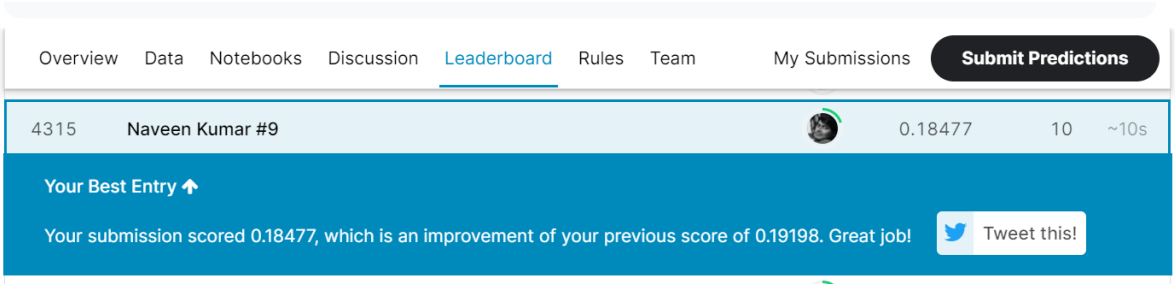
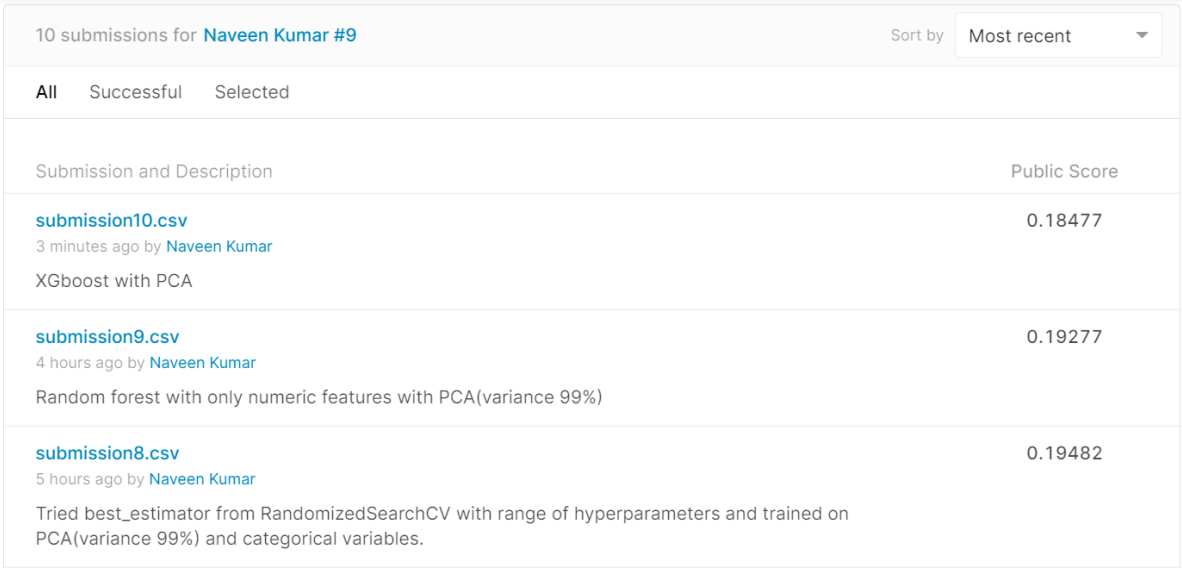
Final Project Report

