FINAL PROJECT REPORT ON

HOUSE PRICE PREDICTION USING ADVANCED REGRESSION TECHNIQUES

Under the guidance of

Prof. Navaneeth Malingan

Submitted by

Tejaswi Kanala

JULY 26, 2020

CHUBB BUSINESS LLP

Table of Contents

Chapter 1
INTRODUCTION

1.1 Introduction

In the project we need to predict sale price of the houses based on the given set of features like LotFrontage, LotShape, Alley, Utilities, FireQC, Building type, Overall quality of the house and so on.

So, In order to solve this problem statement we would be first analysing and understanding the data, preprocess the data and then applying various algorithms to see which one is performing good and able to give good interpretation to the model. Since it is a regression problem we would be using some regression techniques to solve this problem statement.

Chapter 2
LIFECYCLE OF THE PROJECT

2.1 Lifecycle of the project

The lifecycle of the project is involved in various series of steps.

- Understanding the Data
- Exploratory data analysis
    - Data Visualization
- Data Pre processing and Feature Engineering
- Model Building
    - Hyperparameter tuning
    - Principal Component Analysis
- Model Evaluation and Interpretation

Chapter 3
UNDERSTANDING THE DATA

3.1 Features in the Dataset

Our dataset has the features which are responsible to predict the target outcome which is sale of the house.

When we load the dataset we could find around 1460 rows and 81 columns and the features such as ,

MSSubclass : This feature tells about all the categories of the dwelling present

MSZoning : This feature tells us about in which zone like either residential area or commercial area or either agricultural area where the house is present

LotFrontage : This feature tells us about the whether the lot street is connected to the property or not

LotShape : This tells about the shape of the property like whether it is regular or irregular in shape.

Utilities : This feature tells about the type of utilities like all public utilities or electricity utilities etc.. available to the property

Neighborhood : This feature tells about which physical locations are available near to the property

BldngType : This feature tell about the type of the building like whether the property is for single-family or 2-family conversion, Town house end unit and so on

HouseStyle : This feature tells about the style of the property like whether it is one level or second level finished or not and so on

Overall Quality : This feature tells about the overall quality of the property, like what is the overall finishing of the house and material quality with which the house is constructed

YearBuilt : This feature tells in which the house is built

Roof style : This feature tells about the type of the roof like shed ,flat ,gambrel with which the property is built

Exterior : This feature tells with what type of exteriors the house is built like whether with plywood or stone work and all

Exterior Quality : This feature tells about the quality of the exteriors with which it built

BsmtQual : This feature tells about the basement quality of the house like to which height the basement is built and over all basement condition

Heating : This feature tells about the type of the heating like whether it is floor furnace or wall furnace or gravity furnace with which the property is built

Central air conditioning : This feature tells whether there is central Ac is there or not

Bathrooms : This feature tells about the number of bathrooms present whether it is present in basement or in other floors

Bedrooms : This feature tells about the bedrooms which are above the grade

Kitchen : This feature tells about the where the kitchen is present and also the quality of the kitchen

Fireplace quality : This feature tells about the quality of the fire place like how it is made up of and what is the quality of it

Garage built : this feature tells about in which year the garage is built

Garage Type: This feature tells about the type of the garage like whether it is attached to the property or outside of the property or there is no garage

Garage size : This feature tells about the size of the garage like what is the capacity of that garage

Pool Area : This tells about the size of the pool area in the square feet

Pool Quality : This tells about the quality of the pool

Fence quality : This tells about the fence protection around the house like whether the privacy is high or there is no fence at all

MiscFeatures : These are some of the miscellenous features like elevator, tennis court and some other features present to the property or not

Yrsold : In which year the property is sold

SaleType : This tells about the type of the sale like whether it is contract based or real estate based or it is cash based sale

SaleCondition : This tells about the condition of the sale like whether it is normal sale or abnormal sale or sale between the families etc..

So, these are the some of features present in our dataset. So, based on these different features we should use some advanced regression techniques to determine our target feature which is sale price.

Chapter 3
EXPLORATORY DATA ANALYSIS

Now we understood the features of the dataset present. We should perform exploratory data analysis to understand the relationships between our features and the Saleprice. Like how each feature is impacting on our target column.

So, at first we will make some basic data analysis so that we could know if there are any missing values present in our dataset, how numerical features are distributed like whether any transformations are required to do on that , how many categorical are present in our categorical variables, If any outliers are present and finally how these features are impacting on our target column.

3.1 Missing values

We will now try to find out any missing values are present in our dataset. To do first we will check the condition if any one nan present we will retrieve that features and then find out the percentage of missing values in those features.

OBSERVATION:

We do have missing values in our dataset they are LotFrontage ,Alley, MasVnrType, MasVnrArea,BsmtQual,BsmtCond,BsmtExposure,BsmtFinType1,BsmtFinType2,FireplaceQu,GarageType,GarageYrBlt,GarageFinish, GarageQual, GarageCond, PoolQC, Fence, MiscFeature and the highest percentage of missing values present in the columns are Alley , PoolQC, MiscFeature for these features around 96% missing values are present. Since we have missing values we will figure out the relationship between the missing values columns and target column like how it is affecting whether if missing values present the price is increasing or decreasing.

So we could find that for the features like LotFrontage, Alley, MasvnrType, Fence and Misc feature the price increases while missing values are there. So, we should replace this nan values with something meaningful.

3.2 Numerical Features

We will check out from all the features which are the numeric features present so that we could find the distributions for that.

