Design and Development of a Prototype System for Real Time Fraud Detection In Credit Card Transactions Using Ensemble methods

Nyasha Chizampeni

Information Technology,

Harare Institute Of technology

Harare, Zimbabwe

[nchizampeni@gmail.com](mailto:nchizampeni@gmail.com)

Philip Sumburero

Information Technology,

Harare institute of technology

Harare, Zimbabwe

[psumburero@gmail.com](mailto:psumburero@gmail.com)

**Abstract - Fraud is the illegal act of impersonating an individual in order to discredit them or acquire benefits or privileges meant for the said individual. Fraud detection is a strong pillar of FinTech, used as a layer of security to ensure that clients transact in a safe, controlled and monitored environment. Since the innovation of e-business and online transactions, new attack vectors have opened up and people with malicious intent are discovering ways to manipulate these new systems and steal from unsuspecting individuals.**

**There have been significant developments in the research to fight fraud, from rule-based algorithms for dealing with stagnant and slow evolving economic climates to ( batch) machine learning and artificial intelligence for constantly evolving economies, wide array of solutions has been researched and evaluated for this problem.**

**Keywords: FinTech, RT (real time), ARF (Adaptive Random Forest), supervised, unsupervised, batch processing, PCA ( Principle Component Analysis) ,River.**

1. **INTRODUCTION**

Financial institutes and the businesses they offer services to, have the bear the cost and expense in the event of fraud. This is why it is in their best interest to develop flexible systems that have to the ability to the changes in fraudulent activity. In this research, and incremental learning approach was taken in order to improve the performance of the system as a whole.

**Real Time Fraud Detection**

Credit card transactions are the highest sort for means of payment because of their robustness. An introduction of a transaction analysis tool would critically require that the processing time of a transaction be not affected or increased by the tool in discussion.

Theoretical steps for the proposed solution:

* Collect credit card transaction data
* develop a pipeline that will automatically clean, transform and process the dataset
* Perform feature selection using PCA
* Stream one transaction at a time to the algorithm
* Review metrics to ensure that the model is improving

1. DESIGN

**Training and Testing Dataset**

The dataset used in this research contains over 30000 entries and has 15 features after the PCA. This data is used to initially test and train the model but the model will continue to learn as new is fed to the system.

The transaction dataset is in .csv format.

