   
  
**Problem Statement: Contract Amount Prediction**

Develop a machine learning model that accurately predicts the revised contract amounts for contracts, incorporating both the existing dataset and additional external data factors, such as external economic data, that may influence contract values. The objective is to create a robust predictive model that can forecast the revised contract amounts, considering relevant features from the dataset, and incorporating external economic data and other factors that affect contract values.

**Background and Context:**

Accurate prediction of revised contract amounts is crucial for effective administration, budgeting, resource allocation, and decision-making processes. By leveraging historical contract data and external economic indicators, a predictive model can provide valuable insights into potential changes in contract values. External economic data refers to factors such as GDP growth rates, inflation rates, exchange rates, or any other relevant macroeconomic indicators.

**Dataset:**

The dataset includes the following fields: OBJECTID, contract\_number, contract\_description, contract\_status, contract\_type, contract\_subtype, last\_modified, original\_amount, revised\_amount, vendor\_id, vendor\_name, and objectid\_1.

**Link:** [Contract\_dataset.csv](https://dprprod-my.sharepoint.com/:x:/g/personal/ishanp_vconstruct_in/EfRWJBRJy1pEn2DeL5LVxP0BCEJ2wwSINpCC9BMLdfxO7A?e=9udQA4)

**Milestones:**

* 1. **Feature engineering:** search for additional economic data from the World Bank API to complement the given dataset with new relevant features which may help model performance. More detail is provided below (see Tech Stack).
  2. **Exploratory Data Analysis:** perform basic EDA to plot feature distributions and understand the underlying dataset statistics.
  3. **Data augmentation:** increase your dataset size by 30% by implementing statistical data generation techniques. You may use any existing package to fit a multi-dimensional gaussian (or similar) distribution to the dataset using all the provided features. Then, sample new data points using the Monte-Carlo approach. Make sure you discard outliers when they are generated. Plot the dataset again.
  4. **Data preprocessing and cleaning:**prepare data for model ingestion by ensuring the dataset is consistent, there are no missing entries, no outliers, and it is scaled and normalized appropriately.
  5. **Feature selection and model training:** *s*plit your dataset into train, validation and test sets. Create an ML pipeline (using any package of your preference) to train a baseline supervised learning model with a validation approach of your choosing. Employ dimensionality reduction and feature selection to identify the most critical features. Once you have created a training and evaluation pipeline, progressively increase model complexity to improve performance.
  6. **Model evaluation and deployment:** define an appropriate set of metrics to evaluate the final model performance on the test set and plot them. Does the model perform better with or without the added synthetic data?

**Evaluation Criteria:**

The solution will be evaluated based on the following criteria:

* **Code Quality:** Well-structured, documented, and adhering to best practices.
* **Approach Used:** Thoughtful and effective strategy for predicting revised contract amounts.
* **Relevant Tech Used:** Utilizing appropriate machine learning libraries and incorporating external economic data.
* **Model Performance:** Accurate and precise forecasting of revised contract amounts.
* **Performance Optimization:** Efficient processing of data and external factors.
* **Deployment and User Testing:** Deployed solution tested for usability and effectiveness.

**Tech Stack:**

The candidate is encouraged to use appropriate machine learning libraries such as scikit-learn, TensorFlow, or Py Torch for model development. They can also use data manipulation and analysis tools like pandas and NumPy.

The candidate is encouraged to source external economic data from the World Bank API, which provides access to a wide range of economic indicators from various countries, such as GDP growth rates, exchange rates or inflation rates.

World Bank API Doc: <https://documents.worldbank.org/en/publication/documents-reports/api>

Incorporate relevant indicators specific to the United States and integrate World Bank API data with the provided dataset based on the *"last\_modified"* field, which represents the date of modification.

**Constraints and Considerations:**

The model should handle missing data and outliers appropriately during preprocessing. The candidate should also consider scalability and computational efficiency while designing the model. Ethical considerations, such as data privacy and bias, should be addressed throughout the development process.

**Deliverables:**

The deliverables for the hackathon include the full code implementation in Python and preferably in Jupyter notebook format, a saved model file, and a sample output displaying the model's predictions.