

Final Report: Fraud Detection in Financial Transactions

Machine Learning | Date

Project Overview

The goal of this project was to develop a robust fraud detection system for financial transactions. The project encompassed data collection, preprocessing, statistical analysis, machine learning model development, and deployment using streamlit. Additionally, advanced techniques like Natural Language Processing (NLP) and Generative Adversarial Networks (GANs) were integrated to enhance the model's performance.

Week 1: Data Collection and Preprocessing

Tasks

* Data Collection: Gathered financial transaction data that included labeled instances of both fraudulent and non-fraudulent transactions.
* Data Preprocessing: Cleaned the dataset by handling missing values, normalizing features, and ensuring data consistency.

Tools

* Python Libraries: Pandas for data manipulation and NumPy for numerical operations.

Deliverables

* Cleaned Dataset: A preprocessed dataset ready for analysis.
* Data Preprocessing Notebook: Documented steps taken during data cleaning and preprocessing.

Week 2: Statistical Analysis and Machine Learning

Tasks

* Statistical Analysis: Analyzed the distribution of fraud-related features to identify key patterns and insights.
* Machine Learning: Developed and evaluated classification models such as Logistic Regression and Random Forest to detect fraud.

Tools

* Python Libraries: Scikit-learn for model development and Statsmodels for statistical analysis.

Deliverables

* Statistical Analysis Report: Insights derived from the data analysis.
* Model Performance Metrics: Evaluation of the developed models, including accuracy, precision, recall, and F1-score.

Week 3: Advanced Techniques and Streamlit

Tasks

* NLP for Transaction Notes: Applied NLP techniques to analyze transaction descriptions, enhancing feature extraction for the models.
* Streamlit: is a free and open-source framework to rapidly build and share beautiful machine learning and data science web apps. It is a Python-based library specifically designed for machine learning engineers. enables developers to build attractive user interfaces in no time

Tools

* Streamlit: open-source framework to rapidly build and share beautiful machine learning.
* Python Libraries: NLTK and SpaCy for natural language processing.

Deliverables

* Enhanced Model: A fraud detection model that incorporates NLP insights.
* Deployment Setup: streamlit Community Cloud lets you deploy your apps in just one click, and most apps will be deployed in only a few minutes.

Week 4: MLOps, GANs, and Final Presentation

Tasks

* MLOps: Implemented MLflow to manage and track model versions, experiments, and metrics for better model governance.
* GANs for Synthetic Data: Developed a Generative Adversarial Network (GAN) to create synthetic fraud transaction data, aiding in training and validating models.
* Final Report and Presentation: Compiled a comprehensive report detailing all project phases, findings, and methodologies.

Tools

* MLflow: For model management and tracking.
* Python Libraries: TensorFlow or PyTorch for building GANs, alongside Azure services for deployment.

Deliverables

* Deployed Model: A fraud detection system capable of identifying fraudulent transactions.
* Synthetic Data Generation: Utilized GANs to enrich training datasets with synthetic examples.

Conclusion

The project successfully developed and deployed a fraud detection system that leverages advanced data analysis, machine learning techniques, and cloud integration. The use of NLP and synthetic data generation significantly improved model accuracy and robustness. Future work may involve continuous model updates, further enhancement of NLP capabilities, and exploration of additional machine learning algorithms to refine fraud detection accuracy.

|  |  |  |
| --- | --- | --- |
| Team Member | Role | week |
| Ahmed Ali Yahya | Data Collection & Preprocessing | Week 1 |
| Elsayed Osama Elsayed | Statistical Analysis & ML Development | Week 2 |
| Ayman Mohamed Abdelhady | Streamlit & NLP | Week 3 |
| Mhmoud Mohamed | Synthetic Data with GANs | Week 4 |
| Omnia Ashraf eldeeb | MLOps & final report and presentation | Week 4 |

### Project Team Members