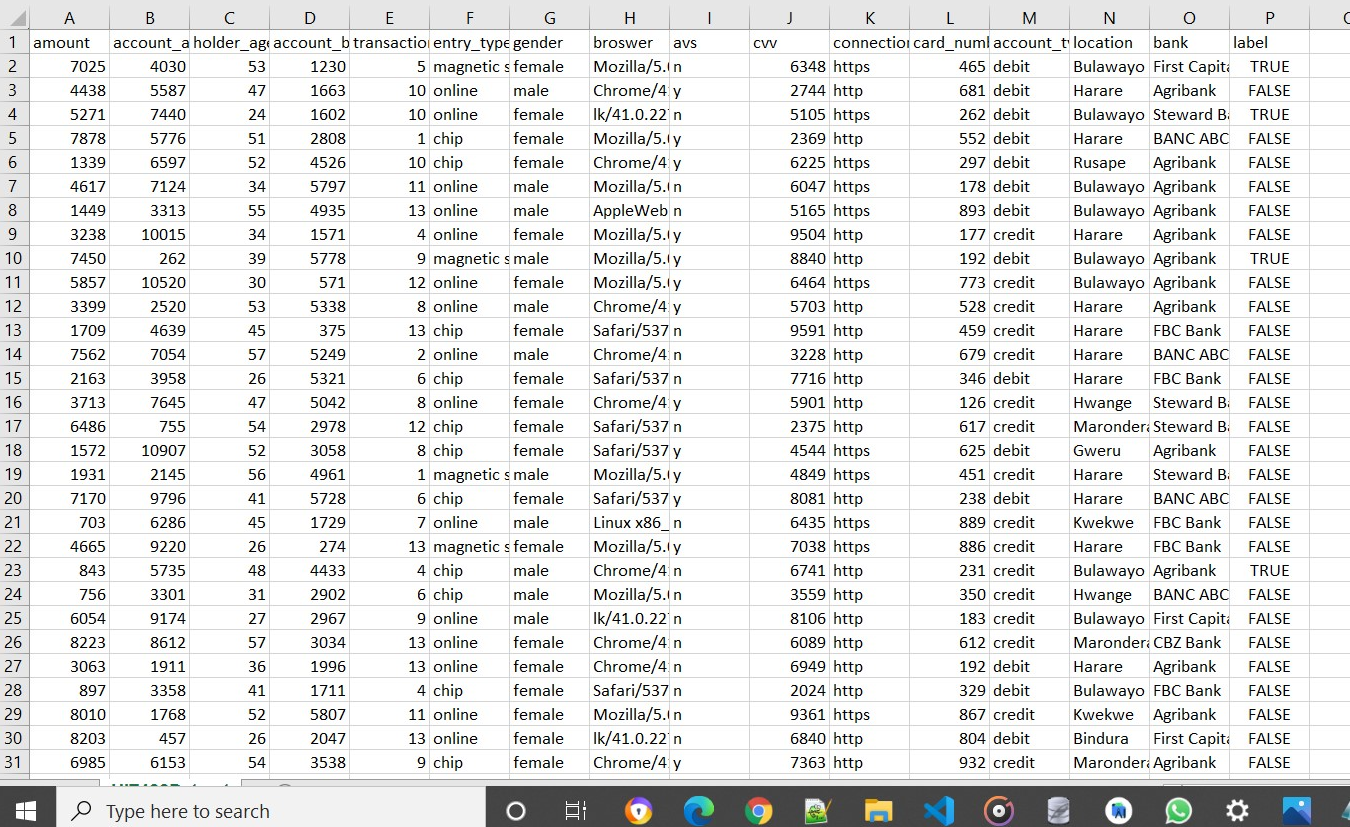
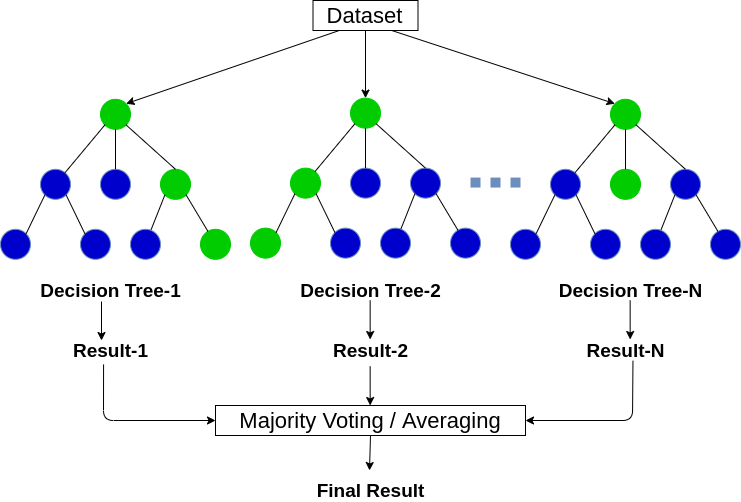
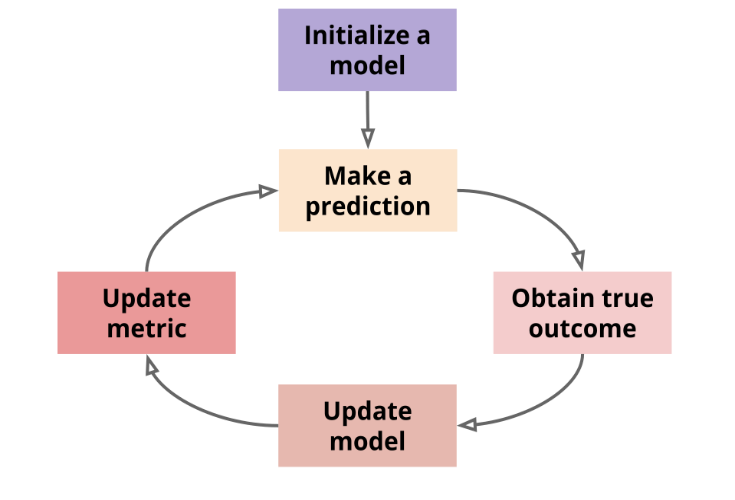


Figure 2: snapshot of credit card transaction data

**Classification of data**



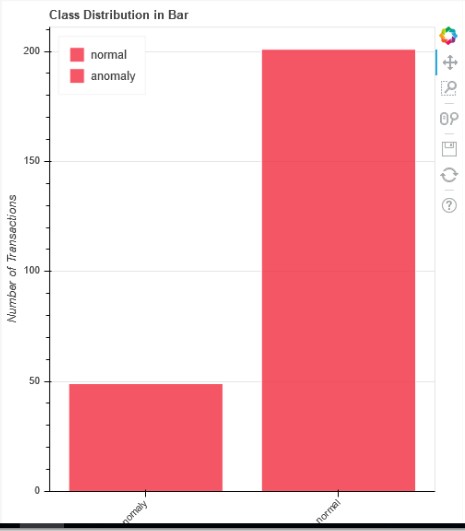
Random forest is a supervised learning algorithm. The "forest" it builds, is an ensemble of decision trees, usually trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result.

