Data Mining and Web Analytics

Candidate Id -210216123

Table of Contents

INTRODUCTION

The report focuses on real estate dataset which talks about different features of the houses that effects the sales price of the house.

The problem underlies in the real estate is determining the price of the property and based on which factors new property sale price can be estimated. The report analysed the different factors which will be impacting the sale price and based on these factors similar property price can be estimated.

The main stakeholders who can benefit from the analysis are Real estate Agent, property Buyers, and the property seller.

Because real estate agents function as a mediator between property buyers and sellers, predictive modelling will assist them in obtaining a range of pricing in various places depending on their attributes, allowing them to assist buyers according to their needs and desired qualities in a home.

Predictive modelling will also benefit property buyers since it will allow them to acquire an estimate of the price of the property they are going to buy in the future, so they will not be disappointed if they find out the property, they bought was expensive.

Similarly, the study will help the property seller because they will be able to acquire an evaluation of their property, as there is a potential that the real estate agent's property price estimation might be incorrect. However, bearing in mind all of the aspects that influence the price of a property will help the seller. Total 10 features or predictors are considered which affect the sale price of the property.

INITIAL EDA

Dataset Type & Dimension -Dataset we have is cross-sectional in nature comprising 1144 records and 79 attributes.

Based on initial modelling and correlation between predictors and target variable, 10 traits were chosen from 79 attributes, and our data looks like this:

MSZoning	LotShape	OverallQual	ExterQual	TotalBsmtSF	1stFlrSF	GrLivArea	TotRmsAbvGrd	GarageType	GarageArea	SalePrice
RL	Reg	7	Gd	856	856	1710	8	Attchd	548	208500
RL	Reg	6	TA	1262	1262	1262	6	Attchd	460	181500
RL	IR1	7	Gd	920	920	1786	6	Attchd	608	223500
RL	IR1	8	Gd	1145	1145	2198	9	Attchd	836	250000

Variables, definitions, their types, and their roles:

The sorts of variables we chose and their roles, as well as their descriptions, are clearly visible in the following presentation.

S.No	Variable	Variable Type	Values	Description	Role
1	**MSZoning**	Nominal	C(all),FV,RH,RL,RM	Determines the sale's general zoning classification (commercial, Floatig village residnetial, Residential high ,low & mdeium density)	Predictor
2	**LotShape**	Nominal	IR1,IR2,IR3,Reg	The feature describes the property's overall shape (Regular, slightly ,moderately irregular)	Predictor
3	**OverallQual**	Ordinal	0-10	Provides a rating for the house's overall material and finish (1 poor & 10 excellent)	Predictor
4	**ExterQual**	Nominal	EX,FA,GD,TA	Assesses the exterior material quality (Excellentm, fair, good, & average)	Predictor
5	**TotalBsmtSF**	Continuous	0-3206	The variables tells us the total basement area in square feet	Predictor
6	**1stFlrSF**	Continuous	334-3228	It tells us about the total 1st floor are in square feet	Predictor
7	**GrLivArea**	Continuous	334-4316	The variable tells us about the above ground living area in square feet	Predictor
8	**TotRmsAbvGrd**	Nominal	2-14	It defines the total number of rooms in a property above the ground	Predictor
9	**GarageType**	Nominal	2Types, Attchd, Basement, Builtin, CarPort, Detchd, NA	The variable indicates the type of garage or the placement of the garage in relation to the property.	Predictor
10	**GarageArea**	Continuous	0-1390	Tells us about the area or size of the garage in square feet	Predictor
11	**SalePrice**	Continuous	34900-755000	selling price of the property	Target

Level of the data

Our dataset is distributed on 5 levels of zone using MSZoning attribute. Levels are C(all)-commercial, FV-floating village residential, RH-residential high density, RL-residential low density, and RM- residential medium density.

Uni-variate Visualization & Bi-variate Visualization

1) MSZoning-In fig 1.1 we can see the distribution of all the categories of MSZoning, the bar chart tells us that the count of Residential Low Density (RL) is much higher in comparison to other MSZoning categories also this signifies class imbalance.

fig 1.2 shows relation between MSZoning (predictor) and sale price(target). RL category has almost similar mean whereas RL covers wide range in terms of prices. Some outliers can be observed under RL & RM category.

