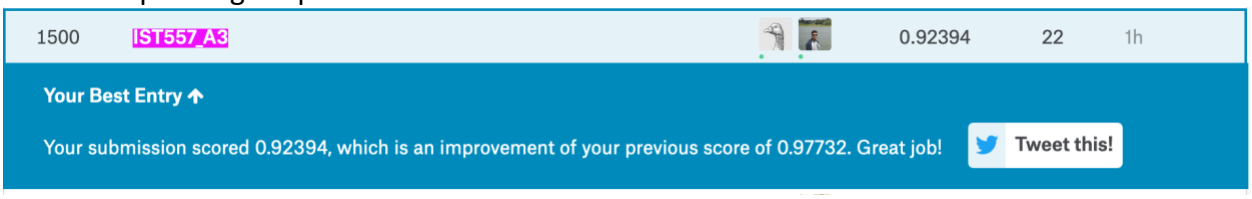
IST 557 Project Progress-2

Mihir Mehta (mzm6664) & Mudit Garg (mxg5783

Best performance:

The best performance- 0.92394

The corresponding snapshot is:



Best Method

Stacking approach using following models:

1. XGBoostRegressor and Random forest as a base learner.
2. Linear Regression as a meta learner

In order to clarify, Stacking is an ensemble of models combined sequentially. Stacking uses a "meta learner" to combine the predictions of "base learners."

The predictions of the base learners are input for the meta-learner.

Motivation:

1. https://www.kaggle.com/anqitu/feature-engineer-and-model-ensemble-top-10/output
2. https://www.kaggle.com/dimitreoliveira/model-stacking-feature-engineering-and-eda

Other Methods

1. At the previous checkpoint, we used XGBoostRegressor which yielded us RMSE of 0.98.
2. After the team #1 presentation, we removed the outliers from our training set, instead of clipping them. This reduced the RMSE to 0.97
3. We fitted different variants of XGBoost Regressor and see if RMSE can be decreased further. We had no luck. We suspect that our model fail to additional delta of information to improve our results.
4. On further reading, we found an approach called stacking and tried to utilize it in our code.
5. We tried blending approach by combining multiple XGBoost regression models' predictions using meta regressors like ridge, lasso, RandomForest Regressors. In fact, our result further deteriorated. It is due to the fact predictions were highly correlated and they did not add any additional value. RMSE was 0.99
6. All these approaches made us think to try regressor different from XGBoost and combine with XGBoost regressor using stacking approach to see if model performance can be improved.
7. Thus, we tried using XGboost, KNN cluster, linear regression, cat boost regressor as base model and linear regression as meta model. This deteriorated the performance even more. It yielded the RMSE of 1.13.
8. We finally tried using the method mentioned as best method. It was just a modification of approach used in point 6. It gave us an RMSE of 0.92394.

Summary

Method	RMSE
XGBRegressor() with outliers clipped	0.98
XGBRegressor() deleting the outlier	0.97
XGBoost, Ridge, lasso, randomForest	0.99
Stacking: XGBoost, CatBoost, KNN, linear regression	1.13
Stacking: XGBoost, random forest, linear regression	0.9234

Future plans

1. We designed features for this iteration using motivation#1. In future, we want to see if we can use features created in iteration 1 to see if accuracy is improved
2. Apart from random forest and xgboost, we are planning to add LSTM and GBR as regressors and see model can be improved.
3. Also, we will see if we can can change meta model – second level model

Contribution:
Mudit Garg- 50%
Mihir Mehta – 50%

References

1. https://www.kaggle.com/dimitreoliveira/model-stacking-feature-engineering-and-eda
2. https://www.kaggle.com/anqitu/feature-engineer-and-model-ensemble-top-10/output
3. https://www.kdnuggets.com/2017/02/stacking-models-imropved-predictions.html
4. https://bradleyboehmke.github.io/HOML/stacking.html