Final Report: Kaggle House Prices Prediction

Team Name: hzm401

Contestants Name:
Hasanuzzaman
Md. Shamim Reza

ID:
1640CSE00512
1640CSE00536

Links:
https://www.kaggle.com/hzm401
https://github.com/hzm401

1)Project Goal

This competition is hosted by data analysis club IIT Palakkad. The challenge is to predict the price of each house given some information related to houses. Goal of competition is to make us familiar with environment of kaggle and basics of regression, and our goal was to learn and do our best in this contest.

1.1 Problem Statement

This playground competition's dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence. With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, we have to predict the final price of each home.

Data: Collection of data objects and their attributes.

Attribute:

An attribute is a property or characteristic of an object
– Examples: eye color of a person, temperature, etc.
– Attribute is also known as variable, field, characteristic, or feature

Object:

A collection of attributes describe an object
– Object is also known as record, point, case, sample, entity, or instance

Data Description

Training dataset contains 1460 training samples and 1459 testing samples. Each sample is represented by 81 features in the training dataset and 80 features in the test dataset.

Data Preprocessing

1. Aggregation
2. Sampling
3. Dimensionality Reduction
4. Feature subset selection
5. Feature creation
6. Discretization and Binarization
7. Attribute Transformation

Aggreation

– Combining two or more attributes (or objects) into a single attribute (or object)

– Purpose

□ Data reduction (Reduce the number of attributes or objects)
□ Change of scale (Cities aggregated into regions, states, countries, etc)
□ More “stable” data (Aggregated data tends to have less variability)

Sampling

Sampling is the main technique employed for data selection.
– It is often used for both the preliminary investigation of the data and the final data analysis.
– Statisticians sample because obtaining the entire set of data of interest is too expensive or time consuming.

Dimensionality Reduction

Purpose

– Avoid curse of dimensionality
– Reduce amount of time and memory required by data mining algorithms
– Allow data to be more easily visualized
– May help to eliminate irrelevant features or reduce noise

Techniques

– Principle Component Analysis
– Singular Value Decomposition
– Others: supervised and non-linear techniques

Future Subset Selection

Another way to reduce dimensionality of data.

Techniques:
– Brute-force approch:
□ Try all possible feature subsets as input to data mining algorithm
– Embedded approaches:
□ Feature selection occurs naturally as part of the data mining algorithm
– Filter approaches:
□ Features are selected before data mining algorithm is run

Feature Selection

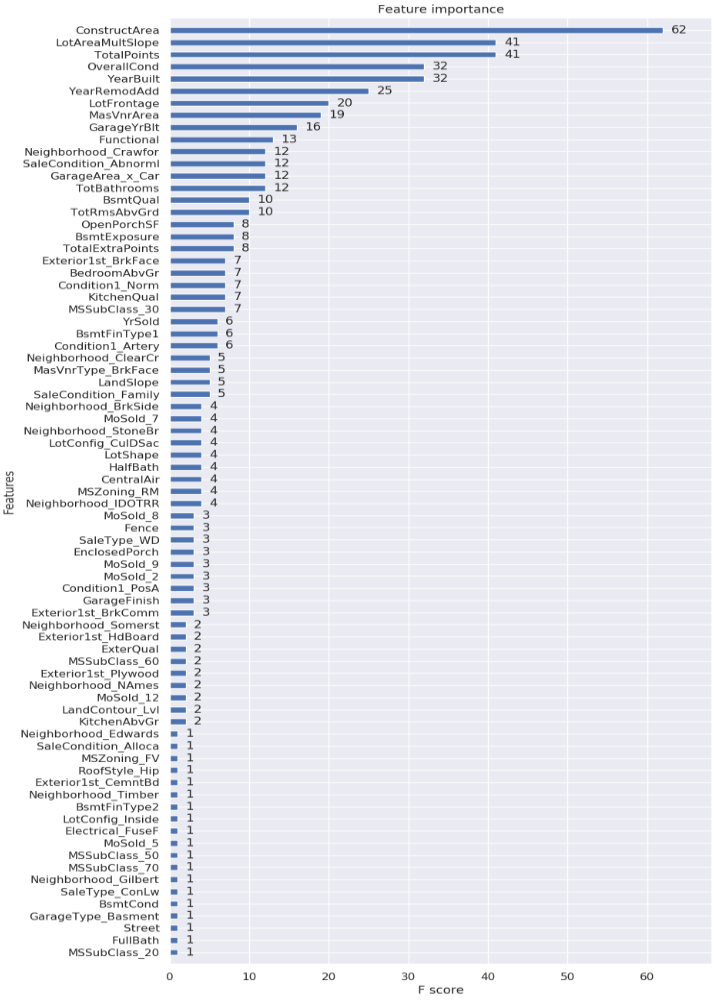
Create new attributes that can capture the important information in a data set much more efficiently than the original attributes.