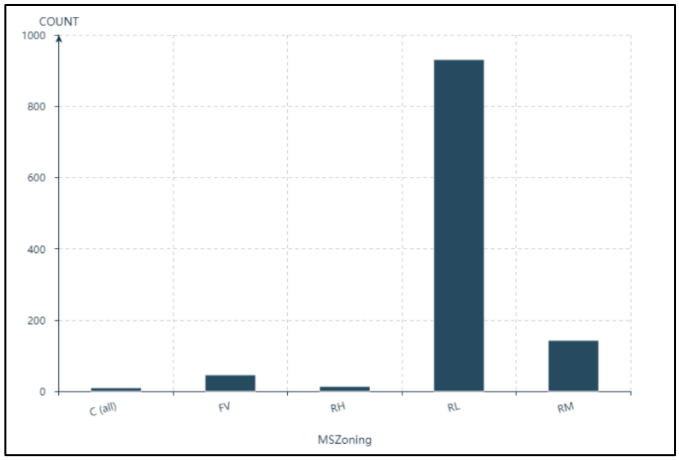


Figure 1.1 Univariate Visualization-MSZoning

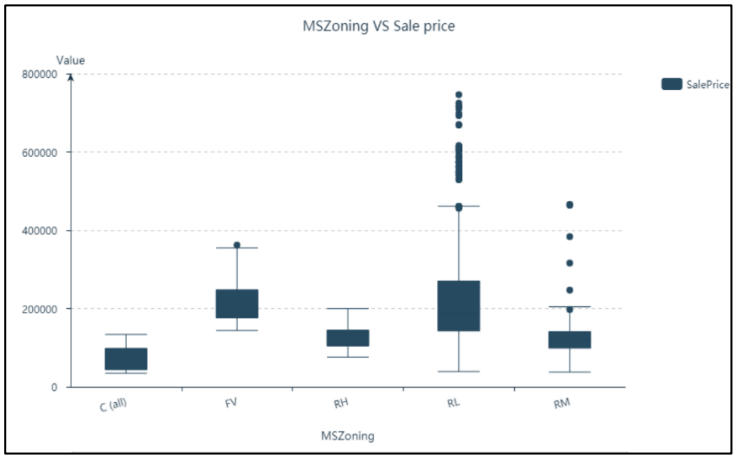


Figure 1.2 Bivariate Visualization-MSZoning vs Sale price

2) LotShape-In fig 1.3 we can see that count of regular lot shape is higher than other categories followed by irregular 1 category. From fig 1.4 average sale price for lot shape type IR3 is higher than other shape which seems surprising as in general regular lot shape property sale price should be high, this may be because of the presence of outliers in the data. IR1 and Reg categories contains some values which are away from mean and quartiles range. We can see some irregular shapes of Lot Shape at high prices.

