Vendor Risk Prediction - Documentation

Project Overview

Goal:

To analyze invoice data and assess vendor risk by predicting whether a vendor's invoice will be **Paid** or **Pending**. Initially, time series forecasting was attempted for monthly spending trends, but the data exhibited white noise, making predictions unreliable. Consequently, the focus shifted to developing a classification model for vendor risk analysis.

Dataset:

- **Invoices Dataset:** Contains details of invoices issued by vendors, including amount, date, and payment status.
- **Payments Dataset:** Records payments made against invoices, including payment dates and methods.
- **Vendors Dataset:** Provides information about vendors, such as name, country, and business category.

Tech Stack:

- **Programming Language:** Python, jupyter notebook
- **Libraries Used:** Pandas, NumPy, Scikit-learn, XGBoost, FastAPI, Matplotlib, Seaborn, Joblib
- **Deployment:** FastAPI,

Exploratory Data Analysis (EDA)

1️⃣ Data Exploration

- Combined all datasets and checked for missing values and data inconsistencies.
- Conducted **pandas profiling** to generate an automated **analytics report (HTML file)** for comprehensive insights.
- Examined statistical distributions and correlations between key features.

2️⃣ Anomaly Detection (Outlier Analysis)

- Applied **visualization techniques** (box plots, histograms) to identify potential outliers.
- Used **Z-score method** and **Interquartile Range (IQR) method** to detect statistical anomalies.

- **Conclusion:** The data did not contain significant outliers, so no further anomaly handling was required.

3️⃣ Time Series Analysis (Discarded Due to White Noise)

- Performed time series analysis on **monthly spending trends** using ARIMA and Prophet.
- Found that the data exhibited **white noise**, meaning there was no predictable pattern for forecasting.
- Concluded that time series modeling would not be effective for this dataset.

Vendor Risk Analysis Model

1️⃣ Feature Engineering

- Extracted meaningful features to enhance prediction accuracy:
  - **Vendor Frequency:** Number of invoices per vendor.
  - **Average Invoice Amount:** Mean value of invoices issued by each vendor.
  - **Total Invoice Amount:** Sum of all invoices per vendor.
  - **Pending Ratio:** Proportion of a vendor's invoices that remain pending.
  - **Invoice Age:** Days since the invoice was issued.
- Encoded categorical variables (e.g., currency) for machine learning compatibility.

2️⃣ Model Training & Comparison

- Implemented multiple classification models:
  - **Logistic Regression** (Baseline Model)
  - **Decision Tree**
  - **Random Forest**
  - **Gradient Boosting (Best Performance)**
  - **XGBoost**
  - **Support Vector Machine (SVM)**
  - **Naïve Bayes**
- Evaluated models based on **accuracy, precision, recall, and F1-score**.
- Gradient Boosting achieved the highest accuracy and was selected as the final model.

Model Deployment

1️⃣ Saving the Best Model

- Used **Joblib** to save the trained Gradient Boosting model as `vendor_risk_model.pkl` for future use.

2️⃣ FastAPI Development

- Created a **FastAPI** application to serve the model predictions.
- Implemented an API endpoint to accept invoice data and return a **risk prediction (Paid or Pending)**.