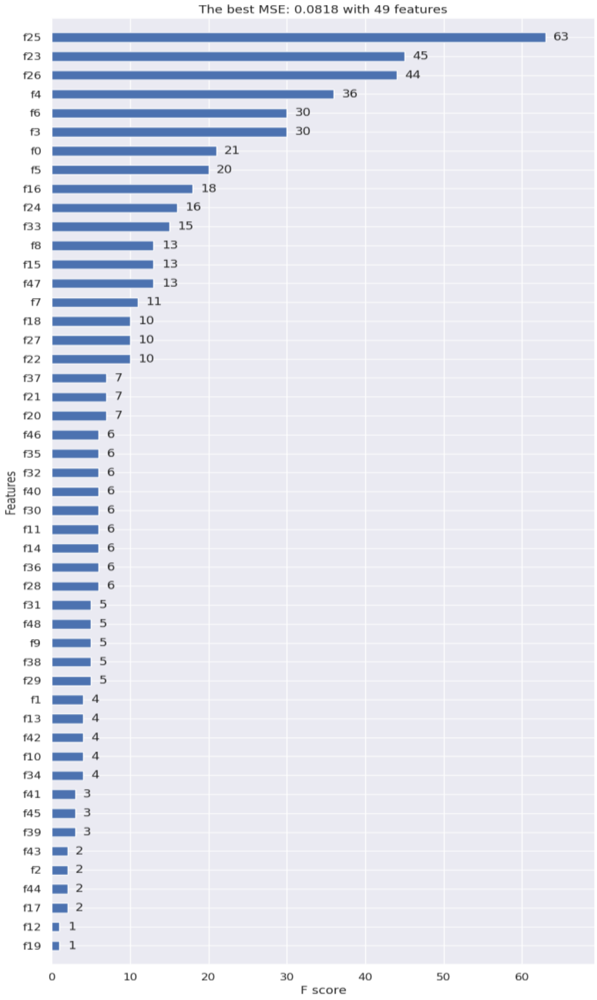
Feature Selection by Gradient Boosting

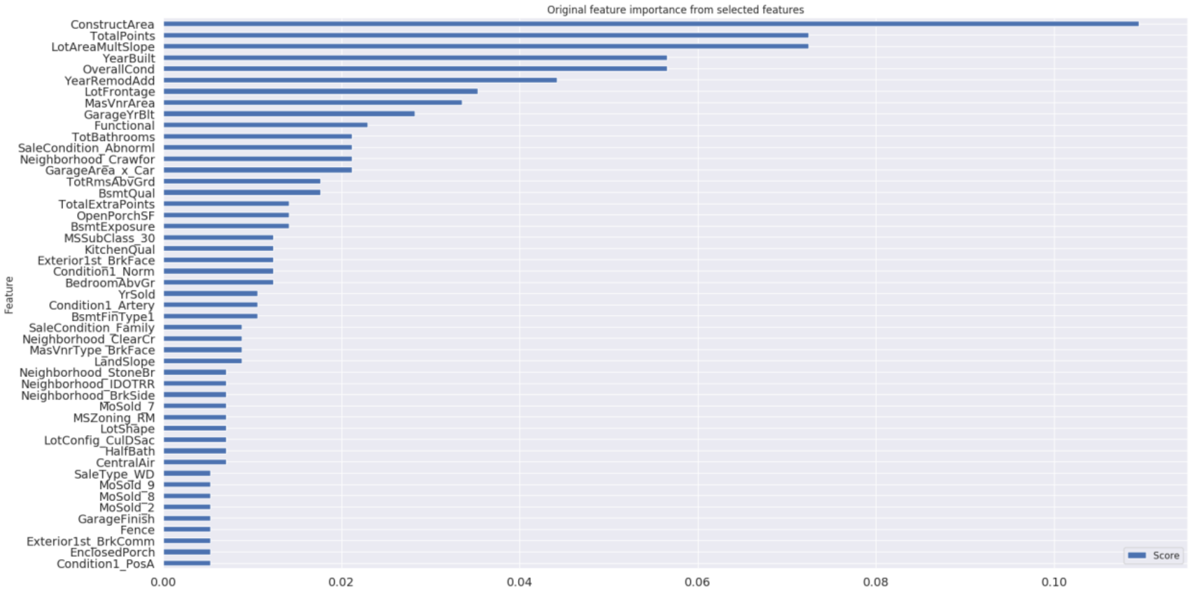
The LightGBM model the importance is calculated from, if 'split', result contains numbers of times the feature is used in a model, if 'gain', result contains total gains of splits which use the feature.

On the **XGBoost** model the importance is calculated by:

- **'weight':** the number of times a feature is used to split the data across all trees.
- **'gain':** the average gain across all splits the feature is used in.
- **'cover':** the average coverage across all splits the feature is used in.
- **'total_gain'**: the total gain across all splits the feature is used in.
- **‘total_cover’:** the total coverage across all splits the feature is used in.


Feature importance
ConstructArea 62
LotAreaMultSlope 41
TotalPoints 41
OverallCond 32
YearBuilt 32
YearRemodAdd 25
LotFrontage 20
MasVnrArea 19
GarageYrBlt 16
Functional 13
Neighborhood_Crawfor 12
SaleCondition_Abnorml 12
GarageArea_x_Car 12
TotBathrooms 12
BsmtQual 10
TotRmsAbvGrd 10
OpenPorchSF 8
BsmtExposure 8
TotalExtraPoints 8
Exterior1st_BrkFace 7
BedroomAbvGr 7
Condition1_Norm 7
KitchenQual 7
MSSubClass_30 7
YrSold 6
BsmtFinType1 6
Condition1_Artery 6
Neighborhood_ClearCr 5
MasVnrType_BrkFace 5
LandSlope 5
SaleCondition_Family 5
Neighborhood_BrkSide 4
MoSold_7 4
Neighborhood_StoneBr 4
LotConfig_CulDSac 4
LotShape 4
HalfBath 4
CentralAir 4
MSZoning_RM 4
Neighborhood_IDOTRR 4
MoSold_8 3
Fence 3
SaleType_WD 3
EnclosedPorch 3
MoSold_9 3
MoSold_2 3
Condition1_PosA 3
GarageFinish 3
Exterior1st_BrkComm 3
Neighborhood_Somerst 2
Exterior1st_HdBoard 2
ExterQual 2
MSSubClass_60 2
Exterior1st_Plywood 2
Neighborhood_NAmes 2
MoSold_12 2
LandContour_Lvl 2
KitchenAbvGr 2
Neighborhood_Edwards 1
SaleCondition_Alloca 1
MSZoning_FV 1
RoofStyle_Hip 1
Exterior1st_CemntBd 1
Neighborhood_Timber 1
BsmtFinType2 1
LotConfig_Inside 1
Electrical_FuseF 1
MoSold_5 1
MSSubClass_50 1
MSSubClass_70 1
Neighborhood_Gilbert 1
SaleType_ConLw 1
BsmtCond 1
GarageType_Basment 1
Street 1
FullBath 1
MSSubClass_20 1
Features
0
10
20
30
40
50
60
F score


The best MSE: 0.0818 with 49 features
f25 63
f23 45
f26 44
f4 36
f6 30
f3 30
f0 21
f5 20
f16 18
f24 16
f33 15
f8 13
f15 13
f47 13
f7 11
f18 10
f27 10
f22 10
f37 7
f21 7
f20 7
f46 6
f35 6
f32 6
f40 6
f30 6
f11 6
f14 6
f36 6
f28 6
f31 5
f48 5
f9 5
f38 5
f29 5
f1 4
f13 4
f42 4
f10 4
f34 4
f41 3
f45 3
f39 3
f43 2
f2 2
f44 2
f17 2
f12 1
f19 1
Features
0
10
20
30
40
50
60
F score


