Team Razor

Name of the Contest: House Prices: Advanced Regression Techniques

Contestants Name & ID:

1. Nazmus Shakib Shaon
   ID: 1640CSE00518
2. MD Jobayer Alam
   ID: 1640CSE00506
3. KH Rakib Ul Islam
   ID:1640CSE00509
4. Akmal Shahrioar Sheum
   ID: 1640CSE00475

Kaggle & Git Account :

Kaggle: https://www.kaggle.com/shakib0/

Git: https://github.com/shakib77/Razor

Project Report:

Introduction:

Project Goal:

In this project, we will develop and evaluate the performance and the predictive power of a model trained and tested on data collected from Kaggle train data. Once we get a good fit, we will use model to predict the monetary price of the houses in test data.

Problem statement:

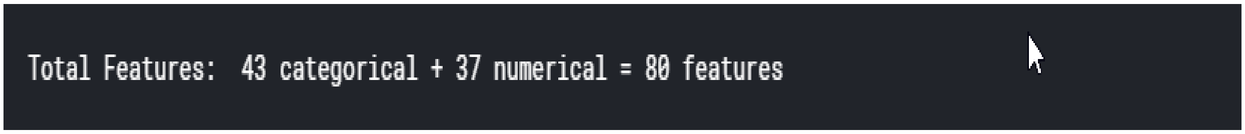
The dataset used in this project comes from the train file in Kaggle . There are 80 features of 1460 houses. We trained our model with these data for finding the sell price of test file houses.

Data Preprocessing:

Handling outlier:

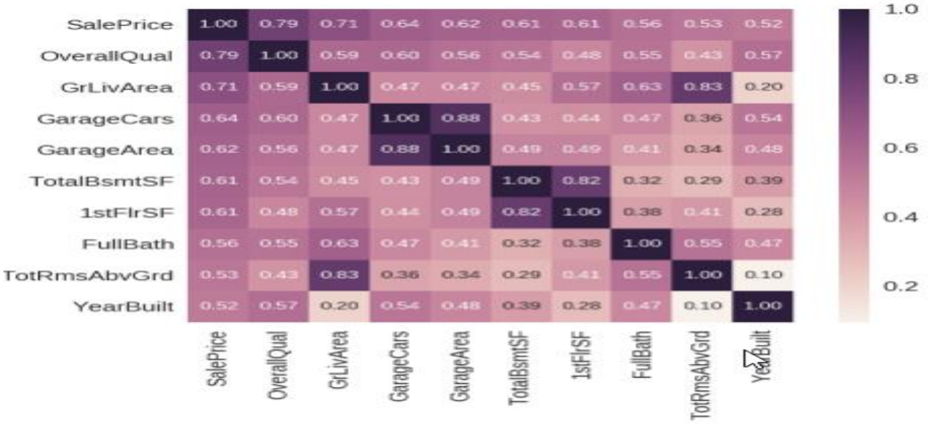
There are two types of features in housing data, categorical and numerical. Take a feature of "Downtown". The response is either "Near", "Far", "Yes", and "No". Back then, living in downtown usually meant that you couldn't afford to live in uptown. Thus, it could be implied that downtown establishments cost less to live in. However, today, that is not the case. So we can't really establish any particular order of response to be "better" or "worse" than the other.

Numerical data is data in number form these features are in a linear relationship with each other.



```
Total Features:  43 categorical + 37 numerical = 80 features
```

Top 10 Heatmap:



	SalePrice	OverallQual	GrLivArea	GarageCars	GarageArea	TotalBsmtSF	1stFlrSF	FullBath	TotRmsAbvGrd	YearBuilt
SalePrice	1.00	0.79	0.71	0.64	0.62	0.61	0.61	0.56	0.53	0.52
OverallQual	0.79	1.00	0.59	0.60	0.56	0.54	0.48	0.55	0.43	0.57
GrLivArea	0.71	0.59	1.00	0.47	0.47	0.45	0.57	0.63	0.83	0.20
GarageCars	0.64	0.60	0.47	1.00	0.88	0.43	0.44	0.47	0.36	0.54
GarageArea	0.62	0.56	0.47	0.88	1.00	0.49	0.49	0.41	0.34	0.48
TotalBsmtSF	0.61	0.54	0.45	0.43	0.49	1.00	0.82	0.32	0.29	0.39
1stFlrSF	0.61	0.48	0.57	0.44	0.49	0.82	1.00	0.38	0.41	0.28
FullBath	0.56	0.55	0.63	0.47	0.41	0.32	0.38	1.00	0.55	0.47
TotRmsAbvGrd	0.53	0.43	0.83	0.36	0.34	0.29	0.41	0.55	1.00	0.10
YearBuilt	0.52	0.57	0.20	0.54	0.48	0.39	0.28	0.47	0.10	1.00

