About me -

Pratyush Patra
+91 9612276838

pratyushpatra13@gmail.com

Task 5:

AutoML and Hyperparameter Optimization: Build a tool that automates the machine learning, pipeline, including feature selection, model selection, and hyperparameter tuning, to streamline, model development.

Project Outline:

The project outline involves the following steps sequentially -

- Installing Libraries
- Importing dependencies
- Data preprocessing
- Feature Selection
- Model Selection and Tuning
- Evaluating the model

The normal ML pipeline looks something the above sequential steps. Auto AI is to be implemented so that we can automize certain specific steps to increase our efficieny and decrease our time in choosing the right model and extract decisive features.

Feature Selection:

In light of the project I have used Random Forest classifier for feature selection due to its capability to assess the importance of individual features within a dataset. Random Forest utilizes an ensemble of decision trees, and during the training process, it measures the contribution of each feature in predicting the target variable. This information is then leveraged for feature selection through methods like `SelectFromModel` or `RFE` (Recursive Feature Elimination). The advantage of Random Forest lies in its ability to handle non-linear relationships, feature interactions, and complex datasets. Moreover, it tends to be less prone to overfitting compared to individual decision trees. While Random Forest was used for illustration, the choice of feature selection method depends on the specific characteristics of the data and the goals of the machine learning task. Different algorithms may be more suitable in various scenarios, and experimentation is often required to identify the most effective approach.

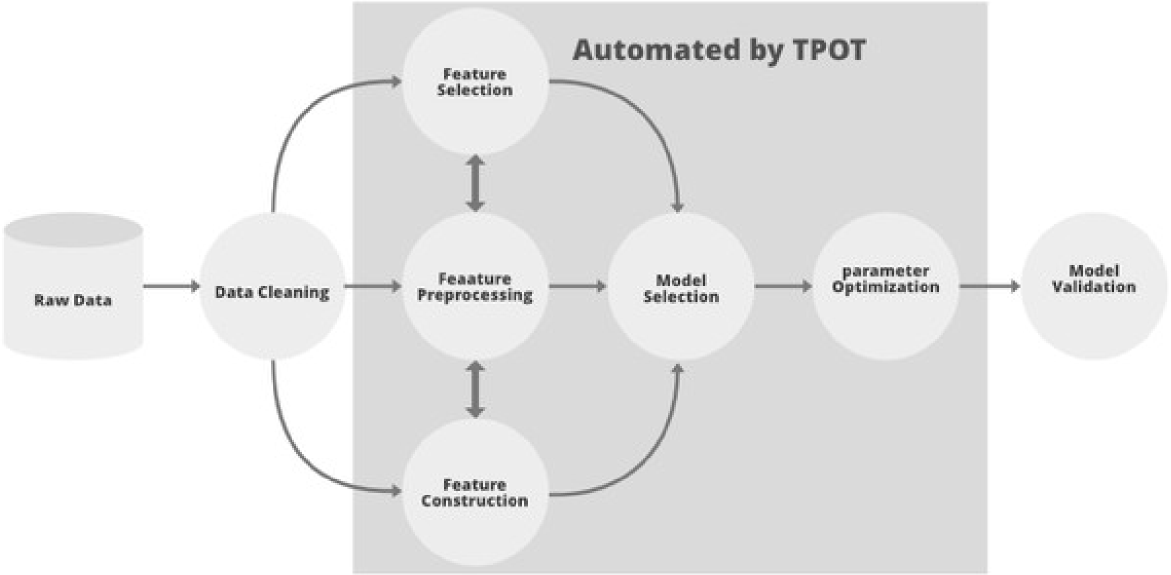
Model Selection:

For model selection I have used TPOT package from python. Tpot is an automated machine learning package in python that uses genetic programming concepts to optimize the machine learning pipeline. It automates the most tedious part of machine learning by intelligently exploring thousands of the possible to find the best possible parameter that suits your data. Tpot is Tpot is built upon the scikit-learn, so its code looks similar to the scikit-learn.

TPOT (Tree-based Pipeline Optimization Tool) is an open-source AutoML tool that automates the process of pipeline optimization for machine learning. TPOT uses genetic programming to automatically search for the best machine learning pipeline, including data preprocessing, feature selection, model selection, and hyperparameter tuning.

Benefits of using TPOT AutoML include:

1. Easy to use: TPOT simplifies the machine learning process by automating many of the repetitive and time-consuming tasks.
2. High-quality pipelines: TPOT generates high-quality pipelines using a range of techniques, including genetic programming and hyperparameter tuning.
3. Customizable: TPOT is highly customizable, allowing users to define various hyperparameters and constraints to generate pipelines that meet their specific needs.
4. Scalable: TPOT can scale to large datasets and distributed environments, making it suitable for big data applications.
5. Open-source: TPOT is open-source and free to use, making it accessible to a wide range of users and organizations.



Use Case: Customer Churn Prediction

Let's consider a use case where you have a dataset related to customer churn in a subscription-based service, and you want to predict whether a customer is likely to churn or not. This is a common business problem, and automating the machine learning pipeline can be beneficial in such scenarios.

Dataset:

For this project I have used a dataset from Kaggle.

Basically, this dataset contains information on various features related to the customer behavior and background with respect to the service. In this use case, a Telecom service provider is the business and the dataset contains customer information.

For better understanding please refer to the kaggle link - https://www.kaggle.com/code/bhartiprasad17/customer-churn-prediction/notebook#-3.-Undertanding-the-data

Goal:

Build a predictive model that automates the machine learning pipeline to predict customer churn based on various features.

The code for the project is below along with other useful links -

Project link - https://colab.research.google.com/drive/1Ti84QWikpnrR5GaCSj5zlS35E2TmzXbK?usp=sharing

TPOT package link - http://epistasislab.github.io/tpot/