Original feature importance from selected features
ConstructArea
TotalPoints
LotAreaMultSlope
YearBuilt
OverallCond
YearRemodAdd
LotFrontage
MasVnrArea
GarageYrBlt
Functional
TotBathrooms
SaleCondition_Abnorml
Neighborhood_Crawfor
GarageArea_x_Car
TotRmsAbvGrd
BsmtQual
TotalExtraPoints
OpenPorchSF
BsmtExposure
MSSubClass_30
KitchenQual
Exterior1st_BrkFace
Condition1_Norm
BedroomAbvGr
YrSold
Condition1_Artery
BsmtFinType1
SaleCondition_Family
Neighborhood_ClearCr
MasVnrType_BrkFace
LandSlope
Neighborhood_StoneBr
Neighborhood_IDOTRR
Neighborhood_BrkSide
MoSold_7
MSZoning_RM
LotShape
LotConfig_CulDSac
HalfBath
CentralAir
SaleType_WD
MoSold_9
MoSold_8
MoSold_2
GarageFinish
Fence
Exterior1st_BrkComm
EnclosedPorch
Condition1_PosA
Feature
Score
0.00
0.02
0.04
0.06
0.08
0.10

Modeling Methods

First, we started to look at different approaches to implement linear regression models, and use hyper parametrization, cross validation and compare the results between different errors measures.

Evaluate Results

Mean Squared Error (MSE)

In statistics, MSE or mean squared deviation (MSD) of an estimator measures the average of the squares of the errors. MSE is a risk function, corresponding to the expected value of the squared error loss. The fact that MSE is almost always strictly positive (and not zero) is because of randomness or because the estimator does not account for information that could produce a more accurate estimate. Which is simply the average value of the SSE cost function that we minimize to fit the linear regression model. The MSE is useful to for comparing different regression models or for tuning their parameters via a grid search and cross-validation.

$$\operatorname{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

Root-Mean-Square Error (RMSE)

The root-mean-square deviation (RMSD) or root-mean-square error (RMSE) is a frequently used measure of the differences between values predicted by a model or an estimator and the valuesobserved.

$$\operatorname{RMSD}(\hat{\theta}) = \sqrt{\operatorname{MSE}(\hat{\theta})} = \sqrt{\operatorname{E}((\hat{\theta} - \theta)^2)}.$$

Mean Absolute Error (MAE)

In statistics, mean absolute error (MAE) is a measure of difference between two continuous variables, is also the average horizontal distance between each point and the identity line.

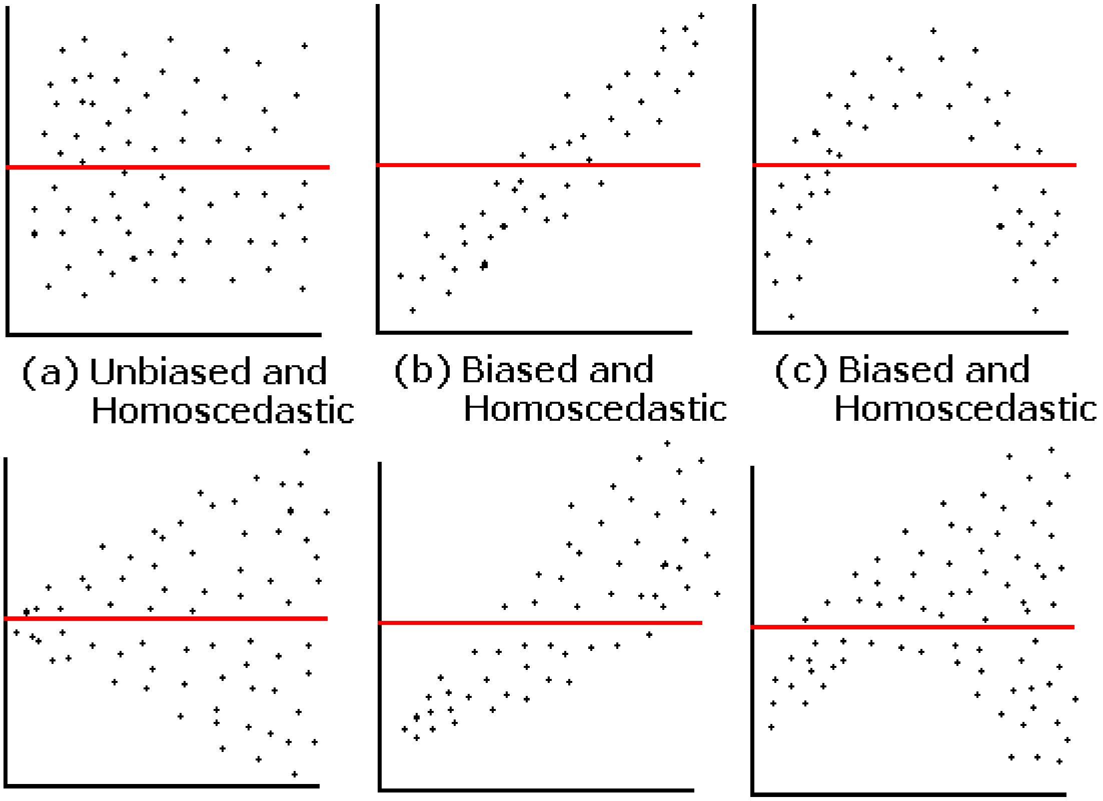
$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} \left| y_i - x_i \right|}{n} = \frac{\sum_{i=1}^{n} \left| e_i \right|}{n}.$$

Coefficient of determination (R^2)

It is the sum of squares of residuals, also called the residual sum of squares: and SStot is the total sum of squares (proportional to the variance of the data): Which can be understood as a standardized version of the MSE, for better interpretability of the model performance (try to say that tree times and faster!). In other words, R^2 is the fraction of response variance that is captured by the model. For the training dataset, R^2 is bounded between 0 and 1, but it can become negative for the test set. If R^2 =1 , the model fits the data perfectly with a corresponding MSE = 0.

Residuals Plots

The plot of differences or vertical distances between the actual and predicted values .Commonly used graphical analysis for diagnosing regression models to detect nonlinearity and outliers, and to check if the errors are randomly distributed .



(a) Unbiased and Homoscedastic

(b) Biased and Homoscedastic

(c) Biased and Homoscedastic

(d) Unbiased and Heteroscedastic

(e) Biased and Heteroscedastic

(f) Biased and Heteroscedastic

- Since Residual = Observed – Predicted positive values for the residual (on the y-axis) mean the prediction was too low, and negative values *mean the* prediction was too high*; 0 means the guess was exactly correct.*
- They're pretty symmetrically distributed, tending to cluster towards the middle of the plot.
- Detect outliers, which are represented by the points with a large deviation from the centerline.
- They're clustered around the lower single digits of the y-axis (e.g., 0.5 or 1.5, not 30 or 150).

