Reduction for Credit Card Fraud Datasets

***Reference***

Lima, R. and Pereira, A., 2017. Feature Selection Approaches to Fraud Detection in e-Payment Systems. *Lecture Notes in Business Information Processing*, [online] pp.111-126. Available at: <https://www.researchgate.net/publication/313731885\_Feature\_Selection\_Approaches\_to\_Fraud\_Detection\_in\_e-Payment\_Systems> [Accessed 11 September 2020].

***Aim***

E-Payment systems tend to generate not just a high volume of records but also a significant number of features for each transaction. Effective feature engineering of such datasets is adversely impacted by the high data imbalance of fraud and non-fraud records. This research paper demonstrates a scenario that shows that certain resampling techniques can counter this imbalance and deliver effective feature reduction and improved fraud detection accuracy.

***Methods***

A real world historical dataset is obtained from a Latin American e-payment system (PagSeguro). All the transactions in this dataset are labelled ‘fraud’ or ‘non-fraud’, although the volume of ‘fraud’ records is only just over 1% of the entire dataset.

A series of Supervised Machine Learning experiments are conducted using a combination of seven re-sampling techniques and three feature selection approaches. Three Classification models (Bayes, Decision Tree, and Logistic Regression) are applied, using 8-Fold Cross Validation, and accuracy is measured primarily using the AUC (Area Under Curve) metrics, with particular attention to the True Positive rate.

***Conclusions***

The experiments conducted by the researchers showed a 57% increase in accuracy when their re-sampling techniques were applied to the dataset before feature reduction computations. This comparison is based on fraud detection results conducted in the past by the PagSeguro Corporation itself.

Within the scope of these experiments, the best results were obtained by a re-sampling technique called ‘Random Undersampling’ followed by the ‘Relief’ approach to feature reduction.

***Gaps in the Research:***

There are no details provided on the fraud detection methods previously employed by the PagSeguro Corporation. The assertion from the researchers that their methods improved on earlier fraud detection attempts by 57% lacks a degree of context.

The number of records in the dataset is not given in this paper, and this is an unusual omission. (This dataset appears in other research papers so it is known to contains over 1 million entries).

***Critique***

The source dataset contains 380 columns and the experiment described in the research paper applies a logical and effective approach to reduce this number of features prior to classification. The difficulties for feature reduction with highly imbalanced datasets are well described the results of the experiment are well tabulated.

A proficient Data Analyst should be able to replicate these experiments with other datasets, and generate new benchmarks. Only three Classification models are applied in these experiments so a more comprehensive suite of algorithms would be desirable to further ground the accuracy assertions of the researchers.

## Assessment of SVM and Random Forest Algorithms in Unbalanced Credit Card Datasets

***Reference***

Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. Decision Support Systems, 50(3), 602–613. <https://doi.org/10.1016/J.DSS.2010.08.008>

***Aim***

To evaluate the effectiveness of Support Vector Machine (SVM) and Random Forest (RF) algorithms in credit card fraud detection in a highly imbalanced dataset. The effectiveness of the ML experiments will be compared against what are considered to be ‘established’ results from prior Logistic Regression (LR) analysis.

***Methods***

Research is conducted through a series of Supervised Machine Learning experiments, with a labelled dataset generated from real work information by an international credit card operator.

The original dataset contained nearly 50 million records but this was reduced to a sample of 350K fraud/non-fraud records, within which fraud records constitute less than 1% of the overall volume of credit card transactions.

The ML experiments conduct an iterative set of resampling, feature engineering, and modelling steps and accuracy is primarily measured by the AUC metric.

***Conclusions***

The ML experiments using RF algorithms generated better fraud detection performance metrics than both SVM and Logistic Regression.

The primary benchmark table in the paper was produced using an oversampled fraud dataset (10% fraud), which produced an AUC value of 0.953 for RF, 0.942 for LR, and 0.908 for SVM.

The tabular and graphical results from many of the other experiments showed that Logistic Regression outperformed many of the SVM runs, despite being considered a more ‘traditional’ model.

***Gaps in the Research:***

The dataset upon which the ML experiments were conducted did not contain any timestamp features. It was therefore not possible to look at time series or sequence analysis and derive any observations. The paper acknowledges this deficiency and sees the inclusion of timestamp data as an important future consideration.

There is also very little effort invested by the researchers in hyper-parameter tuning for any of the algorithms. The authors concede that this could have contributed to the relatively poorer performance by the SVM models.

The source dataset is from 2007, so would not contain more contemporary credit card attributes such as ‘chip and pin’ data, but it is still feature rich and does not contain any obsolete characteristics.

***Critique***

Although this paper is nearly twelve years old, much of the underlying analysis remains valid, particularly as there are still relatively few credit card fraud datasets in general circulation.

This research paper compliments the other articles in this bibliography report in that there is significant upfront emphasis and guidance on re-sampling techniques due to the general imbalance against fraudulent records. The experiments are described in detail in the paper and could be reapplied to newer, more up to data datasets, and still provide valuable insight.

## Just Being Clever with Credit Card Fraud Detection

***Reference***

Lucas, Y., Portier, P., Laporte, L., He, L., Caelen, O., Granitzer, M. and Calabretto, S., 2019. *Towards Automated Feature Engineering For Credit Card Fraud Detection Using Multi-Perspective Hmms*. 1st ed. [ebook] Lyon: Research Gate, pp.4-6. Available at: <https://www.researchgate.net/publication/335600419\_Towards\_automated\_feature\_engineering\_for\_credit\_card\_fraud\_detection\_using\_multi-perspective\_HMMs> [Accessed 1 September 2020].

***Aim***

To determine..

***Methods***

Iterative ML ..

***Conclusions***

….

***Gaps in the Research:***

…

***Critique***

...

## I Mean Really Clever with Credit Card Fraud Detection

***Reference***

Mahmoudi, N. and Duman, E., 2015. Detecting credit card fraud by Modified Fisher Discriminant Analysis. *Expert Systems with Applications*, [online] 42(5), pp.2510-2516. Available at: <https://www.semanticscholar.org/paper/Detecting-credit-card-fraud-by-Modified-Fisher-Mahmoudi-Duman/1cfb2a0a0f11dab8da5dc38d51a6e816f04ac8e3#paper-header> [Accessed 11 September 2020].

***Aim***

To determine ..

***Methods***

Iterative ML workflow ..

***Conclusions***

….

***Gaps in the Research:***

…

***Critique***

...

# CA Guidance (remove before submission)

1. Use **paraphrased** sentences. Do not copy and paste directly from the articles. See a partial example below, with the paper from the first lecture:

**Topic: Examining cyclist behaviour in urban centres**

Reference: **Richardson, M. & Caulfield, B. (2015). Investigating traffic light violations by cyclists in Dublin City Centre. *Accident Analysis and Prevention, 84,* 65-73. doi:10.1016/j.aap.2015.08.011**

Aim: **To examine red light running of cyclists in Dublin City.**

Methods: **An observational survey and an online questionnaire.**

Conclusions: **61.9% of cyclists break the lights: 97.8% of cycle track users (pedestrian green phase); 18.6% of bicycle lane users (motorist phase). The most significant predictors were infrastructure type (cycle track) and cyclist gender (male).**

Gaps in the Research: **Need to account for difficulties people have in reporting incidents while cycling**

Critique: **The Dublin cycle network primarily consists of on-road cycle lanes. The surveys focussed on an even number of junctions with cycle lanes and cycle tracks. The overall rate of infringement is not fairly represented: in reality, it would be lower… (…)**

9 Please observe an overall **maximum of 1500 words** in the assignment.