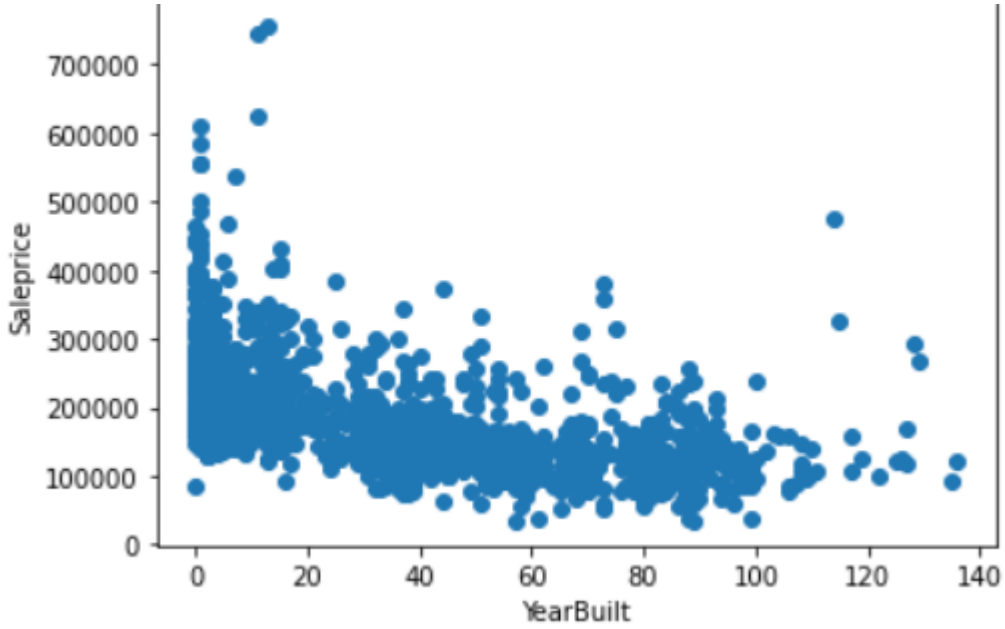
OBSERVATION:

We could see that around 38 are numeric features excluding ID because it is a useless column which is just a unique identification number for the house and there is no information with it. From the observation we could see that the yearbuilt, year sold, garage built year and year remod add are some of the years which are also considered as numeric features.

But generally years though they are in numbers we should not consider them as numeric and do some kind of transformations and all instead we could find some meaningful insight from that year with respect to the target column

After checking the impact of yrsold and the target column as the year goes on the price is decreasing.We, next find the age of the house, age of the garage build, age of like how many years the remodification for the house is done.

From the scatterplot observation we could tell that if the age of the house is more then the price sold for it is less and even same goes for the yrremod add and garage built year as well.

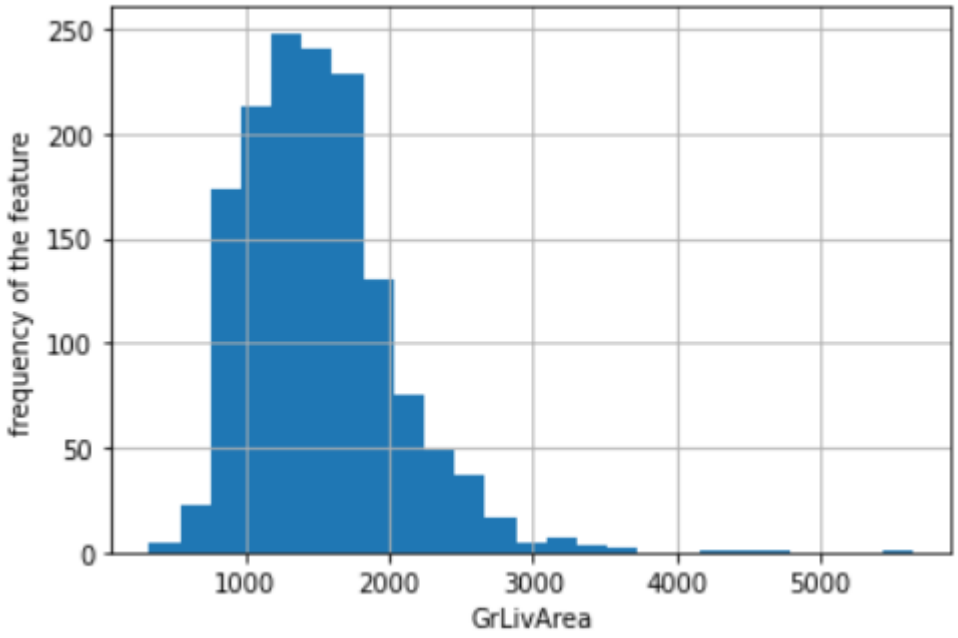


3.2.1 Continuous features

Continuous features are those whose values are in continuous range of numbers. We have 16 continuous features they are, LotFrontage, LotArea, MasVnrArea, BsmtFinSF1, BsmtFinSF2, BsmtUnfSF,TotalBsmtSF,1stFlrSF,2ndFlrSF,GrLivArea,GarageArea,WoodDeckSF,OpenPorchSF,EnclosedPorch,ScreenPorch,SalePrice and For continuous features we will see the distribution of those features. If there is any skewness in the distribution then we must apply some transformations and make it into a gaussian or normal distribution.

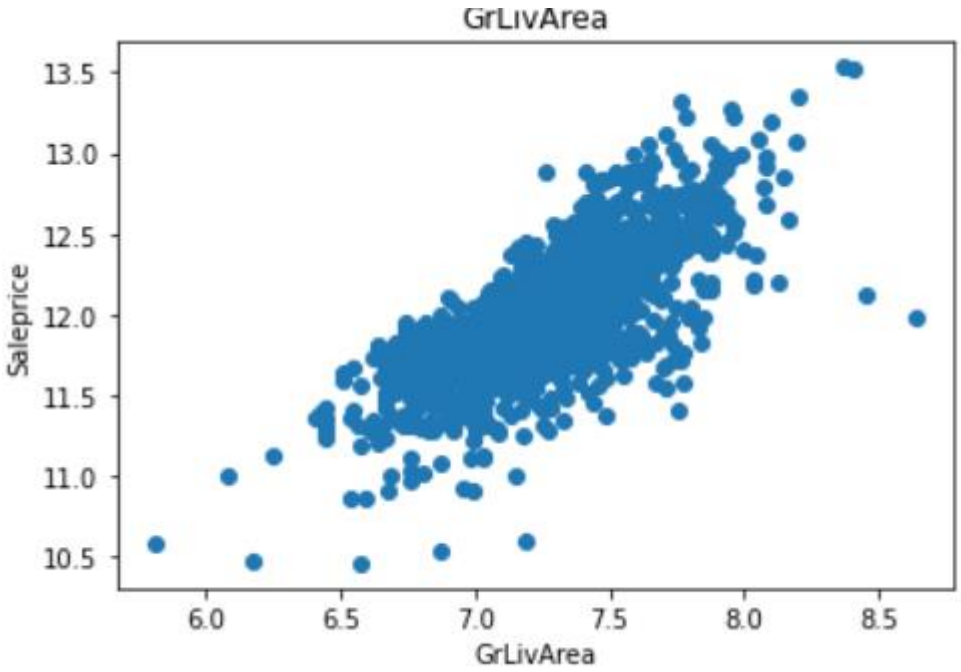
OBSERVATION:

Most the features looked like they are not following the gaussian distribution. From all the features only the sale price a bit is in normal distribution. Remaining features all are either right or left skewed.



So, we will now find the impact of these continuous variable on the sale price. Before that we apply log transform on the features and see impact with respect to target column. We don't apply transform to the columns which has zero because log(0) is not defined.

From the scatterplot we could see that there is a positive correlation for the features like GrLivArea and 1stFlrSF with respect to target variable and moderate correlation for the features like LotArea and LotFrontage.



3.2.2 Discrete features

Discrete features are the type of numerical features but which has the certain range of category. So, in our dataset we filter out the discrete features by using some condition like retrieving the features that has category less than 25.

OBSERVATION:

So, we have got around 17 are the discrete features. They are MSSubClass, OverallQual ,OverallCond, LowQualFinSF, BsmtFullBath, BsmtHalfBath, FullBath, HalfBath, BedroomAbvGr, KitchenAbvGr, TotRmsAbvGrd, Fireplaces, GarageCars, 3SsnPorch, PoolArea, MiscVal, MoSold .

Let's us see how it is affecting the target column that is the sale price.

It is found that :

If the overall quality is high then the sale price is also increased.

For 2-STORY 1946 & NEWER MSsubclass the price is high

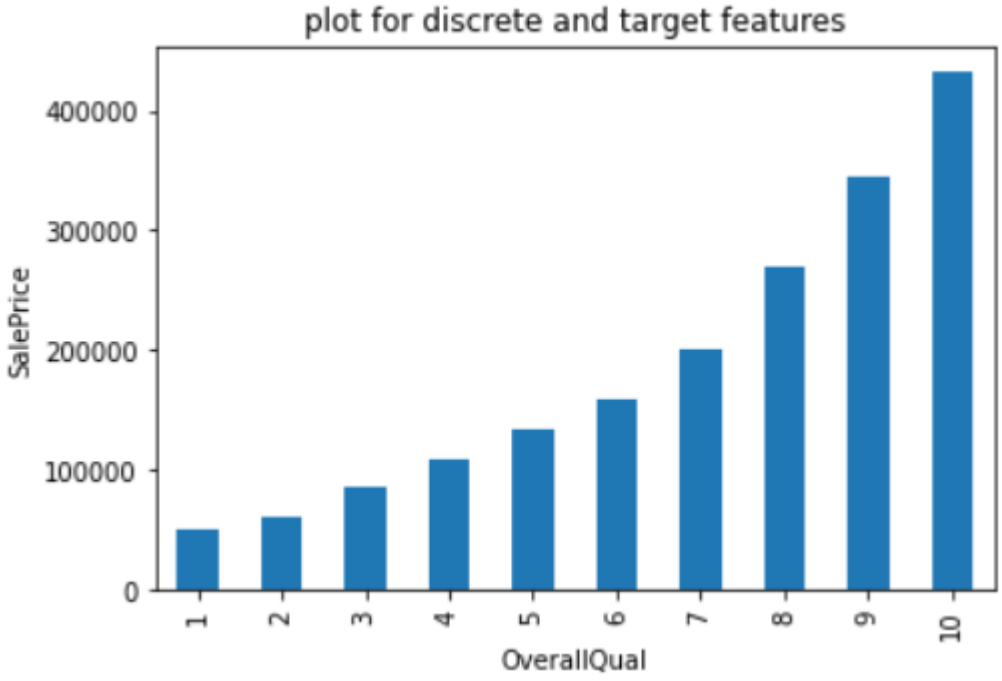
If the overall condition is average and excellent then the price is high

If the bathrooms are full on above grade then the price is high

If garage cars capacity is medium range (3) then the price is high

For pool area (555) the price is more

In 9th month the sales are more that means the sale price is more



3.3 Categorical Features

Now, in our dataset we will find out the number of categorical features present. We filter it out based on the datatype 'O'. So, features whose datatype is object those are categorical features.

So, there around 43 features which are categorical features and we will see how many categories are present in each category. The categorical features are MSZoning, Street, Alley, LotShape,LandContour,Utilities,LotConfig,LandSlope,Neighborhood,Condition1,Condition2 ,BldgType, HouseStyle, RoofStyle, RoofMatl, Exterior1st, Exterior2nd, MasVnrType, ExterQual, ExterCond, Foundation, BsmtQual, BsmtCond, BsmtExposure, BsmtFinType1, BsmtFinType2,Heating,HeatingQC,CentralAir,Electrical,KitchenQual,Functional,FireplaceQ u, GarageType, GarageFinish,GarageQual, GarageCond, PavedDrive, PoolQC, Fence, MiscFeature, SaleType, SaleCondition.

OBSERVATION:

From the observation we have found that for the features Neighborhood there are around 25 different categories, for exterior1st there are around 15 categories, for exterior2nd there are 16 categories and for the sale type there are 9 different categories. And for remaining features there are limited number of categories.

Now, let us find the impact of these categorical features on the target column sale price.