- If we see patterns in a residual plot, it means that our model is unable to capture some explanatory information.
- Non-constant error variance shows up on a residuals vs. fits (or predictor) plot in any of the following ways:
    - The plot has a "fanning" effect. That is, the residuals are close to 0 for small x values and are more spread out for large x values.
    - The plot has a "funneling" effect. That is, the residuals are spread out for small x values and close to 0 for large x values.
    - Or, the spread of the residuals in the residuals vs. fits plot varies in some complex fashion.

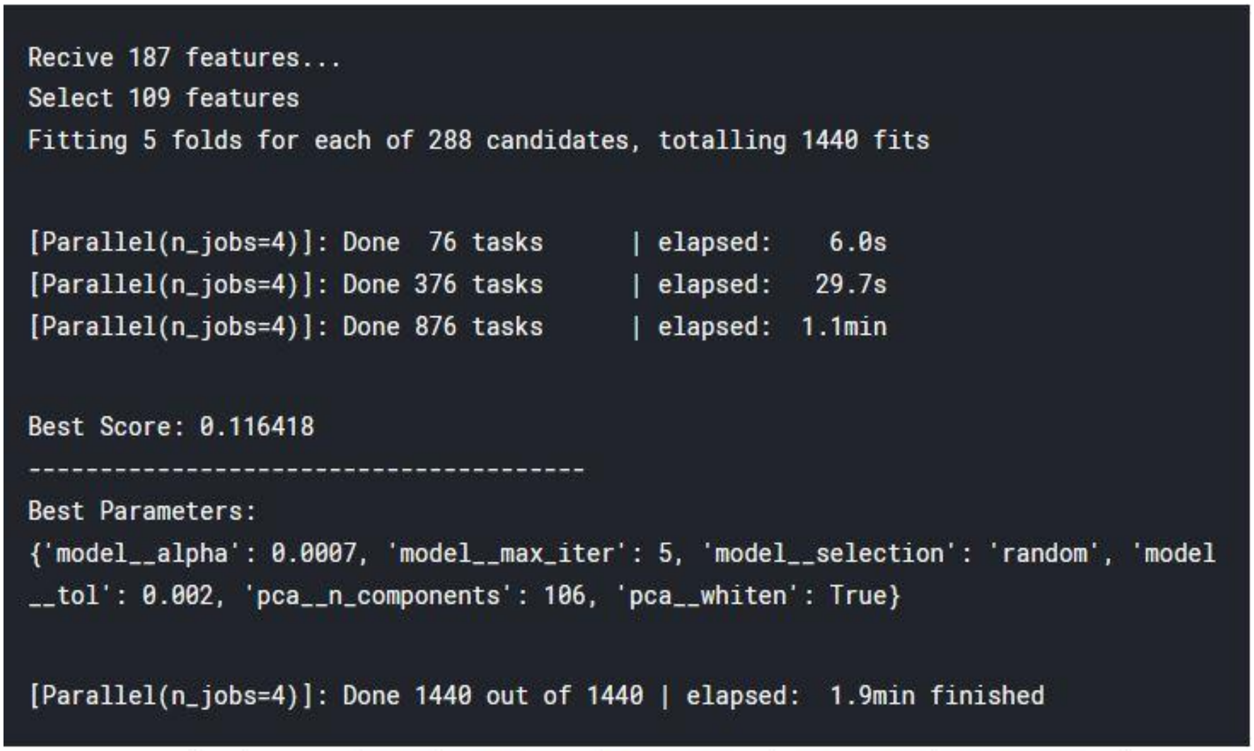
Lasso:

Lasso means Least Absolute Shrinkage and Selection Operator. It is able to achieve both of these goals by forcing the sum of the absolute value of the regression coefficients to be less than a fixed value, which depending on the regularization strength, certain weights can become zero, which makes the Lasso also useful as a supervised feature selection technique, by effectively choosing a simpler model that does not include those coefficients. However, a limitation of the Lasso is that it selects at most n variables if m > n.

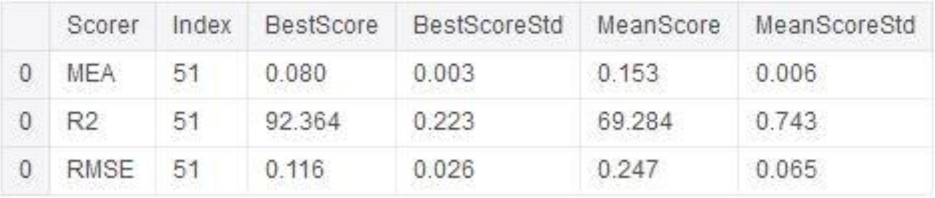
This idea is similar to ridge regression, in which the sum of the squares of the coefficients is forced to be less than a fixed value, though in the case of ridge regression, this only shrinks the size of the coefficients, it does not set any of them to zero.

The optimization objective for Lasso is: (1 / (2 * n_samples)) * ||y - Xw||^2_2 + alpha * ||w||_1

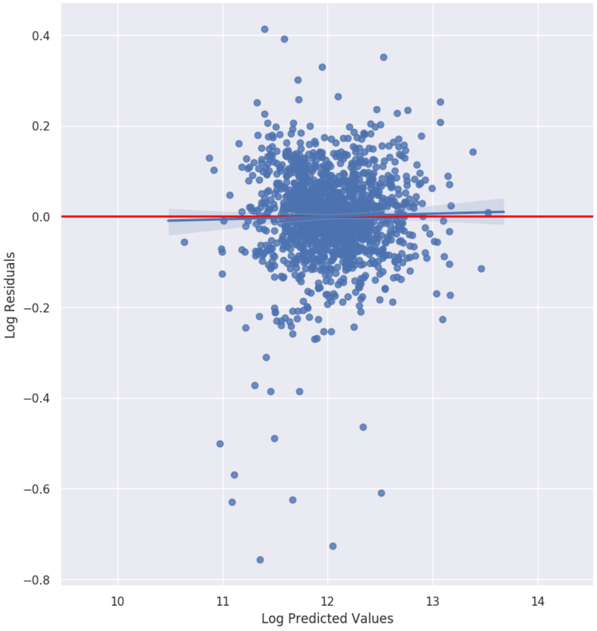
Technically the Lasso model is optimizing the same objective function as the Elastic Net with l1_ratio=1.0, no L2 penalty.