 Using the kappa architecture, the data is treated as a stream. The system does not have to store a historical training set and retrain then model every so often. Another benefit of is that the model is always up to date. The model can deal with concept drift, which happens when the data’s distribution evolves as time goes on. Moreover, the model lifecycle is easier to understand because the learning and prediction steps both process one input at a time. This is architecture was selected because of its similarity to the production environment, where transaction data is served to the API in the form of a data stream.

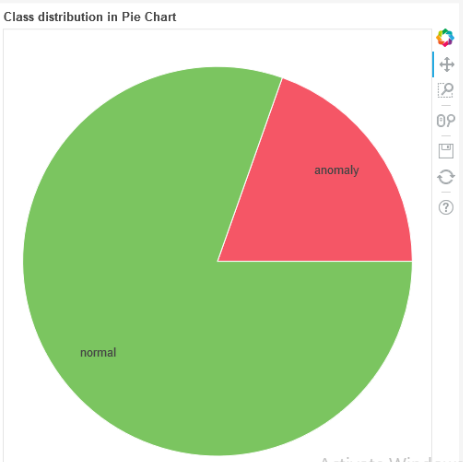
*Figure 3: classification*

**Visualisation of results**

The results of classifications are presented to the user through graphs. Graphs show the distributions of the classes (anomaly and normal transactions)



*Figure 4: Class distribution in Bar Chart*



*Figure 5: Class distribution in Pie Chart*

1. IMPLEMENTATION

**Platforms**

Apache Kafka and spark

Create multiple data streams from the sources of data, ready to be processed as Kafka topics. The Kafka topics are consumed by  
Apache Spark for data analysis and anomaly detection using SVM algorithm. it picks up anomalies such as overcharges and  
undercharges of say duty on certain goods, redirection of funds, unusual amounts for certain already known transactions

Python programing Language

Python is a general-purpose interpreted, interactive, object-oriented, and high-level programming language. This is the language used in the development of the proposed system.

**Frameworks**

Anaconda distribution

Anaconda distribution is an open source data science platform powered by python. It provide over 720 python packages that can be used in the development of any python program. For instance, sci-kit learn, a machine learning package that contains algorithms such as SVM, decision tree, etc.

Django

Django is a high-level python web framework that encourages rapid development and clean, pragmatic design. It follows the model-view-template architectural pattern. In as far as CA is concerned, Django web framework is used as backend of a web dashboard for the continuous monitoring of processing and processed jobs as well as real time visualization

**IDEs**

JetBrains PyCharm community edition

Intelligent Python IDE (instructions developing Editor) with refactoring, debugger, code completion, on-the-fly code analysis and coding productivity orientation. PyCharm integrates well with the above mentioned platforms as well as frameworks that are used to develop the proposed system.

1. DISCUSSION AND RESULTS

The main aim of the study was to detect anomalies in the transaction dataset using algorithms. Literature review showed that algorithms suit for this has applied, but combinations of algorithm are still under study. SVM classification algorithm prove to be the efficient with greater accuracy than most. The system reached a test accuracy of 86%. It shows that the dataset have many features to work with towards the ultimate grouping of transactions which implies that, it require huge amounts of training data in order to raise the accuracy score.

1. FUTURE WORK

The system can be extended to include the ability to continuously learn on its own in case of arrival of new transactional data that haven’t been in the training set.

Furthermore, the proposed prototype, have a 14% chance of misclassifications. However, in case of misclassifications, the classification model can use a predefined set of rules to verify and back up its outcome from prediction.

1. ACKNOWLEDGEMENTS

I wish to acknowledge the assistance received from the following people who made it possible for this document to be put together.

My supervisors, Mr B Masaiti who tirelessly guided me in conducting and compiling the entire project, and my fellow classmates for providing assistance in time of need

1. AUTHOR PROFILES



**Student Author:** Sternford Mavuchi is a fourth-year student at Harare Institute of Technology currently pursuing a BTech Hons degree in Information Technology