**Machine Learning project**

From 1st March 2021 to 15th March 2021, worked about 30hrs/week on the project with aim to build

an end to end 3 machine learning models on a3\_train data set and tested on a3\_test dataset.

I performed Exploratory data analysis to understand the data I am working with. The next step was the creation of baseline models. Having identified a suitable feature set to train our models with, I have trained three baseline models with the default parameters to get an idea of what is the minimum performance that can be achieved before performing feature engineering and optimizing the model’s hyperparameters. Further I made changes to the data that could improve your model’s performance. By selecting best features and dropping correlated columns or making a new feature out of two or more features. The further step was to create machine learning models after applying feature engineering and then comparing the improvement in results. The last section was discussion which included the model’s comparison and further improvements.

The detailed explanation of each steps can be found below and the python code can be found on GitHub (https://github.com/mraunak/Research-Assistant-Work-/blob/main/MachineLearning.ipynb)

1. Introduction

Given the following four data files:

1. a3-train.data2.
2. a3-train. labels
3. a3-test.data4.
4. a3-test. labels

There are 2,000 training rows and 600 test rows, each row has 500 features, and there are only two labels, -1 and 1.

1. Exploratory Data Analysis

The first step in the machine learning pipeline is taking a look at the data to make sure that I understand what I am working with.

Step1: The file was converted into data frame for convenience.

Step2: Segregated the data into continuous numeric data and categorical data.

Step3: Checked if there are any missing values in the given data sets and found there are no missing values. If there were missing values, then we can handle that by replacing missing values by ’mean’ in features with continuous numeric data and by ’mode’ in features with categorical data. There are no categorical data.

Step4: Could not find Unique Identifier Column.

Step5: Checked whether the given datasets are balanced or not, and found the datasets are balanced.

Step6: Found Correlation between Input variables and Target variables. The correlations were not significant.1

3 Baseline Models

Having identified a suitable feature set to train our models with, I have trained at least three baseline models with the default parameters to get an idea of what is the minimum performance that can be achieved before performing feature engineering and optimizing the model’s hyperparameters. Formulated 3 baseline models:

1. Random Forest Classifier

Below are the accuracies before performing feature engineering and optimizing the model’s hyperparameters.

Value of Accuracy train: 0.993000

value of Accuracy test: 0. 6983332.

1. Gradient Boosting Classifier

Below are the accuracies before performing feature engineering and optimizing the model’s hyperparameters.

Value of Accuracy train: 1.000000

value of Accuracy test: 0.840000

1. K-Nearest Neighbors Classifier

Below are the accuracies before performing feature engineering and optimizing the model’s hyperparameters.

Value of Accuracy train: 0.826500

value of Accuracy test: 0.723333

4 Feature Engineering

I made changes to the data that could improve your model’s performance. Maybe dropping correlated columns or making a new feature out of two or more.

1. Normalization of Data set. It did not prove much effective to improve accuracy of my baseline models.

2. Used Select Best function to select important features, on which I trained and tested the three models.

For Random Forest Classifier and K-Nearest Neighbors Classifier selected 15 best features for maximum accuracy and for Gradient Boosting Classifier selected 12 features for maximum accuracy.

5 Model Building

I built three models in which and tried to achieve the highest possible performance on the test set.

Formulated 3 models:

1.Random Forest Classifier

Below are the accuracies after performing feature engineering and optimizing the model’s hyperparameters.

Value of maximum Accuracy train: 0.995000

value of maximum Accuracy test: 0. 880000

2.Gradient Boosting Classifier

Below are the accuracies after performing feature engineering and optimizing the model’s hyperparameters?

value of maximum Accuracy train: 1.000000

value of maximum Accuracy test: 0.870000

3.K-Nearest Neighbors Classifier

Below are the accuracies after performing feature engineering and optimizing the model’s hyperparameters.

Value of maximum Accuracy train: 1.000000

value of maximum Accuracy test: 0.906667

6 Discussion

Describe how the models compared against each other and the baselines.

Was their performance as good as you expected?

Were there any challenges? Is there anything that could improve your model’s performance?

The accuracies of the models were improved by tuning the hyper parameters and by Feature Engineering, for Random Forest, n-estimators were tuned to improve accuracy.

In case of Gradient Boosting Classifier learning rate, n-estimators, max depth were tuned to improve the model accuracy. In case of K Nearest Neighbors n-neighbors, p, leaf size was tuned to improve the model accuracy. Selecting important features improved accuracy as well as the running time. The Models (After tuning the hyperparameters and Feature Engineering) were compared based on their accuracies on test data and found that Gradient Boosting Classifier gave highest accuracy among all three models. The baseline models (Before tuning the hyperparameters and Feature Engineering) were compared based on their accuracies on test data and found that K-Nearest Neighbors Classifier gave highest accuracy among all three models. Yes, there were challenges to tune the hyperparameters and to select best features, optimal hyperparameters combination, a ’for loop code’ was written. There might be some more powerful classification techniques which could have provided better results.