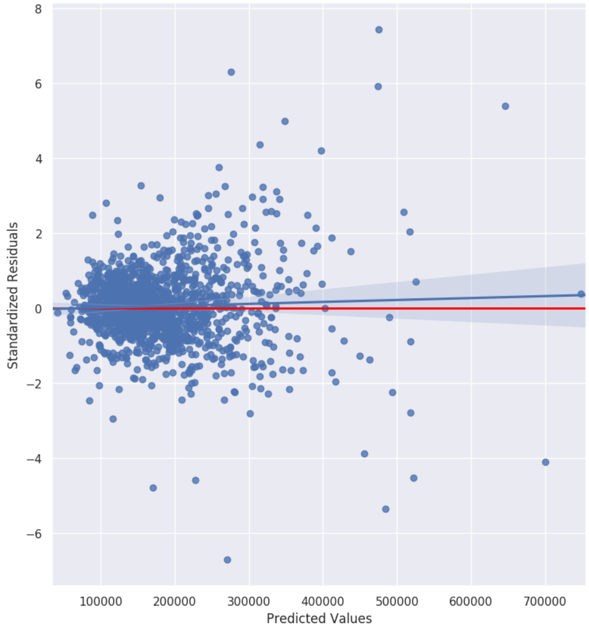


```
Recive 187 features...
Select 109 features
Fitting 5 folds for each of 288 candidates, totalling 1440 fits


[Parallel(n_jobs=4)]: Done  76 tasks      | elapsed:    6.0s
[Parallel(n_jobs=4)]: Done 376 tasks      | elapsed:   29.7s
[Parallel(n_jobs=4)]: Done 876 tasks      | elapsed:  1.1min


Best Score: 0.116418
----------------------------------------
Best Parameters:
{'model__alpha': 0.0007, 'model__max_iter': 5, 'model__selection': 'random', 'model
__tol': 0.002, 'pca__n_components': 106, 'pca__whiten': True}


[Parallel(n_jobs=4)]: Done 1440 out of 1440 | elapsed:  1.9min finished
```



	Scorer	Index	BestScore	BestScoreStd	MeanScore	MeanScoreStd
0	MEA	51	0.080	0.003	0.153	0.006
0	R2	51	92.364	0.223	69.284	0.743
0	RMSE	51	0.116	0.026	0.247	0.065