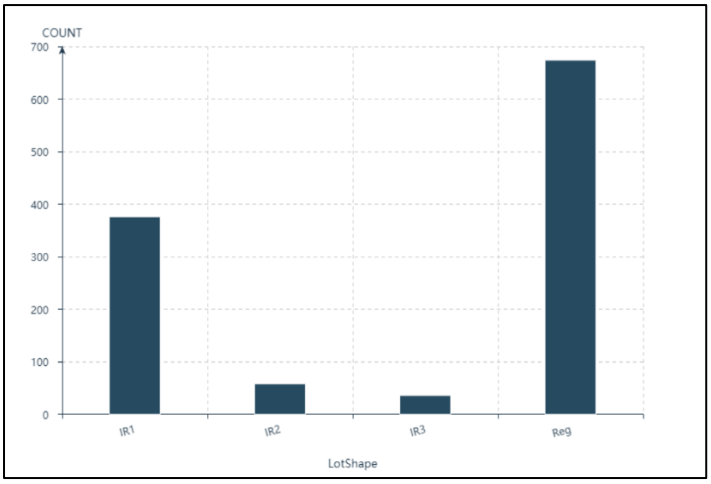


Figure 1.3 Univariate Visualization Lot Shape

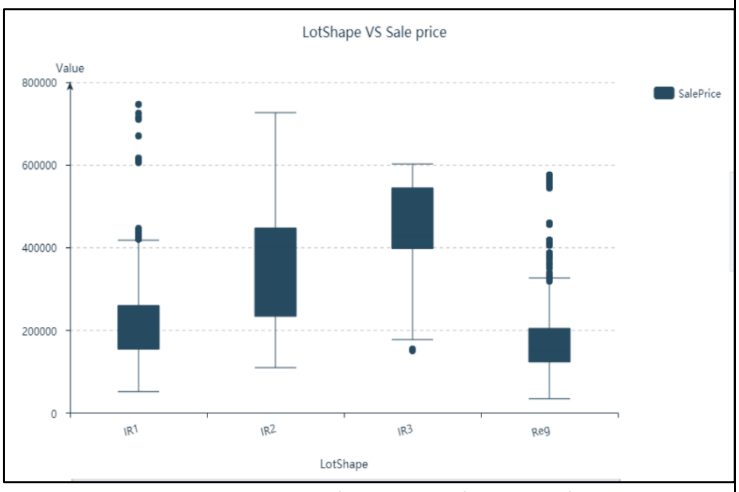


Figure 1.4 Bivariate Visualization Lot shape vs Sale price

3) OverallQual-Considering it as a nominal category the values are not presented in correct form in the dataset due to which it is required to convert them to integers. In fig 1.6 we can see the OverallQual visualization after converting them to integer. We can clearly see that OverallQual is rated 5 or 6 (average and above average) more than other categories.