Then we compared this features with sale price.

Most Correlated Features:

OverallQual: Rates the overall material and finish of the house (1 = Very Poor, 10 = Very Excellent)

GrLivArea: Above grade (ground) living area square feet

GarageCars: Size of garage in car capacity

GarageArea: Size of garage in square feet

TotalBsmtSF: Total square feet of basement area

1stFlrSF: First Floor square feet

FullBath: Full bathrooms above grade

TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)

YearBuilt: Original construction date

Missing data:

PoolQC : data description says NA means "No Pool"

MiscFeature : data description says NA means "no misc feature"

Alley : data description says NA means "no alley access"

Fence : data description says NA means "no fence"

FireplaceQu : data description says NA means "no fireplace"

LotFrontage : Since the area of each street connected to the house property most likely have a similar area to other houses in its neighborhood , we can fill in missing values by the median LotFrontage of the neighborhood.

GarageType, GarageFinish, GarageQual and GarageCond : Replacing missing data with "None".

GarageYrBlt, GarageArea and GarageCars : Replacing missing data with 0.

BsmtFinSF1, BsmtFinSF2, BsmtUnfSF, TotalBsmtSF, BsmtFullBath and BsmtHalfBath: Replacing missing data with 0.

BsmtQual, BsmtCond, BsmtExposure, BsmtFinType1 and BsmtFinType2 : For all these categorical basement-related features, NaN means that there isn't a basement.

MasVnrArea and MasVnrType : NA most likely means no masonry veneer for these houses. We can fill 0 for the area and None for the type.

MSZoning (The general zoning classification) : 'RL' is by far the most common value. So we can fill in missing values with 'RL'.

Utilities : For this categorical feature all records are "AllPub", except for one "NoSeWa" and 2 NA . Since the house with 'NoSewa' is in the training set, this feature won't help in predictive modelling. We can then safely remove it.

Functional : data description says NA means typical.

Electrical : It has one NA value. Since this feature has mostly 'SBrkr', we can set that for the missing value.

KitchenQual: Only one NA value, and same as Electrical, we set 'TA' (which is the most frequent) for the missing value in KitchenQual.

Exterior1st and Exterior2nd : Both Exterior 1 & 2 have only one missing value. We will just substitute in the most common string

SaleType : Fill in again with most frequent which is "WD"

MSSubClass : Na most likely means No building class. We can replace missing values with None

• Feature Engineering:

Some features that may be misinterpreted to represent something it's not.

MSSubClass: 20 1-STORY 1946 & NEWER ALL STYLES

30 1-STORY 1945 & OLDER

40 1-STORY W/FINISHED ATTIC ALL AGES

45 1-1/2 STORY - UNFINISHED ALL AGES

50 1-1/2 STORY FINISHED ALL AGES

60 2-STORY 1946 & NEWER

70 2-STORY 1945 & OLDER

75 2-1/2 STORY ALL AGES

80 SPLIT OR MULTI-LEVEL

85 SPLIT FOYER

90 DUPLEX - ALL STYLES AND AGES

120 1-STORY PUD (Planned Unit Development) - 1946 & NEWER

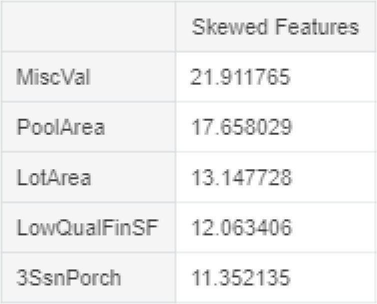
150 1-1/2 STORY PUD - ALL AGES

160 2-STORY PUD - 1946 & NEWER

180 PUD - MULTILEVEL - INCL SPLIT LEV/FOYER

190 2 FAMILY CONVERSION - ALL STYLES AND AGES

Here we fix all of the skewed data to be more normal so that our models will be more accurate when making predictions.

