Project Title: Anomaly Detection in Corporate Credit Card Transactions using Machine Learning

1. Introduction

The rapid growth of the commerce industry has led to an alarming increase in digital fraud and associated losses [1]. Detecting anomalous transaction patterns, such as duplicate payments, fraudulent payments, and transactions that do not follow the company's Delegation of Authority (DOA), is crucial for maintaining a healthy financials of an organization [2].

2. Abstract

This project aims to develop a machine learning model to detect anomalous transaction patterns, such as duplicate payments, fraudulent payments, and transactions that do not follow the company's Delegation of Authority (DOA). The model will be trained on an anonymized or synthetic dataset to protect sensitive information. An automated documentation process will be created to meet SOX reporting standards. The model will be coded in Python, stored on GitHub, and deployed using Amazon SageMaker.

3. Background and Motivation

The IT Audit team at the publicly-traded company is looking to invest in automated data analytics approaches to streamline their work and focus on higher-value tasks. By developing a machine learning model to detect anomalous transaction patterns in corporate credit card data, the team can reduce the time and effort spent on manual review and investigation of potential issues [3], [4].

4. Objectives

1. Develop a machine learning model that accurately detects anomalous transaction patterns
2. Create an automated documentation process that meets SOX reporting standards
3. Provide insights into the most significant variables and relationships in the data
4. Deploy the model on Amazon SageMaker for integration into the IT Audit team's workflow

5. System Architecture

The proposed system architecture comprises three main components: data preprocessing, model development, and model deployment. The data preprocessing component will handle data cleaning, anonymization, and feature engineering. The model development component will involve the implementation and evaluation of various machine learning algorithms. The model deployment component will focus on integrating the trained model into the IT Audit team's workflow using Amazon SageMaker.

6. Methodology

1. Data Preparation [5], [6]
    - Access the corporate credit card transaction dataset
    - Anonymize sensitive records or substitute with synthetic data using Python libraries (Faker, DataSynthesizer)

   - Preprocess data using Python libraries (Pandas, NumPy)
   - Apply feature engineering techniques, such as feature scaling and encoding categorical variables
2. Exploratory Data Analysis [7]
   - Visualize data using Python libraries (Matplotlib, Seaborn)
   - Compute descriptive statistics and create informative plots
3. Model Development [8], [9], [10]
   - Implement supervised learning algorithms using Python libraries (Scikit-learn, TensorFlow)
   - Apply unsupervised learning techniques using Scikit-learn
   - Optimize model performance using feature selection and cross-validation
   - Address potential challenges, such as class imbalance and high-dimensional data
4. Model Evaluation [11]
   - Evaluate model performance using metrics from Scikit-learn (accuracy, precision, recall, F1-score, ROC curves)
   - Assess model generalization ability and balance between bias and variance
5. Documentation Automation
   - Develop an automated approach using Python's docstring and libraries (Sphinx, pdoc3)
   - Ensure documentation meets SOX reporting standards by incorporating specific requirements and guidelines
6. Deployment [12]
   - Code the model in Python and store it on GitHub
   - Deploy the model using Amazon SageMaker

7. Timeline

The project will be completed by December 8, 2024, to align with course requirements.

8. Expected Results

- A machine learning model that effectively detects anomalous transaction patterns
- An automated documentation process that meets SOX reporting standards
- Insights into the most significant variables and relationships in the data
- A deployed model on Amazon SageMaker, ready for integration into the IT Audit team's workflow

9. Limitations and Future Work

- Ensuring the anonymized or synthetic data adequately represents the original dataset
- Adapting the automated documentation process to meet SOX reporting standards
- Exploring advanced feature engineering techniques and deep learning algorithms for improved performance [13], [14]

10. Conclusion

This project proposal presents a comprehensive approach to detecting anomalous transaction patterns in corporate credit card data using machine learning techniques. By leveraging state-of-the-art algorithms, automated documentation processes, and cloud-based deployment, the proposed solution aims to enhance the efficiency and effectiveness of the IT Audit team's fraud detection efforts. The successful implementation of this project will contribute to the overall financial integrity and compliance of the company.

References

[1] A. Herreros-Martínez, R. Magdalena-Benedicto, J. Vila-Francés, A. J. Serrano-López, and S. Pérez-Díaz, "Applied machine learning to anomaly detection in enterprise purchase processes," arXiv:2405.14754 [cs], May 2023, doi: 10.48550/arXiv.2405.14754.

[2] A. R. Khalid, N. Owoh, O. Uthmani, M. Ashawa, J. Osamor, and J. Adejoh, "Enhancing credit card fraud detection: An ensemble machine learning approach," Big Data Cogn. Comput., vol. 8, no. 1, p. 6, Jan. 2024, doi: 10.3390/bdcc8010006.

[3] A. Ali et al., "Financial fraud detection based on machine learning: A systematic literature review," Appl. Sci., vol. 12, no. 19, p. 9637, Jan. 2022, doi: 10.3390/app12199637.

[4] A. Mutemi and F. Bacao, "E-commerce fraud detection based on machine learning techniques: Systematic literature review," Big Data Min. Anal., vol. 7, no. 2, pp. 419–444, Jun. 2023, doi: 10.26599/BDMA.2023.9020023.

[5] A. Teymouri, M. Komeili, E. Velazquez, O. Baysal, and M. Genkin, "DATA5000OMBA – Accessing Data," Carleton University BrightSpace, 2022. [Online]. Available: [URL].

[6] A. Teymouri, M. Komeili, E. Velazquez, O. Baysal, and M. Genkin, "DATA5000OMBA – Pre-Processing Data," Carleton University BrightSpace, 2022. [Online]. Available: [URL].

[7] A. Teymouri, M. Komeili, E. Velazquez, O. Baysal, and M. Genkin, "DATA5000OMBA – Visualizing Data," Carleton University BrightSpace, 2022. [Online]. Available: [URL].

[8] A. Teymouri, M. Komeili, E. Velazquez, O. Baysal, and M. Genkin, "DATA5000OMBA – Machine Learning Introduction," Carleton University BrightSpace, 2022. [Online]. Available: [URL].

[9] A. Teymouri, M. Komeili, E. Velazquez, O. Baysal, and M. Genkin, "DATA5000OMBA – Machine Learning Algorithms - 1," Carleton University BrightSpace, 2022. [Online]. Available: [URL].

[10] A. Teymouri, M. Komeili, E. Velazquez, O. Baysal, and M. Genkin, "DATA5000OMBA – Machine Learning Algorithms - 2," Carleton University BrightSpace, 2022. [Online]. Available: [URL].

[11] A. Teymouri, M. Komeili, E. Velazquez, O. Baysal, and M. Genkin, "DATA5000OMBA – Evaluating Machine Learning Models," Carleton University BrightSpace, 2022. [Online]. Available: [URL].

[12] A. Teymouri, M. Komeili, E. Velazquez, O. Baysal, and M. Genkin, "DATA5000OMBA – Analytics Accelerators," Carleton University BrightSpace, 2022. [Online]. Available: [URL].

[13] A. Y. Shdefat et al., "Comparative analysis of machine learning models in online payment fraud prediction," in 2024 Intelligent Methods, Systems, and Applications (IMSA), 2024, pp. 243–250. doi: 10.1109/IMSA61967.2024.10652861.

[14] A. Teymouri, M. Komeili, E. Velazquez, O. Baysal, and M. Genkin, "DATA5000OMBA – Role of the Data Scientist," Carleton University BrightSpace, 2022. [Online]. Available: [URL].