It is found that,

For MSZoning 'Floating Village Residential' zone the sale price is high

For Street and Alley 'pave' has the high sale price

For LotShape 'Moderately irregular' (IR2) has high sale price

For LandCountour 'Hillside' (HLS) has high sale price

For Utilities 'All public utilities' available has the high sale price

For LotConfig 'Cul-de-sac' has high sale price

For Landslope Moderate Slope and Severe Slope shares almost equal price

For Neighborhood 'Northridge Heights' near to this physical location has high price

For BldngType '1Fam and Twnse' has shared equal cost price

For Roofstyle 'shed' has high price

For roof material 'Woodshngl' has high price

For MasVnrTye 'stone' work has the high price

For exterior if the quality is 'excellent' then the price is high

For foundation 'Pconc' has the high price

If the Basement quality is excellent then the price is high

If the basement condition is good then the price is high

If the basement exposure is good then the price is high

If there is central AC to property the price is high

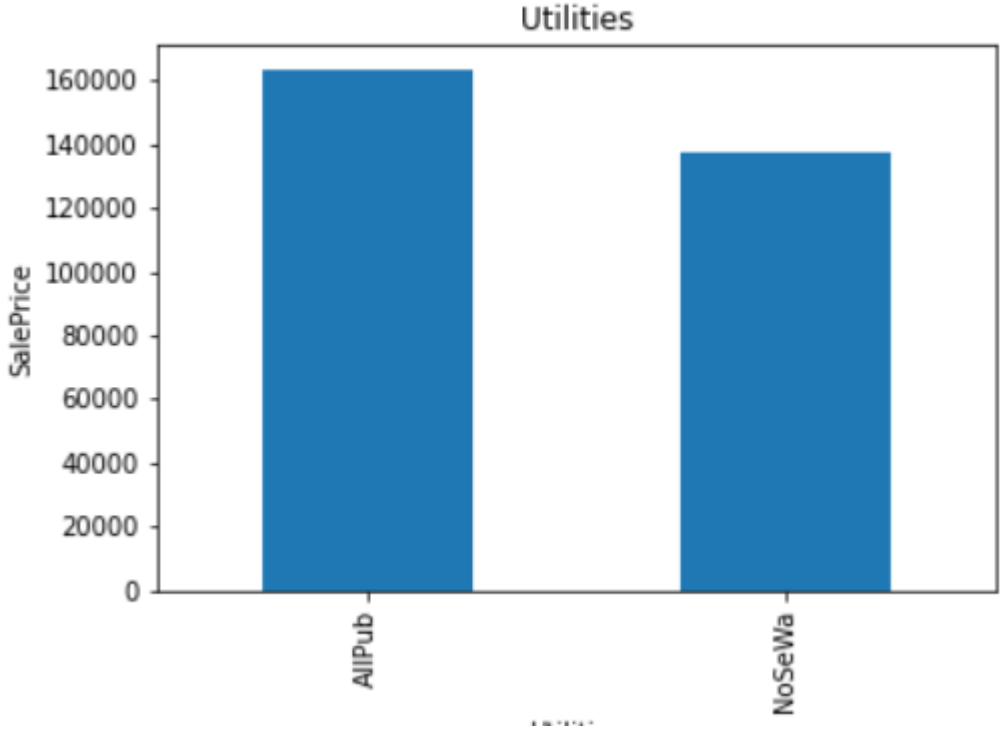
If the kitchen quality is excellent then the price is high

If the fireplace Qu is excellent then the price is high

If the fence is highly protective then the price is high

If the there is tennis court to the property then the price is high

If the sale condition is partial then the price is high.



So, these are the categorical features observations where those are impacting on the target variable salesprice.

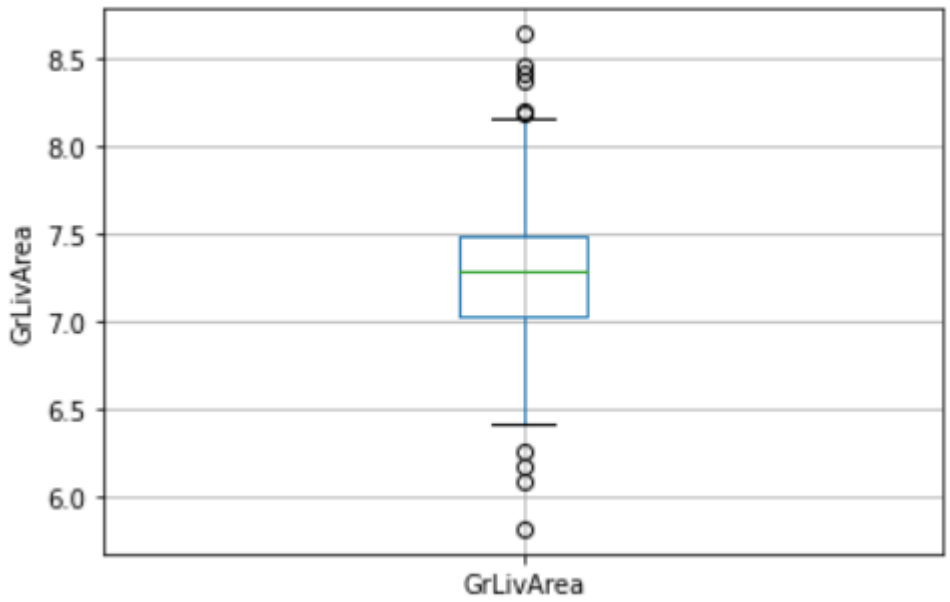
3.4 Outliers

We have now to see whether our data has any outliers present. It's good to identify the outliers.

For, finding the outliers we would be using some boxplot analysis

OBSERVATION:

We could see that there are outliers for the features like LotFrontage, LotArea, 1stFlrSF, GrLivArea.



Chapter 4
DATA PREPROCESSING AND FEATURE ENGINEERING

Till now we made a detailed exploratory data analysis and we found out some of the interesting results. So, to build our model which could able to predict the house prices we must preprocess the data so that our ML model could give some fruitful results.