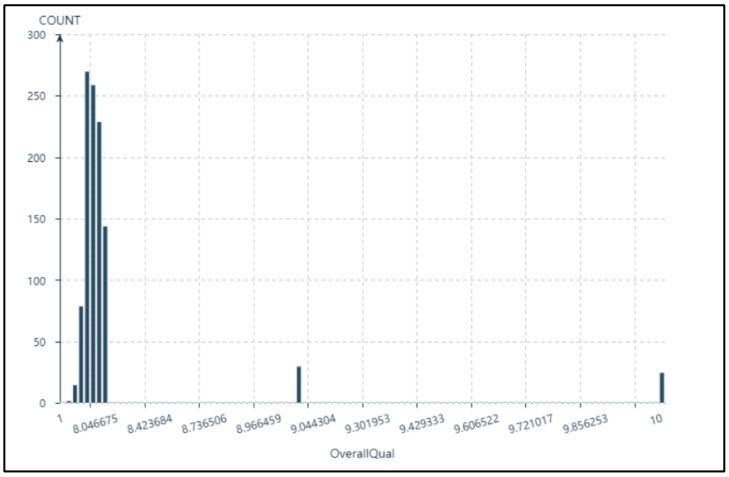


Figure 1.5 OverallQual Univariate Visualization

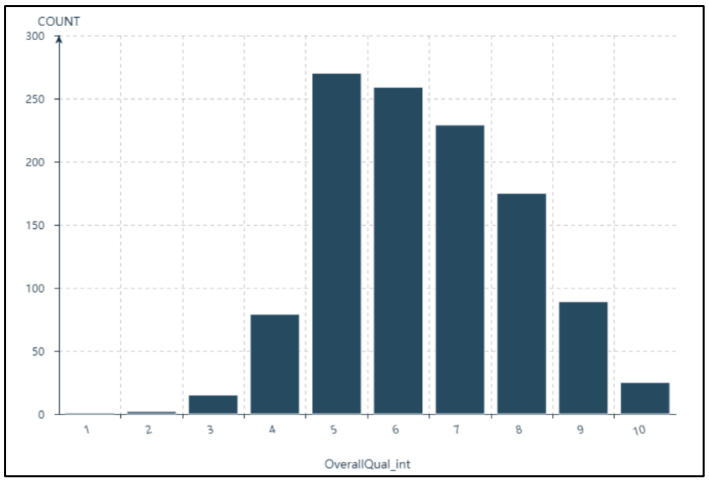


Figure 1.6 Univariate Visualization after converting to integer

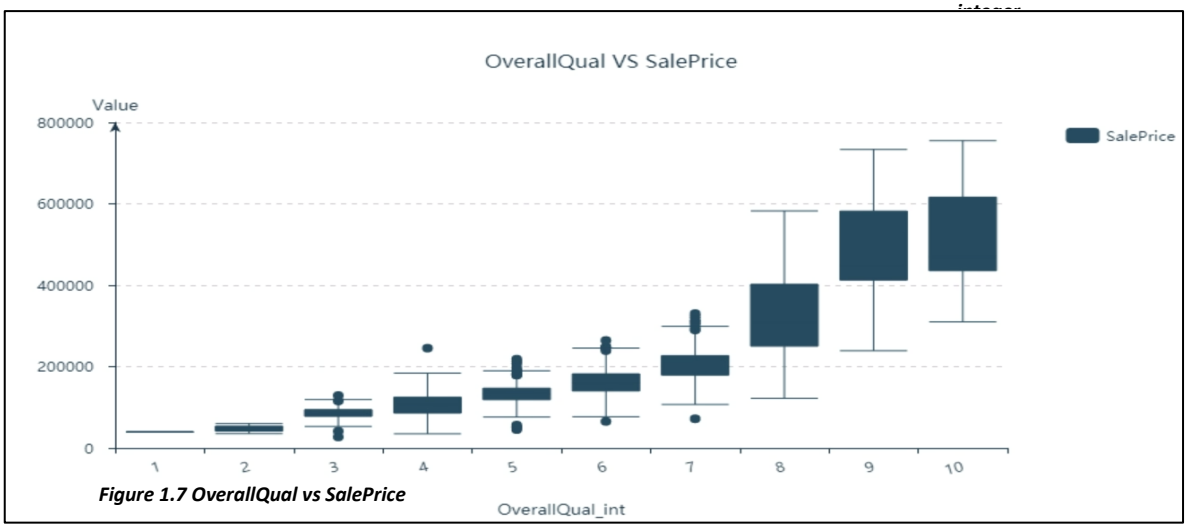


Figure 1.7 OverallQual vs SalePrice

Fig 1,7 shows the bi-variate visualization between sales price and Overall quality of the house. We can see the average price of houses rated 10 is greater than other ratings, here rating 10 refers to very excellent, this is practical as the houses which excellent in quality will have higher price than houses with average quality

4) ExterQual-Fig 1.8 shows the bar plot for exterior quality, from fig we can see number of average quality houses are much more than other houses which either have excellent quality, fair or good quality. Also, from fig 1.9 we can see the box plot showing that average sale price of excellent quality houses is higher than for houses with good or average quality.

