**Project Team Members**

|  |  |  |
| --- | --- | --- |
| Team Member | Role | **Week** |
| Ahmed Ibrahim Abd El Hafeez Muhammed | Data Collection & Preprocessing for ML | Week 1 |
| Mostafa Sobhy Mahmoud Shehab | Statistical Analysis & ML Development | Week 2 |
| Ahmed Hossam Eldien Ahmed Hussain | NLP | Week 3 |
| Nusiba Rabea Abdallah Khalf | Azure AI Fundamentals | Week 3 |
| Mostafa Khaled Hossam Eldien | GANs for Synthetic Data | Week 4 |

**Project Overview**

* This project focuses on developing predictive maintenance models for industrial equipment.
* The goal is to predict potential equipment failures, thereby optimizing maintenance schedules, reducing downtime, and lowering operational costs.
* The dataset comprises historical maintenance records and various equipment-related attributes, which were preprocessed and analyzed to build robust predictive models.

**Week 1: Data Collection and Preprocessing**

**Tasks**

* Data Collection: Obtain financial transaction data, including labeled fraudulent and non-fraudulent transactions.
* Data Preprocessing: Clean and preprocess the data, addressing missing values and normalizing features.

**Tools**

* Python (Pandas, NumPy).

**Deliverables**

* Cleaned and preprocessed dataset.
* Data preprocessing notebook.

**-- By Mostafa Sobhy ---**

**Week 2: Statistical Analysis and Machine Learning**

**Tasks:**

**Statistical Analysis**: Perform statistical analysis to understand the distribution of fraud-related features.

**Machine Learning**:

* Develop and evaluate classification models for fraud detection (e.g., Logistic Regression, Random Forest).

**Tools:**

* Python (Scikit-learn, Statsmodels).

**1. Feature Engineering**

**1.1 Feature Engineering**

* The dataset was inspected to understand variable types and identify issues like missing values.
* Missing values were addressed by imputing numerical columns with their mean and categorical columns with their mode.

**1.2 Scaling**

* Continuous variables were scaled.

**2. Exploratory Data Analysis (EDA)**

**2.1 Data Distribution and Correlation**

* Data Distribution: Visualizations revealed key patterns, helping to identify skewed features and outliers that could affect the model's accuracy.
* Correlation Analysis: A heatmap displayed correlations among features, revealing strong relationships that could influence the target variable. Highly correlated variables were either combined or reduced to prevent multicollinearity and boost model reliability.

**2.2 Key Insights**

* Key failure indicators include equipment age usage frequency, and operational environment.
* Higher usage rates, in particular, showed a strong correlation with equipment **failure, suggesting they should be prioritized in modeling efforts.**

**3. Modeling**

**3.1 Model Selection**

* Multiple models were evaluated to identify the best-performing one:
* Logistic Regression: Used as a baseline for interpretability.
* Decision Tree: Selected for its ability to capture non-linear relationships.
* Random Forest: Chosen to reduce overfitting through ensemble methods, improving stability.
* Gradient Boosting: Tested for its high accuracy in complex datasets.

**3.2 Hyperparameter Tuning**

* Grid search and cross-validation optimized parameters such as learning rate, tree depth, and the number of trees.
* This improved the model’s accuracy and generalizability across different data splits.

**3.3 Evaluation Metrics**

**Models were evaluated using the following metrics:**

* Accuracy: Measures overall predictive performance.
* Precision and Recall: Help assess relevance and completeness of positive predictions.
* F1-Score: Balances precision and recall, useful for cases with class imbalance.
* ROC-AUC Score: Measures the model’s ability to distinguish between failure and non-failure cases.

**3.4 Final Model Selection and Accuracy**

* The Random Forest model achieved the highest accuracy (approximately X%), with strong precision, recall, and ROC-AUC scores. Cross-validation results confirmed that this model generalizes well across data splits, making it the optimal choice.

**4. Model Evaluation and Results**

**4.1 Performance Summary**

* The final Random Forest model achieved an accuracy of X%, precision of Y%, recall of Z%, and a high ROC-AUC score, effectively predicting equipment failures with minimal false positives.

**5. Conclusions and Recommendations**

**5.1 Key Findings**

* High equipment usage, operational environment, and lack of regular maintenance were strong predictors of failure.
* The correlation between usage and failure underscores the importance of monitoring this factor closely.

**Week 3: Advanced Techniques and Azure Integration**

**Tasks**

* NLP for Transaction Notes: Apply NLP techniques to analyze transaction descriptions or notes
* Azure AI Fundamentals: Deploy the fraud detection model using Azure Machine Learning or Azure Synapse.

**Tools**

* Azure Machine Learning, Python (NLTK, SpaCy)

**Deliverables**

* Enhanced fraud detection model with NLP integration.
* Deployment setup on Azure.

**Week 4: MLOps, GANs, and Final** **Presentation**

* GANs for Synthetic Data: Developed a Generative Adversarial Network (GAN) to create synthetic fraud transaction data, aiding in training and validating models.

**Tools**

* Python Libraries: PyTorch for building GANs, alongside Azure services for deployment.

**Deliverables**

* Synthetic Data Generation: Utilized GANs to enrich training datasets with synthetic examples.