A logo of a globe with a graduation cap

Description automatically generatedA black background with white text

Description automatically generated

DIGITAL EGYPT PIONEERS INITIATIVE

(DEPI)

Round (1) 2024

* Technical Company: CLS
* Technical Track: IBM Data Science
* Group code: CLS CAI1\_AIS3\_S1e
* Technical Instructor: ENG/ Mohamed Soliman

Team Members:

* **Emad Mostafa Srag Eldein**
* **Omar Hussein Mohamed Sahwan**
* **Ebrahem Osama Hassan Mohamed Omran**
* **Abdelrahman Mohamed Youssef Zahran**

**Project Idea: Personalized Recommendation System**

A personalized recommendation system project focuses on developing an algorithm to predict and suggest items that users are likely to engage with based on their past behaviors, preferences, or demographics. The system analyzes user interactions, such as clicks, purchases, ratings, or browsing history, to generate personalized recommendations. These recommendations can span various domains like e-commerce (product recommendations), entertainment (movies or music), or content platforms (articles or videos).

**Project Idea: Store Sales - Time Series Forecasting**

Store Sales - Time Series Forecasting is a data science challenge focused on predicting future retail sales using historical data and external factors. This involves modeling sales trends over time to generate accurate forecasts based on past performance and relevant influences.

**The project involves:**

* **Problem Definition:** Predict future sales for retail stores using historical and external data.
* **Data Collection**:
* **Train Dataset:** Historical sales with store numbers, product families, and promotions.
* **Test Dataset:** Similar features without sales for evaluation.
* **Additional Datasets**: Information on holidays, oil prices, store details, and transactions.
* **Data Preprocessing:** Clean and prepare data by handling missing values and encoding categorical variables.
* **Feature Engineering**: Create new features, including calendar-based variables.
* **Exploratory Data Analysis (EDA):** Analyze data through visualizations and summary statistics.
* **Data Visualization**: Illustrate sales trends and relationships.
* **Modeling:**
* **XGBoost Model:** Manages complex interactions.
* **K-Means Model:** Cluster customers based on sales and promotions
* **Arima Model:** used for analyzing and forecasting time series data that exhibits trends, cycles, or seasonality.
* **Model Evaluation:** Use metrics like RMSE, MAE, and R² for performance assessment.
* **Results Interpretation:** Compare models and provide strategic insights.
* **The project's goal is to develop accurate sales forecasting models for Corporación Favorita in Ecuador using advanced techniques, with a focus on predictive modeling, data analysis, performance evaluation, providing strategic insights, and continuous improvement**

**1. Overview**

This project focuses on predicting outcomes using a machine learning approach. It uses multiple datasets, including sales data, holiday events, oil prices, and store information, to develop predictive models. The goal is to build models that accurately forecast sales based on historical data and external factors.

**Datasets:**

* \*Train.csv\*: Contains historical sales data.
* \*Test.csv\*: Contains sales data for testing.
* \*Holidays\_events.csv\*: Information on holidays.
* \*Oil.csv\*: Oil price data.
* \*Stores.csv\*: Store-related data.
* \*Transactions.csv\*: Store transactions data.

**2. Data Preprocessing**

* Missing data was handled using imputation techniques where necessary.
* Label encoding and feature scaling were applied to transform the data for model training.
* Datasets were split into training and testing sets.

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

During the EDA phase, the data was explored to understand its distribution, missing values, and key statistics. Below are the key steps performed:

* Displayed top records and key statistics (mean, std, min, max).
* Checked for missing values and data shapes.
* Visualized the distributions of key features using seaborn and matplotlib.

**5. Model Evaluation**

The models were evaluated using the following metrics:

* \*Accuracy\*: Measures the correctness of classification.
* \*R2-Score\*: Determines the fit of regression models.
* \*Mean Absolute Error (MAE)\* and \*Mean Squared Error (MSE)\*: To measure prediction error.
* \*Confusion matrices and classification reports were used to assess performance.
* The silhouette score measures how similar an object is to its own cluster compared to other clusters, with values ranging from -1 (poor clustering) to 1 (well-clustered).

**6. Results**

The Random Forest model performed the best, with an R2-score of X on the test data, and showed the lowest mean squared error (MSE). The results suggest that ensemble methods like Random Forest and Gradient Boosting work well for this dataset.

**7. Conclusion**

This project demonstrates the importance of data preprocessing and model selection. The findings indicate that certain external factors, such as oil prices and holiday events, significantly impact sales. Future work could focus on feature engineering and model optimization to further improve accuracy.