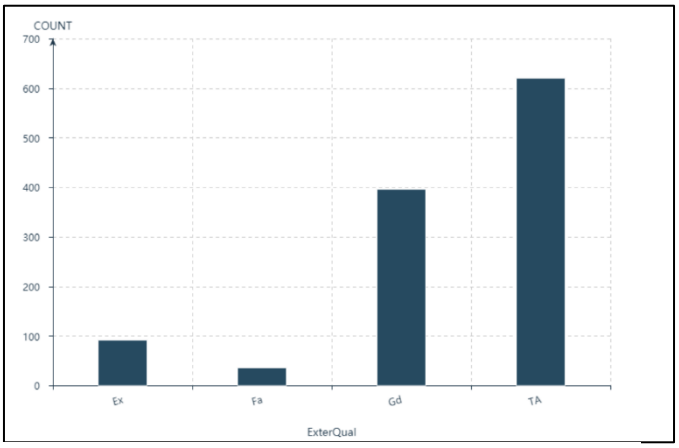


Figure 1.8 Univariate Visualization of Exterior Quality

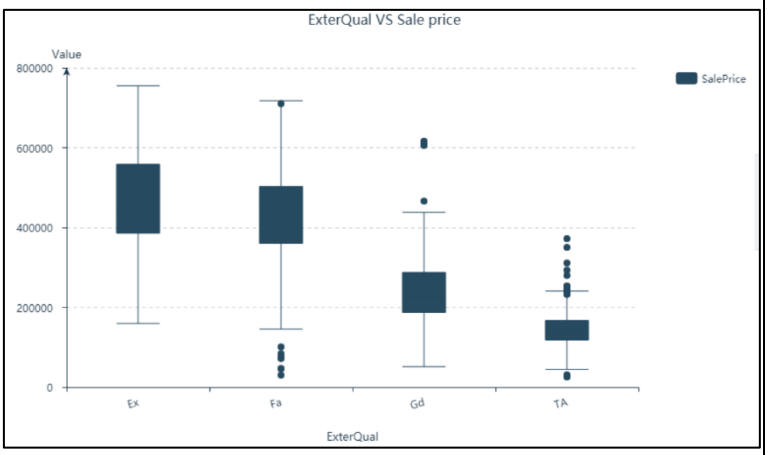


Figure 1.9 Bivariate Visualization Between ExterQual vs SalePrice

5) TotalBsmtSF-Fig 1.10 shows the distribution of the variable, we can see that TotalBsmtSF is not normally distributed and is right skewed, also if we see bivariate visualization between TotalBsmtSF and SalePrice we can observe that the sale price is increasing with the increase in total basement are in square feet. Also, presence of extreme can be observed a its too far from other records, at ~6000sqft.

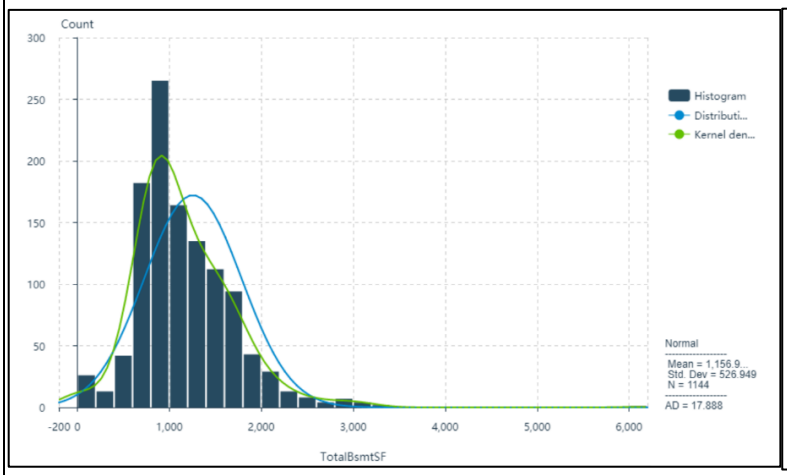


Figure 1.10 Histogram for TotalBsmtSF

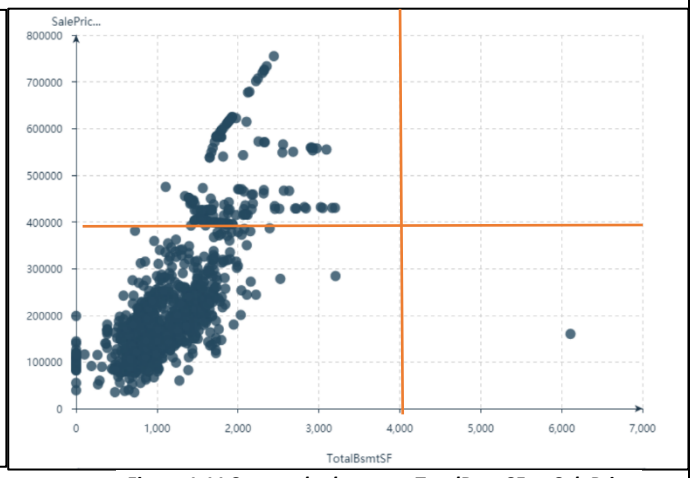


Figure 1.11 Scatterplot between TotalBsmtSF vs SalePrice

6) 1stFlrSF-We can see from fig 1.12 that variable 1stFlrSF is right skewed and is not normally distributed also its distribution is fitting kernel distribution which confirms skewness. From fig 1.13 scatterplot shows that sale price increases with increase in area of 1stFlr square feet. Under right bottom quarter there is one irregularity can be observed at ~4600 might require elimination of this value.