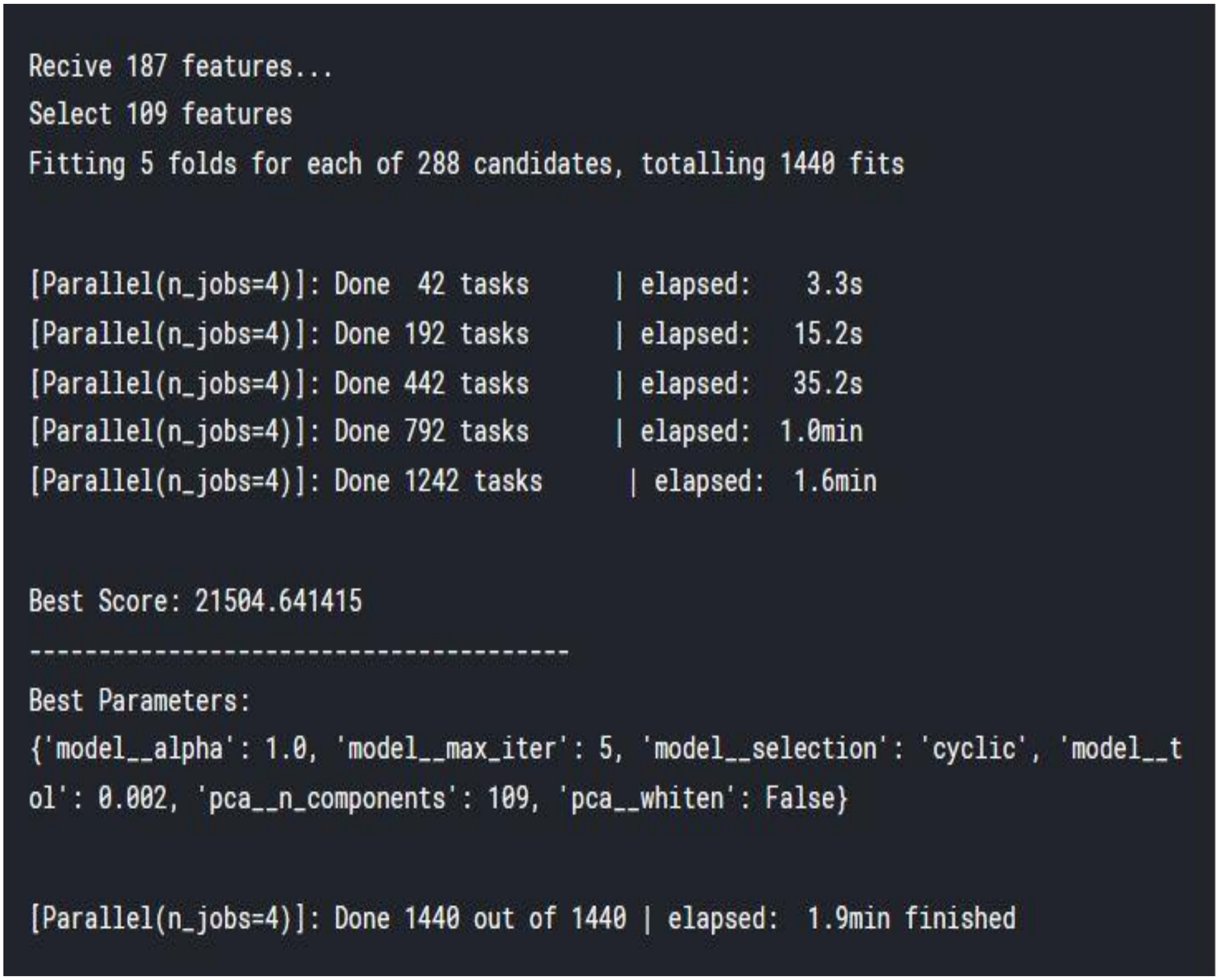




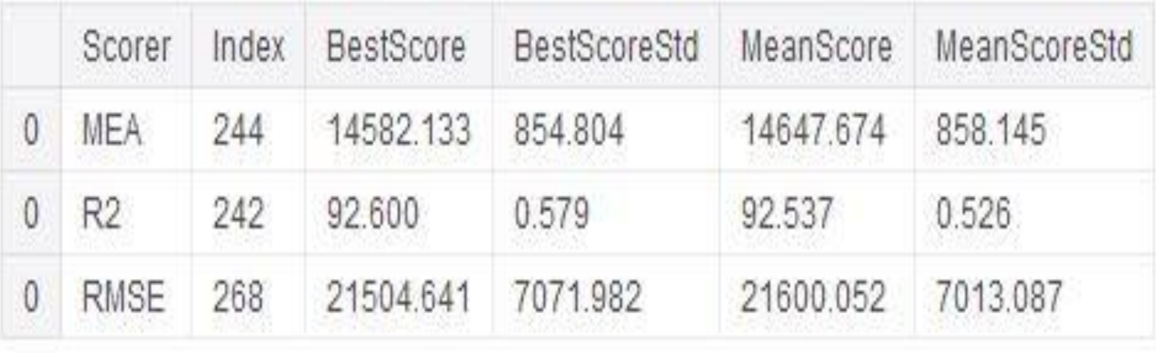
As you can see our Lasso has good performance with RFEcv selection features plus polynomials features, with: MAE 0.080, RMSE 0.1164 and R^2 of 92.36%.

So we decided to remove some outliers , some of the biggest deviations from the log observations perspective.

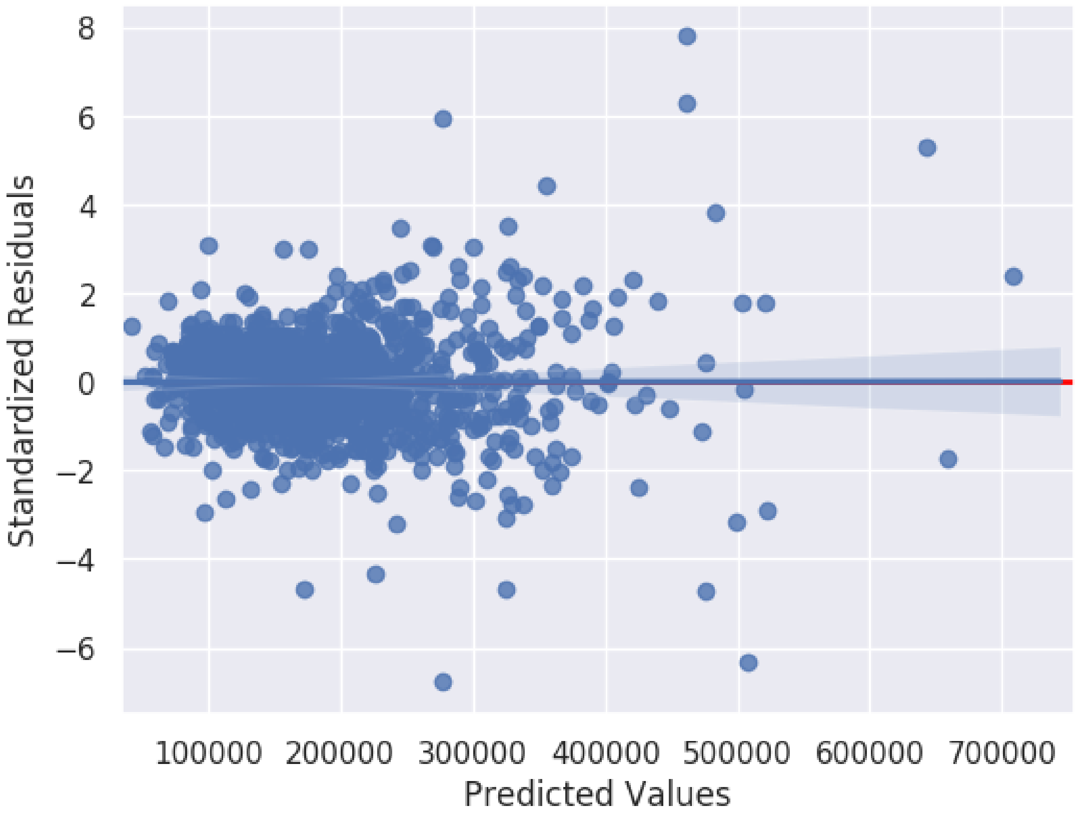
After removing some outliers the result was impressive, shown below:



```
Recive 187 features...
Select 109 features
Fitting 5 folds for each of 288 candidates, totalling 1440 fits


[Parallel(n_jobs=4)]: Done  42 tasks      | elapsed:    3.3s
[Parallel(n_jobs=4)]: Done 192 tasks      | elapsed:   15.2s
[Parallel(n_jobs=4)]: Done 442 tasks      | elapsed:   35.2s
[Parallel(n_jobs=4)]: Done 792 tasks      | elapsed:  1.0min
[Parallel(n_jobs=4)]: Done 1242 tasks      | elapsed:  1.6min


Best Score: 21504.641415
----------------------------------------
Best Parameters:
{'model__alpha': 1.0, 'model__max_iter': 5, 'model__selection': 'cyclic', 'model__t
ol': 0.002, 'pca__n_components': 109, 'pca__whiten': False}


[Parallel(n_jobs=4)]: Done 1440 out of 1440 | elapsed:  1.9min finished
```



