Project_82_phase_04

1. Ensemble:

Ensemble is a general approach to machine learning that seeks better predictive performance by combining the predictions from multiple models.

Ensemble methods used for classification are:

a. Max. Voting

It is mainly used for classification problems. The method consists of building multiple models independently and getting their individual output called ‘vote’. The class with maximum votes is returned as output.

b. Extreme Gradient Boosting(XGB)

XGBoost is a decision-tree-based ensemble Machine Learning algorithm that uses a gradient boosting framework.

Advantages:

I. It is Highly Flexible and uses the power of parallel processing.
II. It is faster than Gradient Boosting and supports regularization.
III. It is designed to handle missing data with its in-build features.
IV. The user can run a cross-validation after each iteration.

Disadvantages:

I. Difficult interpretation, visualization tough
II. Over fitting possible if parameters not tuned properly.
III. Harder to tune as there are too many hyper parameters.

c. Gradient Boosting

Advantages:

I. train faster especially on larger datasets,
II. most of them provide support handling categorical features,
III. Some of them handle missing values natively.

Disadvantages:

i. Prone to over fitting.
ii. models can be computationally expensive
iii. Hard to interpret the final models.

d. Light GBM

Advantages:

I. Faster training speed and higher efficiency.
II. Lower memory usage
III. Better accuracy than any other boosting algorithm and compatibility with Large Datasets

Disadvantages of Light GBM

i. Over fitting: Light GBM split the tree leaf-wise which can lead to over fitting as it produces much complex trees.

ii. Compatibility with Datasets: Light GBM is sensitive to over fitting and thus can easily overfit small data.

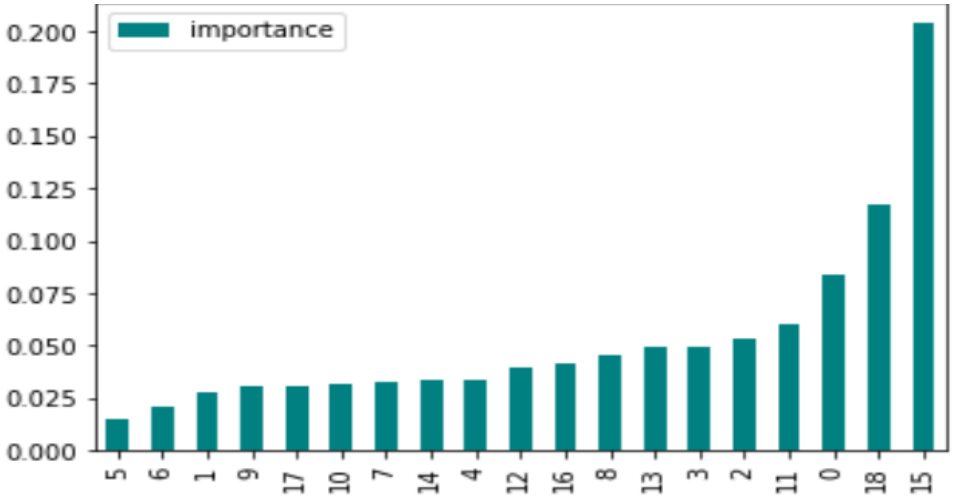
	Model	Train data_F1 score	Test data_F1score
0	K Nearest Neighbours	0.75	0.68
1	Randomforest	0.87	0.74
2	Decision Tree	0.76	0.74
3	MAx Voting ensemble	0.82	0.77
4	XGboost	0.91	0.79
5	Gradient boost	0.77	0.76
6	Light GBM	0.79	0.77

For Max voting ensemble method, K nearest neighbors, Decision Tree and Random Forest are used. The F1_score of max voting ensemble model is better than decision tree and k Nearest Neighbor. The test data F1_score is significantly improved to 77%. XGboost ensemble has highest Train data F1_score 0.91 and test data F1_score 0.79 compared to all the models.

2. Feature importance and Ranking

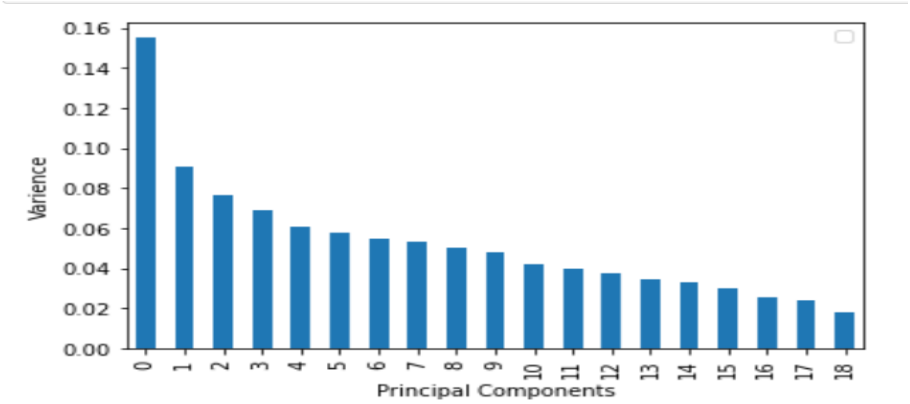
For improving the performance of the model, most important features are selected based on their variance. Random forest regressor and Principal component analysis is used for selecting the model.

Random Forest Regressor:



Features are plotted in increasing order in the above bar graph with column numbers. Only high importance features are selected for improving the performance.

Features with highest variance is selected with feature column numbers and stored in another data frame for modeling.
Using the above method for feature selection, modeling with top 12 important features on base line model K-Nearest neighbors F1_score increased on train data increased from 0.75 to 0.83 and on test data F1_score increased from 0.68 to 0.83
Principal Component Analysis:



Using the above method for feature selection, modeling with top 12 principal components on base line model K-Nearest neighbors F1_score increased on train data increased from 0.75 to 0.79 and on test data F1_score increased from 0.68 to 0.74

From the both feature selection methods used on same base line model K nearest neighbours model performance significantly improved with the features selected from random forest regressor.