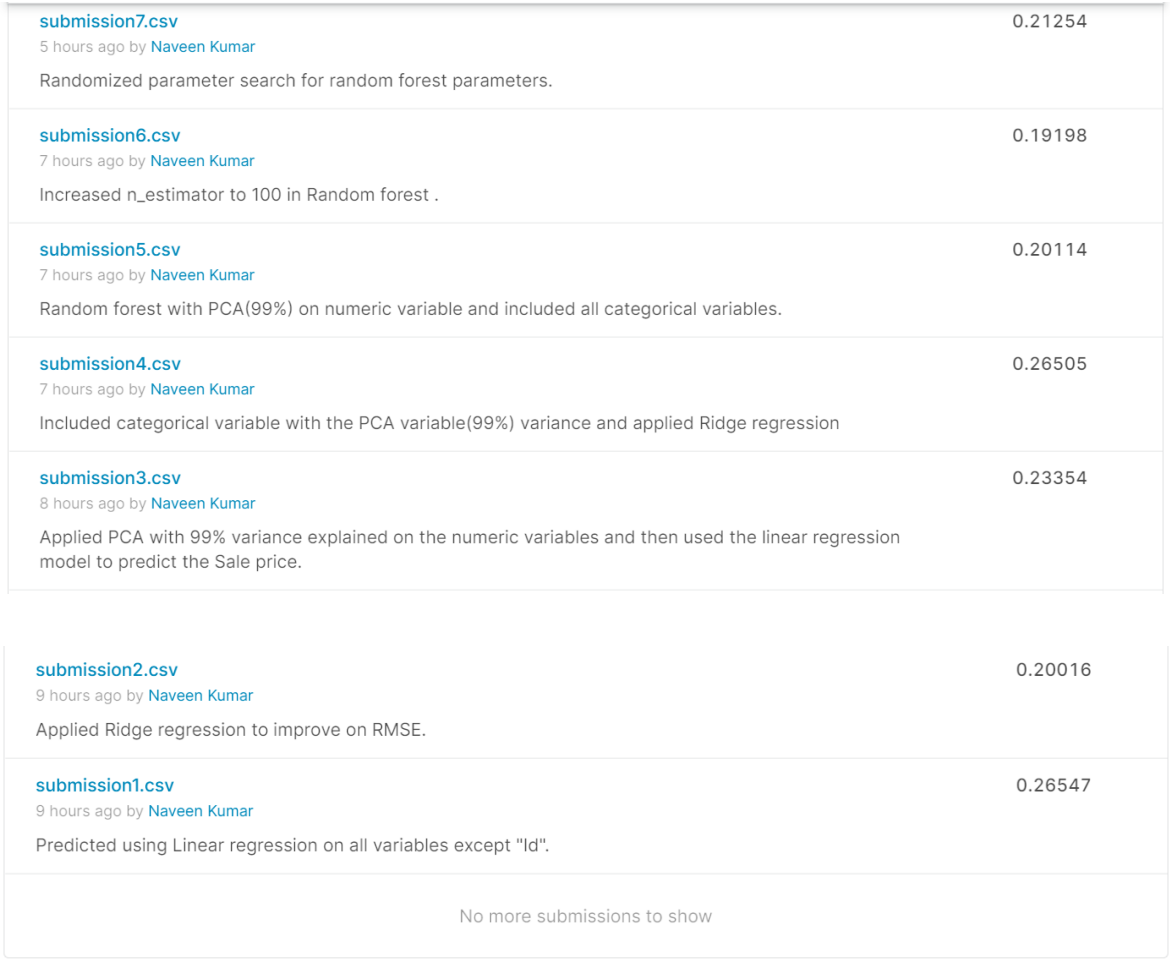
Done By – Naveen Kumar #9 (Kaggle Username)

- Kaggle Leaderboard:



- Kaggle Submissions





- Details of Changes made in Each Submission

1. Submission 1:

   Started with Importing the train data. Identified the categorical variables and then and encoded them to make numeric. Here, NaN has been coded as -1 so that we didn't had much data loss.

   Then identified the numeric variables and interpolate them with average of the respective columns.

   After preparing the data, split the train data into training and testing (70:30). Defined the Linear Regression Models and fit the model with training data. Once trained, predicted the SalePrice on the testing split and evaluated using Mean squared Error.

   MSE: 0.02661

   Generated the submission prediction and submitted on Kaggle. **Score: 0.2654**

2. Submission 2:

Experimented with Ridge regression for different Alpha (0.01,0.1,1, 10,100) values.

Below are the corresponding MSEs:

Alpha: 0.01
MSE: 0.0265939155743259

Alpha: 0.1
MSE: 0.026407385462780854

Alpha: 1
MSE: 0.025200426721766273

Alpha: 10
MSE: 0.023356314364559625

Alpha: 100
MSE: 0.02270194557790402

Alpha: 1000
MSE: 0.025127733454935623

We get the best MSE for alpha: 100. We fit the Ridge model and generated prediction for test data and submitted on Kaggle.

Kaggle Score: 0.20016

3. Submission 3:

Tried PCA for the numeric variables. Chose n_component =0.99 in order to get variance explained 99%.
Considered only numeric variables from PCA and modeled linear regression.

MSE: 0.024264
Kaggle Score: 0.23354

4. Submission 4:

   Included PCA and other categorical coded variables and modelled Ridge regression with alpha 100.

   MSE: 0.022261
   Kaggle Score: 0.26505

5. Submission 5:

   Used random forest model with (n_estimators = 10) on the PCA and categorical variables included in the training set.

   MSE: 0.023376
   Kaggle Score: 0.20114

6. Submission 6:

   Used random forest model with (n_estimators = 100) on the PCA and categorical variables included in the training set.

   MSE: 0.022072
   Kaggle Score: 0.19198

7. Submission 7:

   Tried the random forest (n_estimator= 100) with raw numeric variables without PCA and categorical coded variables.

   MSE: 0.023670
   Kaggle Score: 0.21254

8. Submission 8:

   Applied RandomizedSearchCV on range of Hyperparameters for Random Forest Model. Here are the hyperparameter dictionary.

{'n_estimators': [100, 311, 522, 733, 944, 1155, 1366, 1577, 1788, 2000], 'max_depth': [30, 40, 50, 60, 70, 80, 90, 100, 110, 120, None], 'max_features': ['auto', 'sqrt'], 'min_samples_split': [2, 3, 5, 10], 'min_samples_leaf': [1, 2, 4], 'bootstrap': ['True', 'False']}

Used PCA and categorical coded variables for training and the Best Estimator for Evaluation.

MSE: 0.0226022
Kaggle Score: 19482

9. Submission 9:

Used only numeric variables and calculated PCA with 99% variance explained. Used this data to train the above mentioned Random forest model.

MSE: 0.022108
Kaggle Score: 19277

10. Submission 10:

Used the PCA on numeric variable with 99% variance explained and combined the categorical coded data to create training and test split. Fit the XGBoost model on the training data.

MSE: 0.019410
Kaggle Score: 0.18477

- **Program File Included**: Final_project.ipynb