Project Proposal - Financial Fraud Detection Tool (Transaction)

MSCI 446 - Intro to Machine Learning

Group 23

Senan Gaffori - 20949022

Dev Shah - 20940056

William Girdwood - 20964572

John Greenough - 20939341

**Project Option -**

**2- Empirical evaluation: Pick an application area of interest (healthcare, energy, finance, etc.), collect data and propose ML solutions to the problem. Note that you can collect data by surveying or generating synthetic data.**

**Project Goal:** Financial Fraud Detection Tool (Transaction)

**Identify need:**

**Societal Importance:**

The societal importance of your project is multifaceted and significant. At its core is Consumer Protection, where the primary benefit is safeguarding individuals from unauthorized and fraudulent transactions. This is increasingly vital as our society leans more towards cashless transactions, and maintaining consumer confidence in this digital shift is essential. Additionally, your project plays a crucial role in Reducing Financial Crime. By detecting and preventing fraudulent transactions, it not only addresses the direct issue of financial fraud but also contributes to a broader societal objective of curtailing financial crimes, which are often linked to funding other illicit activities. Furthermore, there is an aspect of Enhanced Trust in Digital Transactions. As the prevalence of digital transactions grows, the security of these transactions becomes paramount. Ensuring their safety is a key factor in sustaining and nurturing public trust in our financial systems.

**Environmental Impact:**

In terms of environmental impact, the project contributes indirectly but notably through the Digitalization of Financial Services. Effective fraud detection mechanisms support the transition towards digital financial services. Compared to traditional, paper-based systems, digital services are more environmentally friendly. By reducing the reliance on paper and promoting digital transactions, your project aids in diminishing the overall environmental footprint of financial services.

**Importance for the Economy:**

The project's significance for the economy is profound. One of the primary economic impacts is Minimizing Financial Losses. Financial fraud in transactions represents a substantial cost to both businesses and consumers, running into billions annually. By reducing these losses, your project directly benefits the economy. Another critical aspect is ensuring Stability in Financial Markets. Secure transactions are fundamental to the stability of financial markets. Improved fraud detection mechanisms enhance this security, contributing to the overall health and stability of financial systems. Lastly, the project promotes Efficient Financial Operations. Resources that financial institutions currently spend on addressing and rectifying fraud issues can be more effectively utilized if fraud is detected and prevented at an early stage. This not only saves resources but also allows financial institutions to focus on more productive and innovative activities.

**Approach:**

The plan to approach the problem of Financial Fraud is to be able to seamlessly identify fraud in financial records and transactions using machine learning. This general approach is broken down into the processes of cleaning large-scale data frames, recognizing outliers, and developing mathematical approaches that allow the model to be scalable and applicable to different users, the machine learning model will then be trained on available public datasets. After this, we will test, and use reinforcement learning to improve our model. Then repeat the testing for as many iterations as necessary to make the tool effective.

We are planning on using readily available datasets that have been previously used for auditing processes, and fraud detection. We will analyze this data and make sure it is up to our standards. This standard will include: cleaning duplicates or irrelevant data, fixing all structural errors, filtering any unwanted outliers (which is minimal because we want to identify the outliers), handling any missing data or important features, and finally, we will validate the data and quality assurance. The potential ML approaches we are considering are; supervised learning, and reinforced learning. This will allow us to direct the ML more efficiently because we already have an idea of how we are trying to categorize data. We will use reinforced learning so that the model has a tighter fitting model without becoming overfitting.

**Comparison and Validation:**

To compare and validate our model, we will need to find comprehensive data sets with many different variables to properly train our model. Our data set must be large enough that when we split it into training sets and test sets, both sets are substantially large to train and test our model. To select the method for testing we must take into account the type of data that we collect. We would need to take into account factors that cause underfitting, and overfitting within different Machine Learning Model types. To validate our results, we would use the test to check the accuracy of the model.

**Potential risks:**

The undertaking of this project will come with various challenges, including technical risks, as well as risks relating to human factors such as resource and scheduling management. Of the technical risks that we face, one of the most significant is the data quality and the bias of that data. For this project we plan on using externally available, open source financial data sets. These datasets can pose several challenges relating to bias, including historical and demographic biases. Past this, the datasets may also be imbalanced, with few fraudulent transactions, leading to a less accurate detection model.

Another significant technical challenge to be faced in a technical context is the feature selection of our model. Many of the datasets to be used are incredibly large and contain several features, properly selecting these features to ensure an accurate and precise model. Outside of the technical risk and challenges we will face when undertaking this project, there will also be several other risks we face. These include scheduling our development to accommodate our schedules and other classes. Other challenges that we must overcome during the development of this project will include proper communication and project management implementation in order to fully build and deliver this project on time.

**Design thinking and project management**:

We have implemented a Gantt chart to organize our workflow throughout the project. The tasks were divided up into the following categories: collecting data, cleaning data, organizing data for training and testing, building our model, testing and reworking the model and finally presenting.

The first four phases of our project development are related to the data collection and cleaning of our model. It is important we dedicate ample amounts of time to these phases of development since a machine-learning model is only as good as the data it gets.

We have also dedicated large amounts of time to the development and testing of the model, this is in order to ensure that should any issues arise, we have adequate time to address them.

For a more detailed understanding of our workflow and scheduling please take a look at the attached Gantt chart below.