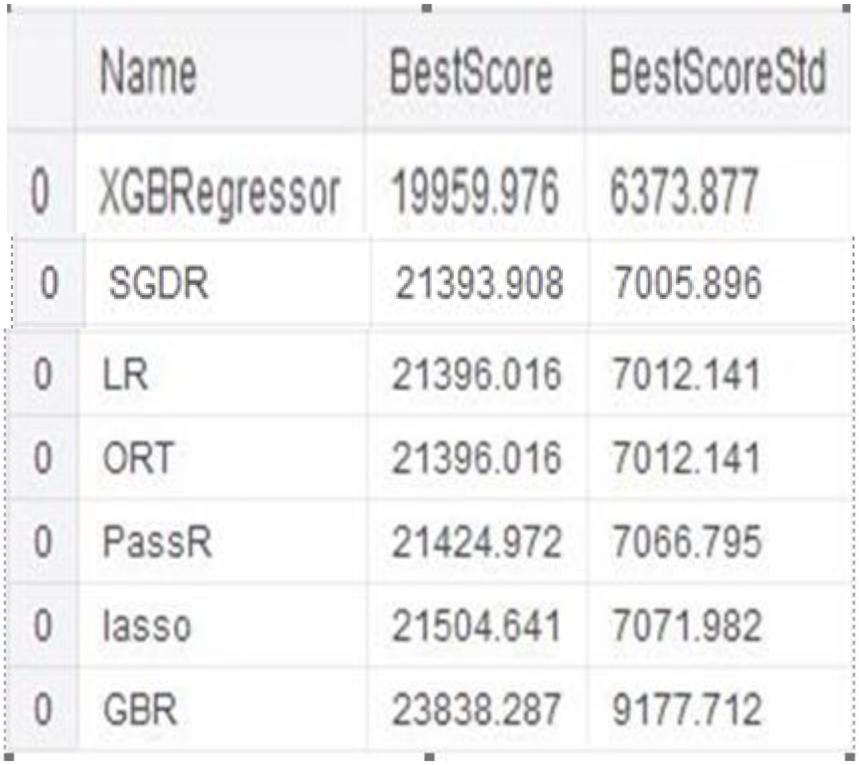
	Scorer	Index	BestScore	BestScoreStd	MeanScore	MeanScoreStd
0	MEA	244	14582.133	854.804	14647.674	858.145
0	R2	242	92.600	0.579	92.537	0.526
0	RMSE	268	21504.641	7071.982	21600.052	7013.087



Modeling:

We used multiple regressions and got the best score from the models hyper parameterization.
Used regressions ar given below:

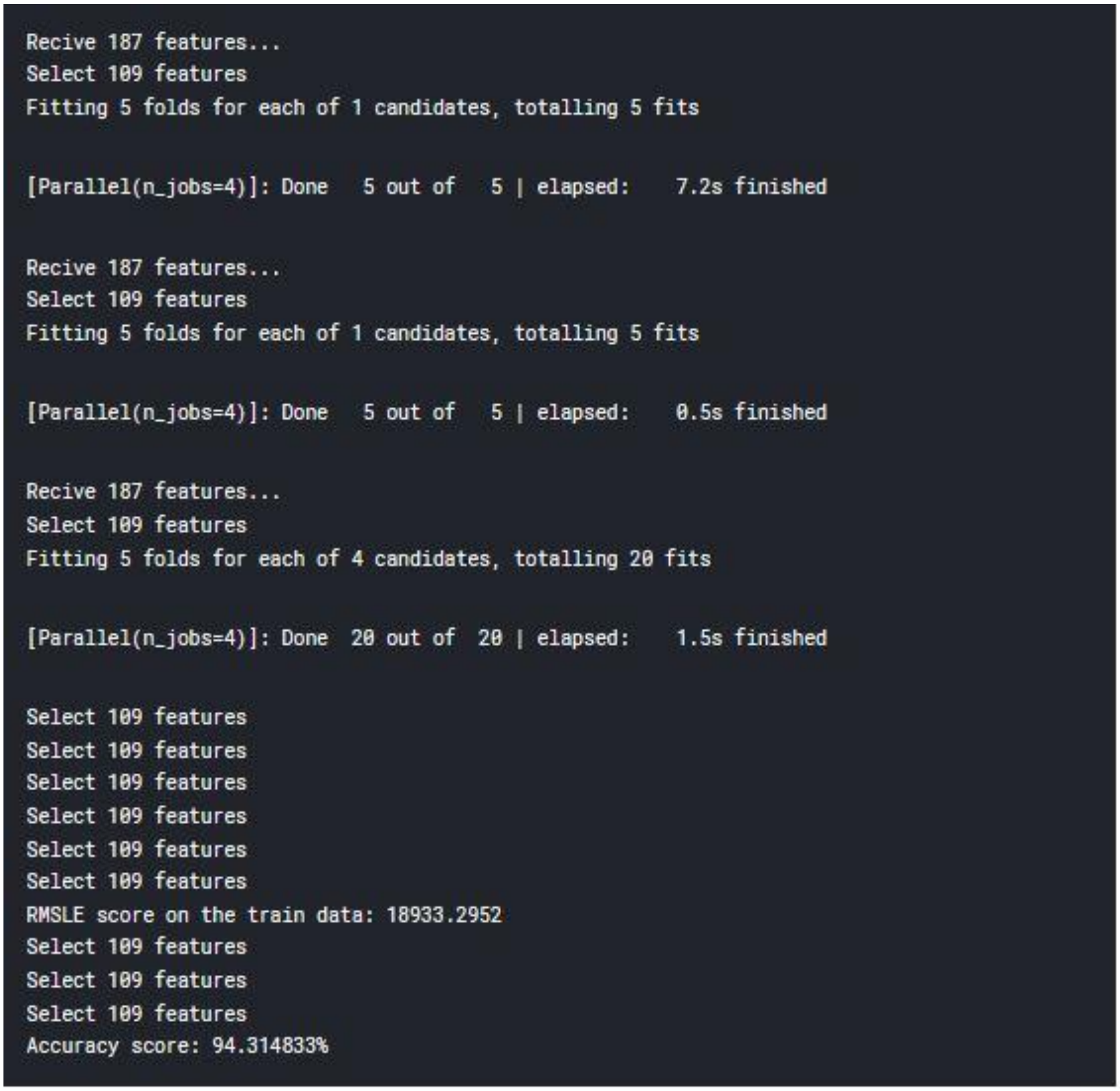
- XGB Reggressor
- SGDR
- LR
- ORT
- PassR
- Lasso
- GBR etc.