4.1 Handling missing values

As, in the exploratory data analysis we found there are missing values present both in the categorical and numerical features. So, we should fill these missing values with some useful values.

For, categorical features it is a good practise to replace the missing values with some new label. So, here I have replaced the categorical feature with ‘missing’ label.

For numerical features before filling the missing values we should check if there are any outliers present, if outliers are present then we could replace the missing values using the median. So, I have replaced the missing values in numerical features using median and created a new column stating that if there is missing value then 1 else 0.

Next we move on to the four different date features. As we saw in the EDA as the Year sold is increasing then there is decrease in the price. Generally if it is sold as latest it must increase in the price but here it is decreasing. So, we would consider other factors.

Making a comparisonand extracting the years for yrbuilt, yrremod and garagebuilt with yrsold and so now in the place of years we have the ages for house built, house remodelled and garage built.

4.2 Log norm transformation

As, in EDA we have seen about the distribution of the numerical features. Many of the numeric features are skewed. So, we should make it to a gaussian distribution for that we do the log normal transformation.

So, the transformed features are LotFrontage, LotArea, 1stFlrSF, GrLivArea, SalePrice.

4.3 Label encoding for categorical features

For categorical features we will do label encoding because before feeding our data to any ML model we should convert everything to numbers. Before that we will eliminate some features whose categories are very less and replace it with 'rarlabel'. We set threshold limit of 0.01 and replace it.

Finally we have label encoded the categorical features.

4.4 Feature Scaling for numerical features

We do feature scaling for the numeric features because if we have the numeric features of different ranges then it would be difficult for the model for function approximation. So, in order to avoid that we do feature scaling using MinMax scaler and scale down the values to the range of 0-1. So, atlast we do feature scaling for all the numerical features.

After doing these all preprocessing and feature engineering then we would be storing this into a new csv file which we can use that data for further model building, evaluation and interpretation.

Chapter 5
MODEL BUILDING AND OPTIMIZATION

Now our data is processed and ready to build the model which could able to predict our target variable SalePrice. Since it is a regression problem we could use some of the advanced regression techniques like Lasso regression, XGBoost regressor , RandomForest regressor and some dimensionality reduction techniques like PCA.

I have first used the base models to train the data and later did some hyperparameter tuning to see whether it is yielding any better results or not.

The metrics that are considered while building the model are:

1) **RMSE** : Root mean square error is a frequently used measure of the differences between values predicted by a model or an estimator and the values observed.
   It also tells about how much we are far from the actual value. It is one the best metric considered for the regression problems

2) **R2 score** : It is the measure of goodness fit measure for the regression models. It is statistic measure indicates the percentage of the variance in the dependent variable that the independent variables explain collectively. But mainly I have considered RMSE.

Let's move on to the models and see how each model performed it's job

5.1 Lasso Regression

This is the type of the regression that uses shrinkage. It performs both feature selection and regularisation that prevents overfitting. It does the feature selection but not considering the variables whose coefficient is zero. The regularisation parameter (lambda) is tuned by hyperparameter optimization. I did select features using lasso regression select from model.

So, when lasso regression is applied on the data containing 84 columns it selected only 41 features so it literally didn't consider rest of the 43 features.

It gave a pretty good score with good RMSE value. We will discuss the model interpretation and evaluation in the next section. So, this is the good advantage of lasso regression it did shrink the model and try to prevent the overfitting of the model. On top of that the main thing is it uses the regularisation parameter lambda (0.001)

5.2 Random Forest Regressor

It is a bagging ensemble technique that has multiple decision trees which aggregate the results and produce the outcome which is the average of all the models that are used in that. I did use the default random forest regressor considering all the features and using the random forest by tuning the hyperparameters using grid search.

The hyperparameters that are tuned using grid search with folds 5 are:

n_estimators : the number of trees required

min_sample_leaf : minimum samples required at the tree

max_depth : It is the depth of tree till if the depth is too high then there is chance of overfit

The results and evaluation will be discussing in the next section. Last I have generated feature importance plot to visualize the top important features that this algorithm considering for predicting the target outcome. I got RMSE score a better one but this complete model with and with out hyperparamter tuning slightly seems to be overfit.

5.3 XGBoost Regressor

XGBoost comes under the boosting technique which basically learns from the weak learners. Xgboost is used because of its high execution speed and for good model performance. It train on a weak model and the one's which are wrongly predicted are sent to next model which are in sequential and until it get good accuracy it keep on improving the weak model.I did use xgboost by considering all the features and also tuned the hyperparameters by using randomized search cv. The results we will be discussing in next section.

The hyperparameters considered for building this model are:

n_estimators : the number of trees required for model

max_depth : It is used to control overfit and learn relation for one particular sample

booster : It select which type of model to run in iteration whether gbtree or gblinear. In our case it chooses gbtree.

learning_rate : Making model robust by shrinking the weights

min_child_weight : defines minimum sum of the weights required in child. If value is higher leads to underfit. So if values are high then it prevent model from learning the relations

gamma : this parameter used it takes split which gives positive reduction in loss function.It makes algorithm conservative.

At last even I have plotted the top 10 important features considered by this model for predicting the target outcome. After that even I have tried fitting the model with the top 10 features as well. Of all these models XGBoost gives me the best RMSE score with good prediction results as well.

5.4 PCA (dimensionality reduction)

PCA is nothing but the principal component analysis which is used to reduce the dimension of the features without losing much information from them. Since we have 84 columns I just tried PCA to reduce dimension for the features. I applied PCA on dataset containing 84 features and compressed the dimension to 50 by retaining the 95% of the total variance. After that I trained XGBoost regressor and Random Forest regressor by hyperparameter tuning using randomised search cv on that features whose dimension is reduced to 50 features. The results will be discussed in the next section.

Chapter 6
MODEL EVALUATION AND INTERPRETATION

Now we will see about the model evaluation and interpretation and to see which model performed well to our data.

For all the models, we considered

We considered ‘y’ dependent variable as our SalePrice

And ‘x’ to be our independent features which drops the ID column and target column.