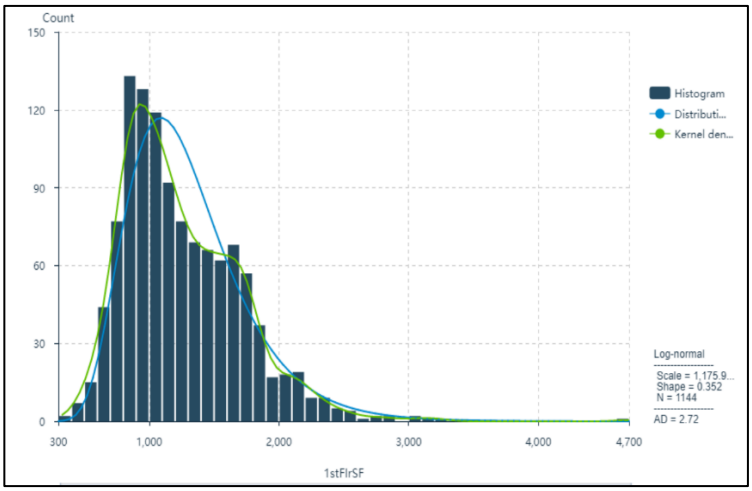


Figure 1.12 Histogram for 1stFlrSF

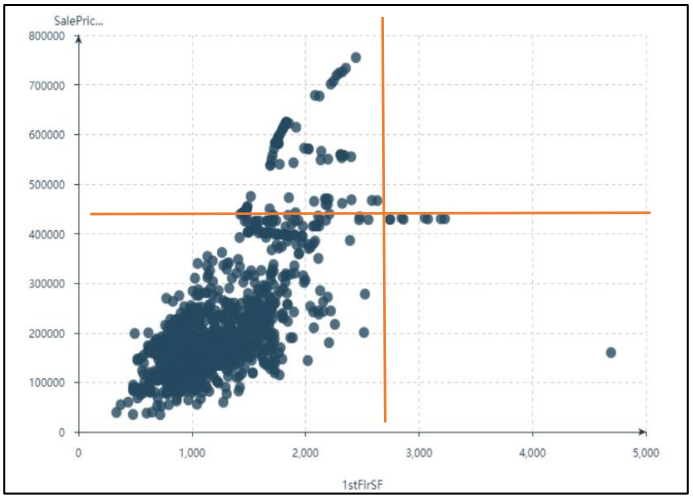


Figure 1.13 Scatterplot between 1stFlrSF and SalePrice

7) GrLivArea- Histogram depicts values are not normally distributed and is right skewed there are 1 or 2 properties having GrLivArea of 5000 & 6000. Fig 1.15 shows a scatterplot, and we can observe that the price of the house increases with the increase in GrLivArea it shows increasing trend. Most of the houses GrLivArea Is clustered between 500 to 2500. If we dissect scatterplot into four quarters, we can see under right bottom quarter there is one irregularity which is not part of the trend i.e., along with few outliers.

