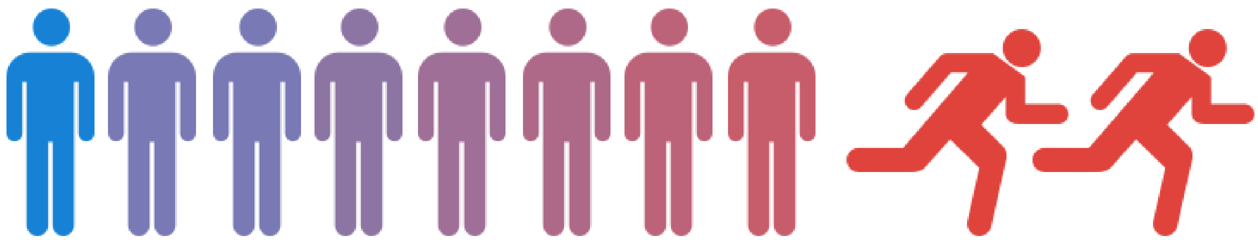
Customer Churn Project



Approach:

The project aims to predict customer churn for a telecommunications company using machine learning. Customer churn is a critical issue for telecommunications companies, as it can lead to a loss of revenue and market share. Therefore, predicting customer churn can help companies take proactive measures to retain customers and improve their business performance.
The project involves several steps, including data preprocessing, feature selection/engineering, model selection, hyperparameter tuning, and evaluation metrics.

Data Preprocessing:

The Telco Customer Churn dataset was loaded using pandas. The dataset contained 7043 rows and 21 columns. The data was preprocessed using the Preprocessing class in the preprocessing.py file. The preprocessing steps included dropping unnecessary columns, converting categorical variables to numerical variables using one-hot encoding, and scaling the data using StandardScaler. Data preprocessing is an essential step in machine learning, as it helps to clean and transform the data into a format that can be used by machine learning algorithms.

Feature Selection/Engineering:

No feature selection or engineering was performed in this project. Feature selection and engineering are techniques used to select or create relevant features that can improve the performance of machine learning models. However, in some cases, the dataset may already contain relevant features, and feature selection/engineering may not be necessary.

Model Selection:

The project used an EnsembleModel class in the model.py file to train an ensemble of three different machine learning models: RandomForestClassifier, GradientBoostingClassifier, and LogisticRegressionCV. The models were chosen based on their ability to handle classification tasks and their performance on similar datasets. Ensemble learning is a technique that combines multiple machine learning models to improve their performance and reduce overfitting.

Hyperparameter Tuning:

The hyperparameters for each model were tuned using GridSearchCV in the SingleModel class of the model.py file. The hyperparameters were chosen based on their ability to improve model performance and reduce overfitting. Hyperparameter tuning is an essential step in machine learning, as it helps to optimize the performance of machine learning models by selecting the best hyperparameters.

Evaluation Metrics:

The performance of the models was evaluated using accuracy, confusion matrix, and ROC curve. The accuracy score was used to measure the overall performance of the models. The confusion matrix was used to measure the number of true positives, true negatives, false positives, and false negatives. The ROC curve was used to measure the trade-off between true positive rate and false positive rate. Evaluation metrics are essential in machine learning, as they help to measure the performance of machine learning models and compare them to other models.

Test Results:

The TestModel class in the test.py file was used to test the accuracy and functionality of the EnsembleModel class. The test_accuracy method tested the accuracy of the ensemble model, and the test_models_has_coefs method tested that the LogisticRegressionCV estimator had coefficients. Testing is an essential step in machine learning, as it helps to ensure that the models are working correctly and producing accurate results.

Conclusion:

The project successfully predicted customer churn for a telecommunications company using an ensemble of three different machine learning models. The models were trained using preprocessed data, and their hyperparameters were tuned using GridSearchCV. The performance of the models was evaluated using accuracy, confusion matrix, and ROC curve. The project demonstrated the importance of data preprocessing, model selection, hyperparameter tuning, and evaluation metrics in machine learning. By following these steps, machine learning models can be optimized to produce accurate and reliable results.