6.1 Lasso model evaluation

Now we will see how lasso model performed on our data.

Now lasso is fitted to the model with lambda (0.001) and retrieved the selected 41 features and then with those selected features I have trained the model

We got,

Test RMSE : 0.14 and the invert score of log transform is : RMSE: 37983

Train RMSE : 0.11

So, from this lasso model we could interpret that our model is predicting good and performed with rmse of 37983.

6.2 Random Forest regressor evaluation

Now we will see how random forest regressor performed on our data

First default random forest tuned on all the features.

We got,

Test RMSE : 0.13

Train RMSE : 0.05

Second random forest is tuned by using grid search CV with cv of 5 folds

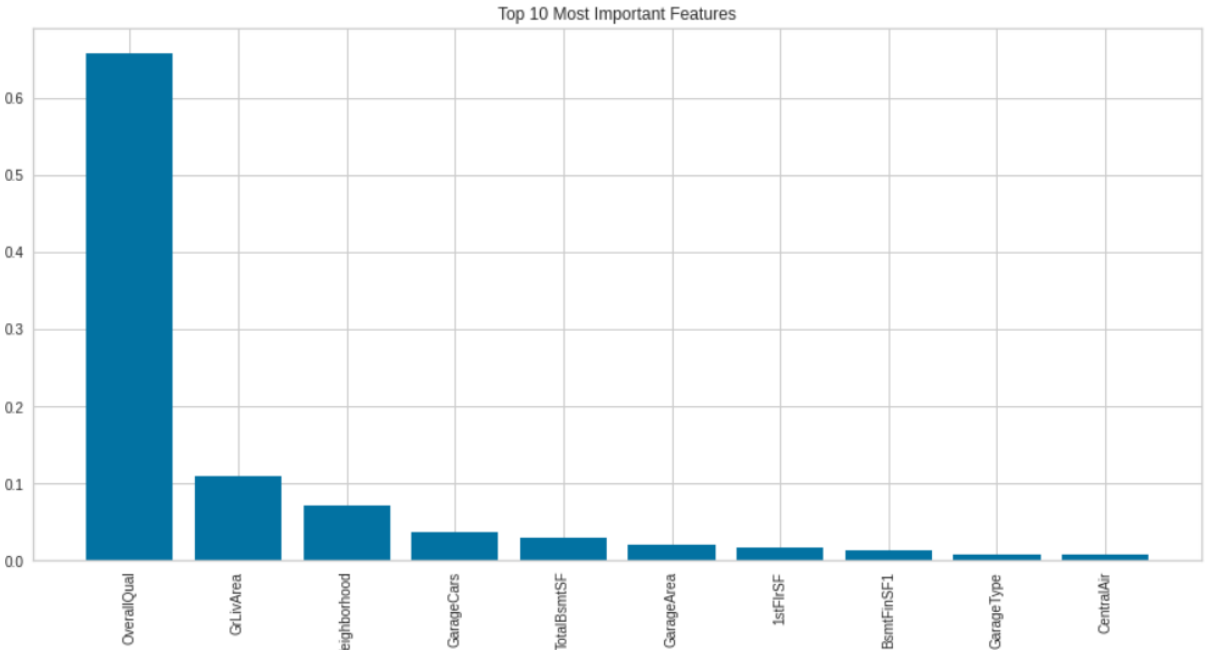
We got,

Test RMSE : 0.15

Train RMSE : 0.05

We could see that random forest is overfitting in both the cases. This maybe because we have around 43 categorical variables and 2 to 3 categorical features has around 15-16 categories. So, maybe because of this random forests are favour in those attributes with more levels. So, it is not suitable for this type of data

The top 10 features that random forest considered are:



6.3 PCA evaluation

I have used this dimensionality reduction technique and reduced the features 50 and with 95% retain in the variance.

First applied XGBoost regressor with hyper parameter tuned using randomised search CV

We got,

Test RMSE : 0.20 and the invert score of log transform is RMSE: 43871

Train RMSE : 0.14

Next applied random forest regressor with hyper paramter tuned using randomised search

We got,

Test RMSE : 0.18

Train RMSE : 0.06

Here, we can tell that xgboost is performing well a bit when PCA is applied. Random forest is overfitting in this case as well.