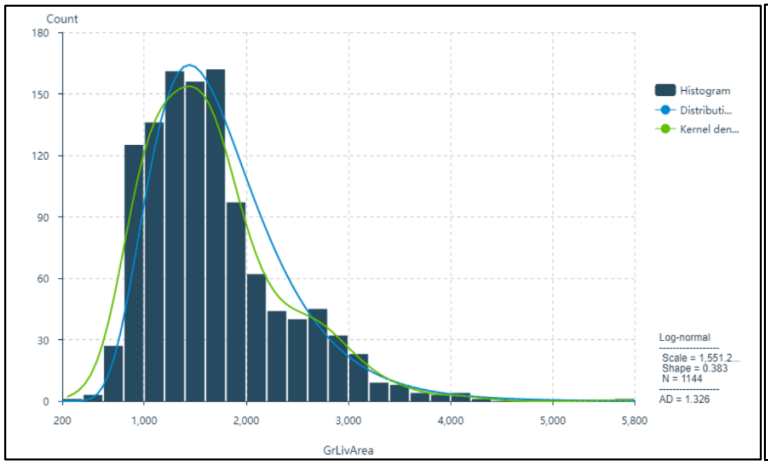


Figure 1.14 Histogram for GrLivArea

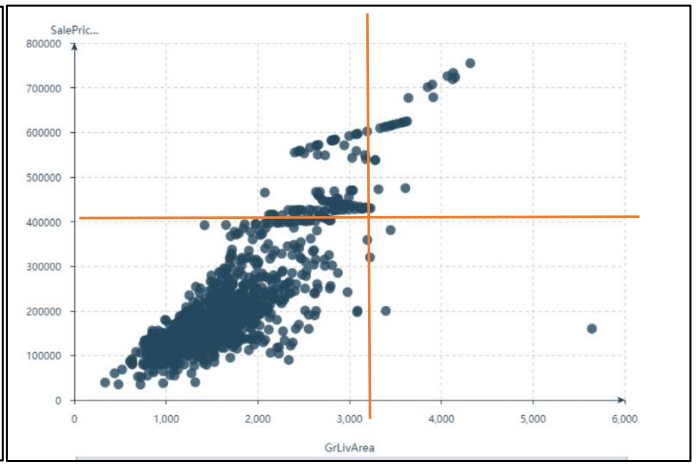


Figure 1.15 Scatterplot between GrLivArea and SalePrice

8) TotRmsAbvGrd-Since the variable is nominal and from the visualization. In fig 1.17 we can see the correct distribution of the category after Converting into integer. We can say from fig 1.17 that most of the houses in dataset consists of 6 or 7 rooms.

Fig 1.18 shows tells that the mean price for the houses with 11 rooms is higher than the houses with a smaller number of rooms however houses with 12 rooms have low sale price this may be due to some data inconsistency or may be houses with 12 rooms are in non-residential area where prices are low. Presence of outliers can also be observed under 4,5,6,7,8 & 11 rooms.

