Cairo University
Faculty of Engineering

Department of Computer
Engineering

Machine Learning
Project Proposal

Team 17

Name	ID	Sec	BN
Abdelhamed Emad	9202758	1	34
Abdelrahman Hamdy	9202833	1	38
Ziad Sherif	9202586	1	26
Zeyad Tarek	9202588	1	27

Date: 22/03/2024

Presented to: Eng. Mohammed Shawki

Ideas:

From Highest to Lowest Priority

1ST: Credit card fraud detection

(We really want you to choose this idea)

1. Problem Description:

Problem: Credit card fraud detection is a critical task for financial institutions to protect their customers from unauthorized transactions. The challenge lies in identifying fraudulent transactions among a large volume of legitimate ones while minimizing false positives.

Problem Definition and Motivation: The goal is to develop a **machine learning model** capable of accurately **detecting fraudulent credit card transactions** based on transaction data. This is motivated by the need to prevent financial losses for both customers and credit card companies, as well as to maintain trust in the financial system. Credit card fraud detection is crucial for protecting customers and minimizing financial losses. Machine learning models, including supervised algorithms like logistic regression and random forests, as well as unsupervised techniques like anomaly detection, are employed to identify fraudulent transactions. Evaluation metrics such as precision, recall, and AUC-ROC ensure the effectiveness of these models. Once trained, the selected model can be deployed to monitor real-time transactions, aiding financial institutions in promptly detecting and mitigating fraudulent activity.

2. Evaluation Metrics:

Area Under the Precision-Recall Curve (AUPRC): Given the highly unbalanced nature of the dataset, AUPRC is a more appropriate metric than accuracy. It provides a comprehensive measure of the model's ability to detect fraud while controlling for false positives.

3. Dataset and References:

- **Dataset:** The dataset contains transactions made by credit cards in September 2013 by European cardholders. It includes 492 frauds out of 284,807 transactions, with highly unbalanced classes.
- **References:**

- The dataset is sourced from Kaggle: Credit Card Fraud Detection
- Additionally, a simulator for transaction data has been released as part of the practical handbook on Machine Learning for Credit Card Fraud Detection (Fraud Detection Handbook).

2ND: Movie Recommendation System

1. Problem Description:

Problem: Building a movie recommendation system that suggests movies to users based on their preferences and viewing history. The system aims to enhance user experience by providing personalized movie recommendations.

Problem Definition and Motivation: With the vast amount of content available on streaming platforms, users often face decision paralysis when selecting movies to watch. A recommendation system can help alleviate this issue by offering personalized suggestions tailored to each user's tastes and preferences. This project aims to develop such a recommendation system to improve user engagement and satisfaction.

2. Evaluation Metrics:

- Accuracy: Measure how accurately the recommendation system predicts users' preferences for movies.
- Precision and Recall: Metrics to evaluate the relevance of recommended movies to users' interests.
- Mean Average Precision (MAP): Average of precision at different cutoff levels, providing a comprehensive measure of recommendation quality.

3. Dataset and References:

- **Dataset:** The dataset contains movie ratings provided by users on a scale of 1 to 5. It includes information about users, movies, and ratings.
- **References:** The dataset is sourced from Kaggle: https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset

3RD: Intrusion detection

1. Selected problem:

Problem: Intrusion detection is a critical aspect of cybersecurity, aiming to identify unauthorized access, misuse, or abnormal activities in a computer network. The challenge is to develop an effective intrusion detection system (IDS) capable of accurately distinguishing between normal and malicious network traffic.

Problem Definition and Motivation: The goal is to build a machine learning-based IDS to enhance network security by detecting and preventing cyber-attacks. This is motivated by the increasing sophistication of cyber threats and the need to safeguard sensitive data and systems from unauthorized access and exploitation.

2. Evaluation Metrics:

- **Accuracy:** Measures the overall correctness of the model's predictions.
- **Precision:** Calculates the proportion of true positives among all positive predictions, indicating the model's ability to avoid false alarms.
- **Recall:** Measures the proportion of true positives that were correctly identified by the model, indicating its ability to detect actual intrusions.
- **F1 Score:** Balances precision and recall, providing a single metric to evaluate the model's performance.

3. Dataset and References:

- **Dataset:** The dataset contains network traffic data, including features extracted from network packets, to classify normal and malicious activities. It is sourced from Kaggle: https://www.kaggle.com/datasets/hassan06/nslkdd.
- **References:** The proposed project is inspired by research articles and methodologies in the field of cybersecurity and intrusion detection systems.