So, here we choose xgboost as the best with RMSE of 43871

6.4 XGBoost regressor evaluation

Now we will see how xgboost performed on our data

First applied default Xgboost with all features

Test RMSE : 0.13

Train RMSE : 0.08

Next applied xgboost with hyperparameter tuning

Test RMSE : 0.14

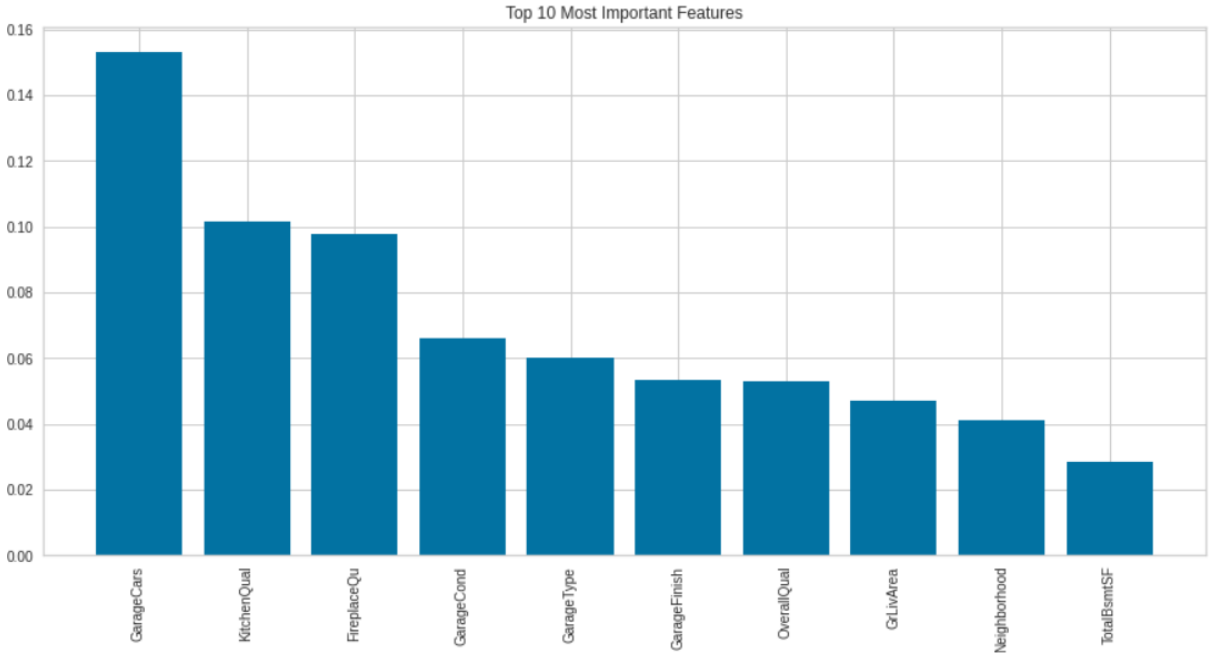
Train RMSE : 0.10

Applied xgboost algorithm with selected features

Test RMSE : 0.14

Train RMSE : 0.12

The top 10 features that xgboost considered as:



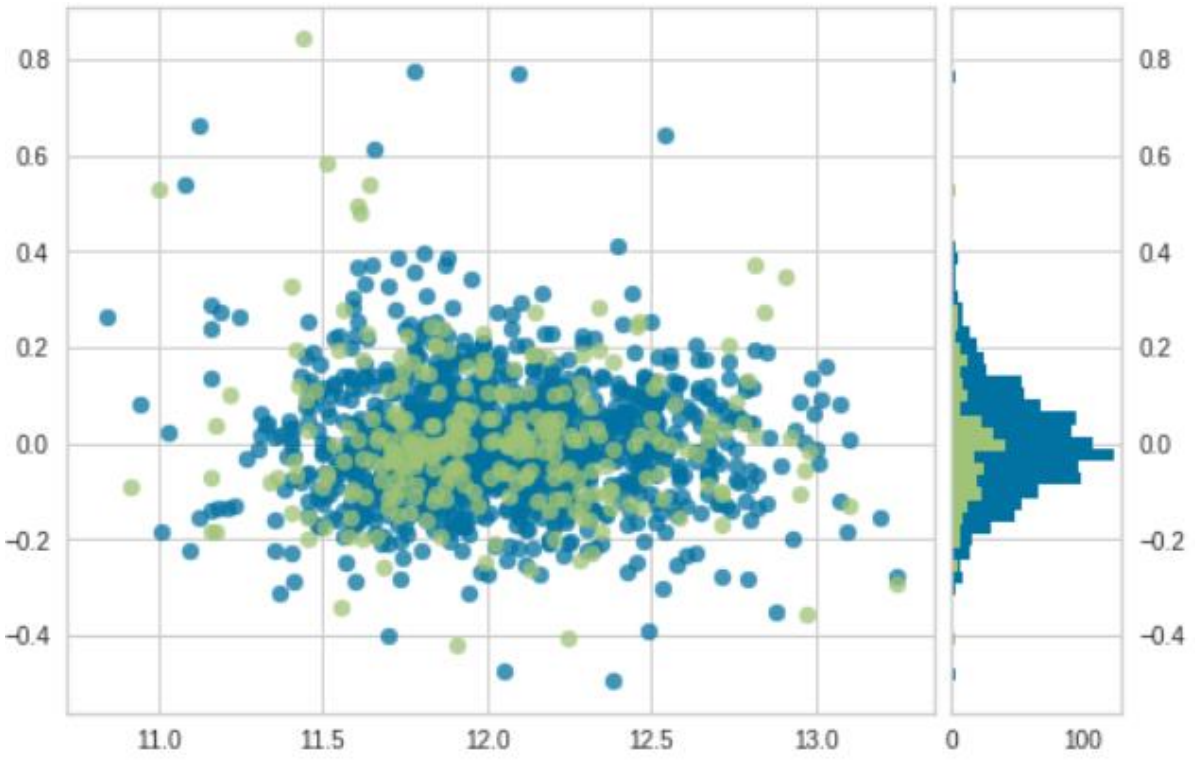
As, we can see base xgboost with all features are overfitting. Next when it is tuned it could give some better result , but when it is trained with selected features then we got a very good RMSE value and it is predicting good.