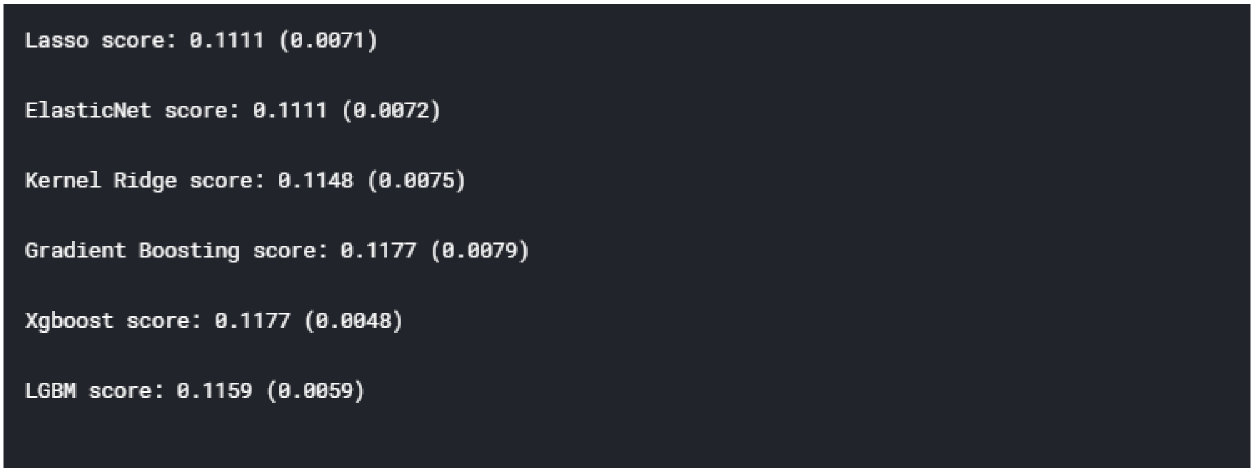


	Skewed Features
MiscVal	21.911765
PoolArea	17.658029
LotArea	13.147728
LowQualFinSF	12.063406
3SsnPorch	11.352135

We found 59 skewed numerical features.

• Modeling Methods:

cross-validation:



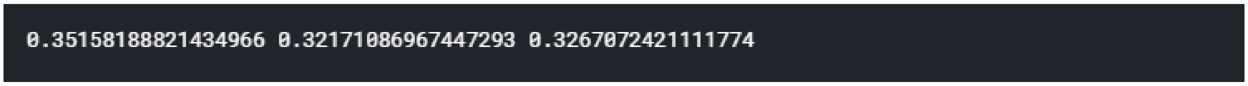
```
Lasso score: 0.1111 (0.0071)

ElasticNet score: 0.1111 (0.0072)

Kernel Ridge score: 0.1148 (0.0075)

Gradient Boosting score: 0.1177 (0.0079)

Xgboost score: 0.1177 (0.0048)

LGBM score: 0.1159 (0.0059)
```

Here, we stack the models to average their scores.

Then we average ENet, GBoost, KRR, and lasso. We added in XGBoost and LightGBM later.

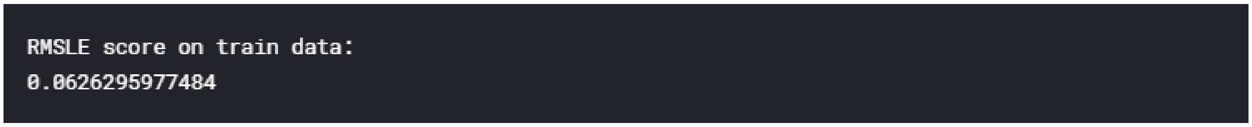
Then we'll use it as a meta-model.

