Project Design Phase-I
Solution Architecture

Date	October 2023
Team ID	Team-592444
Project Name	Anticipating Business Bankruptcy
Maximum Marks	4 Marks

CSV Data:

The raw CSV data source serves as the architecture's starting point. This data comprises the necessary information for the machine learning challenge. Any organized data, such as a dataset with numerous attributes and goal variables, could be used.

Data Preprocessing:

Data preprocessing is an important stage in machine learning that ensures the quality and consistency of the data. This stage consists of the following components:

- Data cleaning: Deals with missing values, outliers, and discrepancies in data.
- Data transformation: Includes feature engineering, categorical variable encoding, scaling, and normalization.
- Feature selection: Selecting relevant features to improve model performance.

Dividing Data into Train and Test Sets:

The dataset is separated into two subgroups after data preprocessing:

Training Data: This subset of data is utilized to train the machine learning model. It usually contains a bigger amount of data, approximately 70-80% of the dataset.

Test Data: This subset is used to assess the model's performance. It typically contains the last 20-30% of the dataset.

Algorithm Selection:

Choose the machine learning algorithm or model that is most suited for the task at hand. The algorithm chosen relies on the nature of the task (classification, regression, etc.) and the data characteristics.

Upon testing with various algorithms including decision trees and random forests, the chosen algorithm for our solution is **XGBoost** due to high accuracy scores.

Training the Model:

Using the training data, the chosen algorithm is trained. This entails feeding the algorithm the features and their matching target variables for the algorithm to understand the underlying patterns in the data.

Model Evaluation:

The model's performance is tested using test data after training. Metrics for evaluating problems vary depending on the kind, but popular metrics include accuracy, precision, recall, F1-score for classification, and mean squared error, R-squared for regression.

Decision Point and Final Model Building:

At this point, a decision is taken depending on the outcomes of the evaluation. A conditional branch is included in the architecture to test the model's performance:

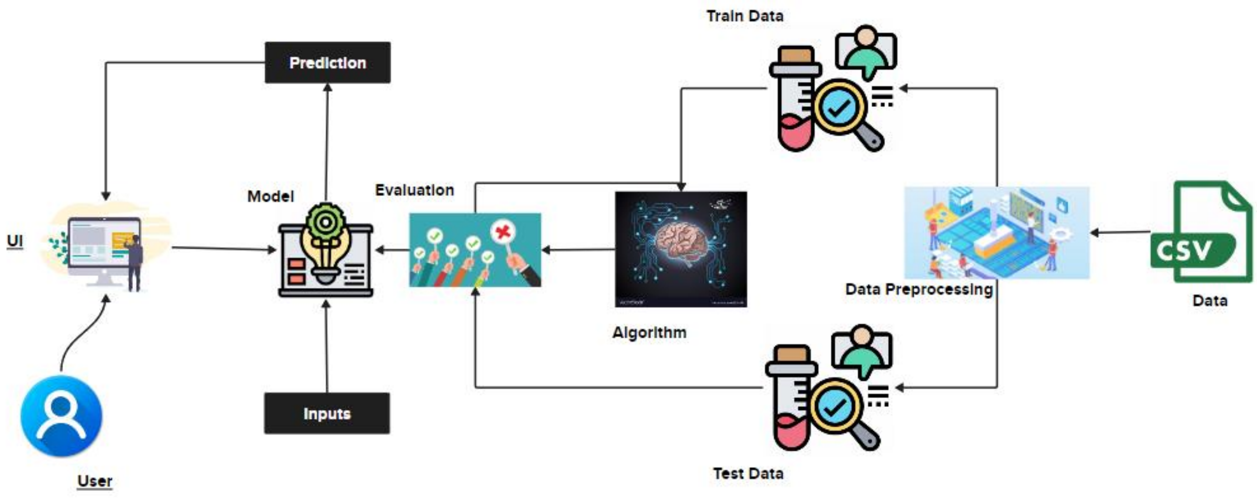
1. If the model's performance is unsatisfactory, it returns to the "Algorithm Selection" step, where you can choose a different algorithm, fine-tune hyperparameters, or re-run data preparation.
2. If the model passes established requirements for example an accuracy > 0.85 for a classification test, the system builds the final model.

If yes: The architecture, then uses the full dataset (train + test) to generate the final model. This aids in making the most of the given data for model training.

Deployment and Continuous Monitoring:

If the model performs well in the final evaluation, it can be used in the real world. As new data becomes available, continuous monitoring and changes may be required to maintain model performance.

Solution Architecture Diagram:


Train Data
Prediction
Model
Evaluation
UI
Data Preprocessing
CSV
Data
Algorithm
User
Inputs
Test Data