	Name	BestScore	BestScoreStd
0	XGBRegressor	19959.976	6373.877
0	SGDR	21393.908	7005.896
0	LR	21396.016	7012.141
0	ORT	21396.016	7012.141
0	PassR	21424.972	7066.795
0	lasso	21504.641	7071.982
0	GBR	23838.287	9177.712

Stacking the Models

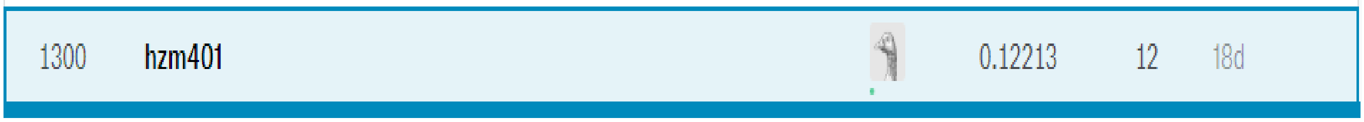
After modeling we stacked the models and averaged base model score.



```
Recive 187 features...
Select 109 features
Fitting 5 folds for each of 1 candidates, totalling 5 fits

[Parallel(n_jobs=4)]: Done   5 out of   5 | elapsed:    7.2s finished

Recive 187 features...
Select 109 features
Fitting 5 folds for each of 1 candidates, totalling 5 fits

[Parallel(n_jobs=4)]: Done   5 out of   5 | elapsed:    0.5s finished

Recive 187 features...
Select 109 features
Fitting 5 folds for each of 4 candidates, totalling 20 fits

[Parallel(n_jobs=4)]: Done  20 out of  20 | elapsed:    1.5s finished

Select 109 features
Select 109 features
Select 109 features
Select 109 features
Select 109 features
Select 109 features
RMSLE score on the train data: 18933.2952
Select 109 features
Select 109 features
Select 109 features
Accuracy score: 94.314833%
```

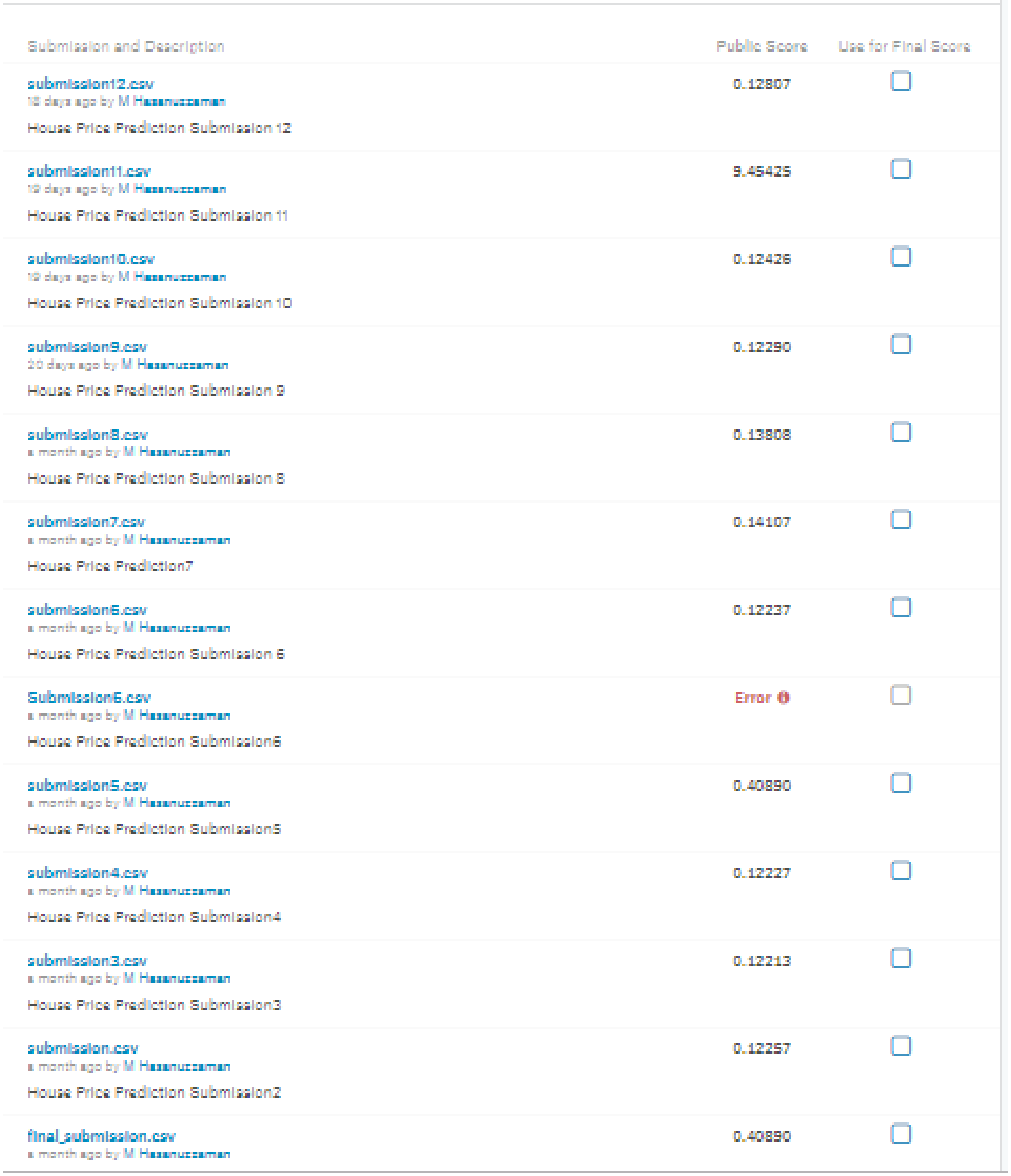
Result & Discussion:

So at the end we got this score for prediction as you can see below:



1300	hzm401	0.12213	12	18d

A list of submission given below:



Submission and Description	Public Score	Use for Final Score
submission12.csv 18 days ago by M Hasanuzzaman House Price Prediction Submission 12	0.12807	□
submission11.csv 19 days ago by M Hasanuzzaman House Price Prediction Submission 11	9.45425	□
submission10.csv 19 days ago by M Hasanuzzaman House Price Prediction Submission 10	0.12426	□
submission9.csv 20 days ago by M Hasanuzzaman House Price Prediction Submission 9	0.12290	□
submission8.csv a month ago by M Hasanuzzaman House Price Prediction Submission 8	0.13808	□
submission7.csv a month ago by M Hasanuzzaman House Price Prediction7	0.14107	□
submission6.csv a month ago by M Hasanuzzaman House Price Prediction Submission 6	0.12237	□
Submission6.csv a month ago by M Hasanuzzaman House Price Prediction Submission6	Error ❗	□
submission5.csv a month ago by M Hasanuzzaman House Price Prediction Submission5	0.40890	□
submission4.csv a month ago by M Hasanuzzaman House Price Prediction Submission4	0.12227	□
submission3.csv a month ago by M Hasanuzzaman House Price Prediction Submission3	0.12213	□
submission.csv a month ago by M Hasanuzzaman House Price Prediction Submission2	0.12257	□
final_submission.csv a month ago by M Hasanuzzaman	0.40890	□