Ensemble: To get our weights for each model, we took the inverse of each regressor and average it out of 100%



```
0.35158188821434966 0.32171086967447293 0.3267072421111774
```

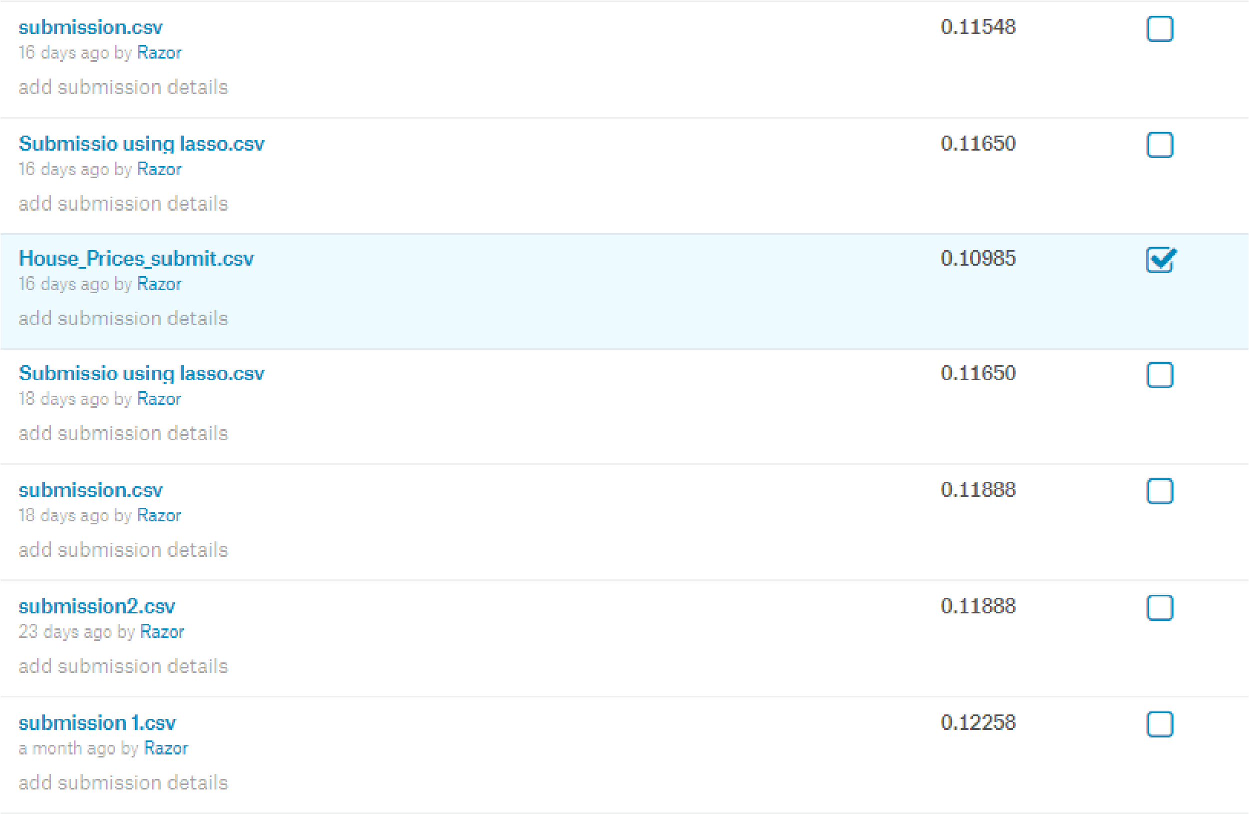
RMSE on the entire Train data when averaging.



```
RMSLE score on train data:
0.0626295977484
```

• Results & Discussion:

Here all our submission & achieved score



submission.csv 0.11548 □
16 days ago by Razor
add submission details

Submissio using lasso.csv 0.11650 □
16 days ago by Razor
add submission details

House_Prices_submit.csv 0.10985 ☑
16 days ago by Razor
add submission details

Submissio using lasso.csv 0.11650 □
18 days ago by Razor
add submission details

submission.csv 0.11888 □
18 days ago by Razor
add submission details

submission2.csv 0.11888 □
23 days ago by Razor
add submission details

submission 1.csv 0.12258 □
a month ago by Razor
add submission details