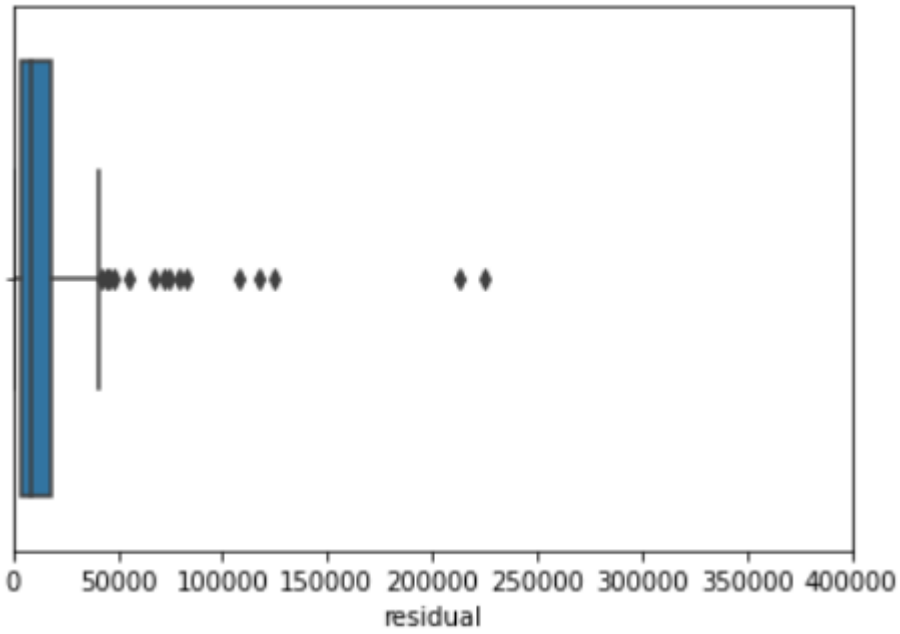
So, of all the models I could tell,

6.5 Final consideration

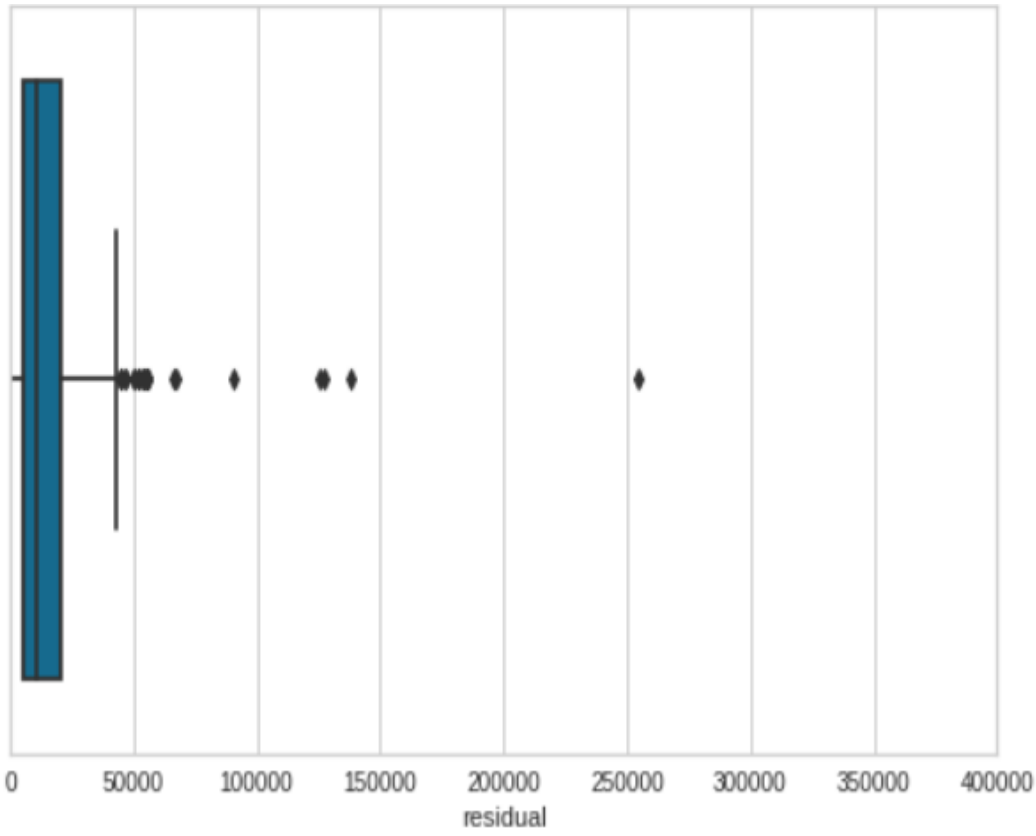
First XGboost with selected features performs well with RMSE : 0.14

It is very much closer to zero. Xgboost performed well because the core idea of this algorithm is to train on residuals so it learn on weak learners


0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
11.0
11.5
12.0
12.5
13.0
0
100

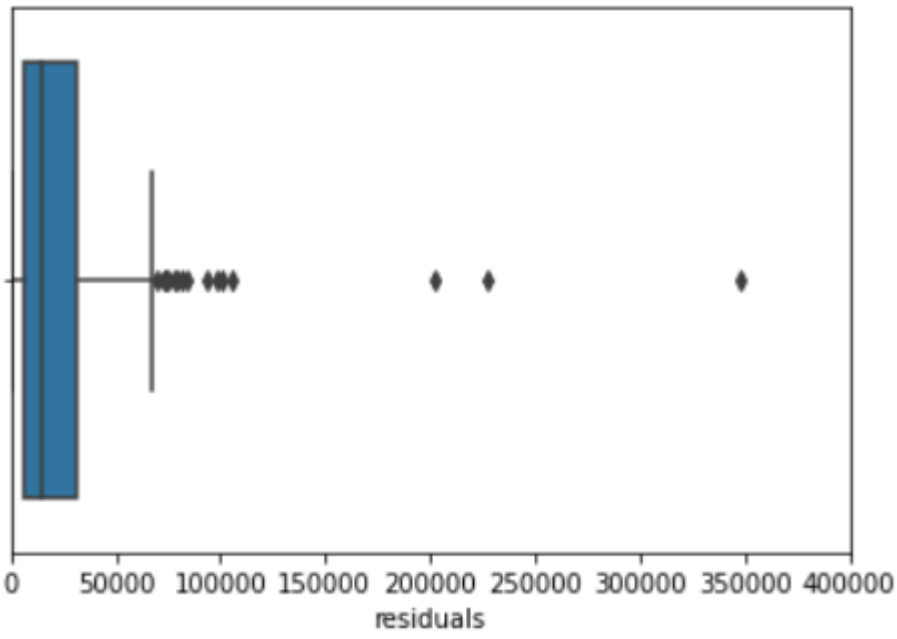

0
50000
100000
150000
200000
250000
300000
350000
400000
residual

Second comes the Lasso regression with RMSE of 0.14



Here it is a way closer to zero but not more than the first model.

Third comes a bit better xgboost with PCA applied with RMSE of 0.20.



Algorithm	Test RMSE	Train RMSE
Xgboost with selected features	28361$ (0.14)	24298$ (0.12)
Lasso regression	37983$ (0.14)	25445$ (0.11)
PCA with XGBoost	43871$ (0.20)	31955$ (0.14)