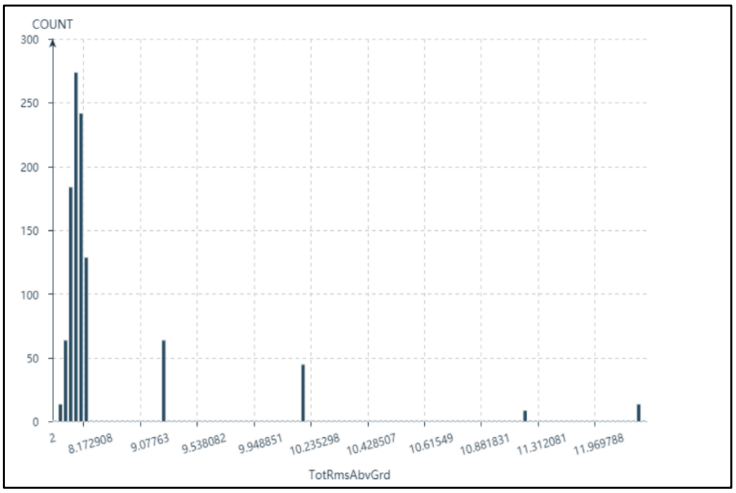


Figure 1.16 Univariate visualization for TotRmsAbvGrd

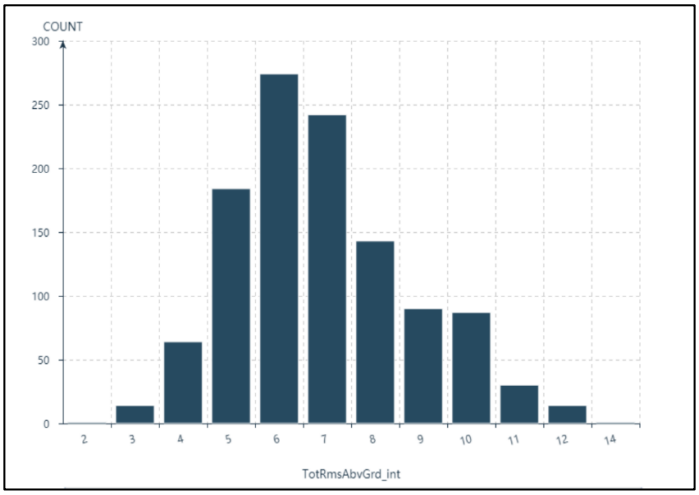


Figure 1.17 Univariate Visualization after converting to Integer

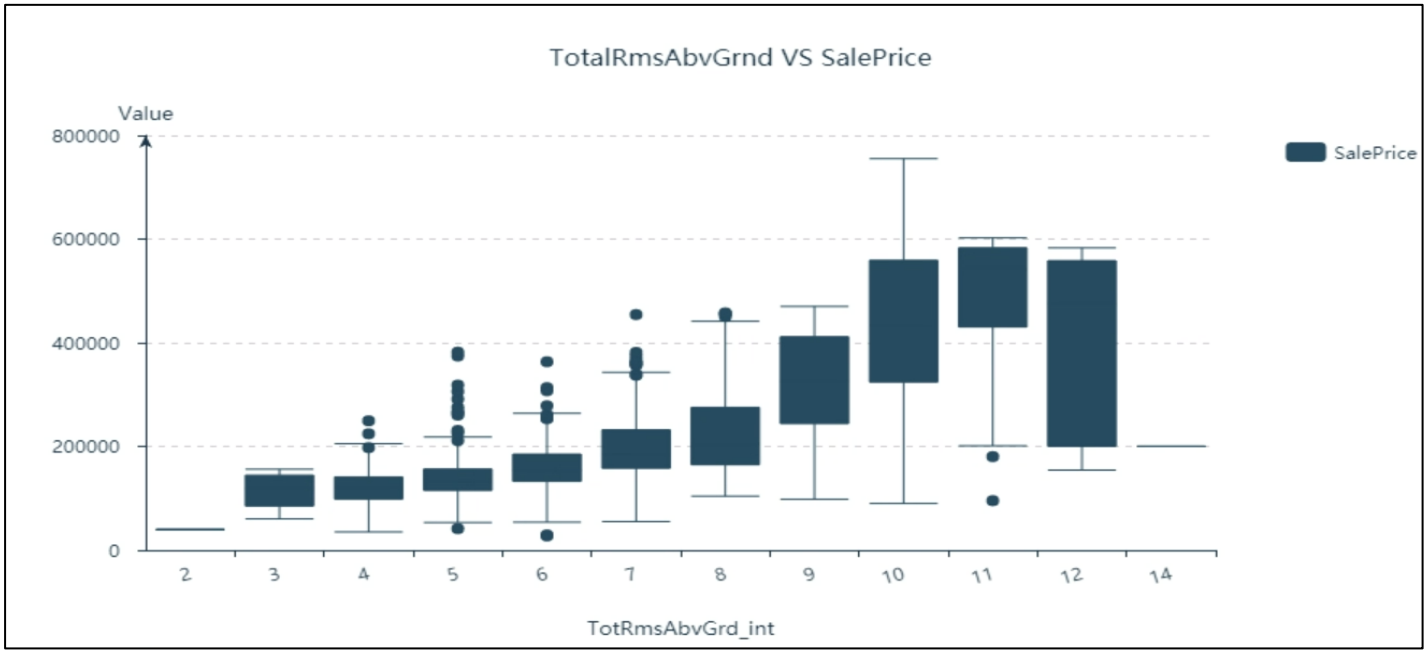


Figure 1.18 Boxplot for TotRmsAbvGrd and SalePrice

9) GarageType-from figure 1.19 We can say that large number of houses in dataset have attached garage instead of other types Moreover there are some houses which has no garage, and this can be observed from NA type having ~25 houses. Figure 1.20 shows that the average prices of houses having basement garage are higher than for other types of garages. Few outliers can be seen in attached, basement, detached and NA category of GarageType.

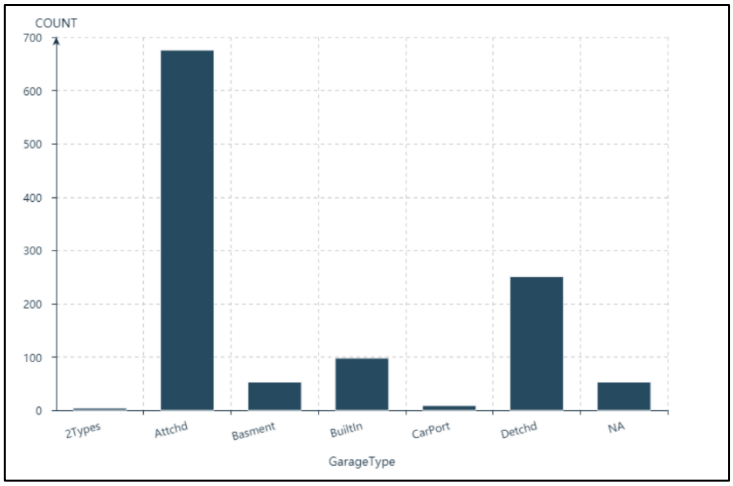


Figure 1.19 Bar plot for GarageType

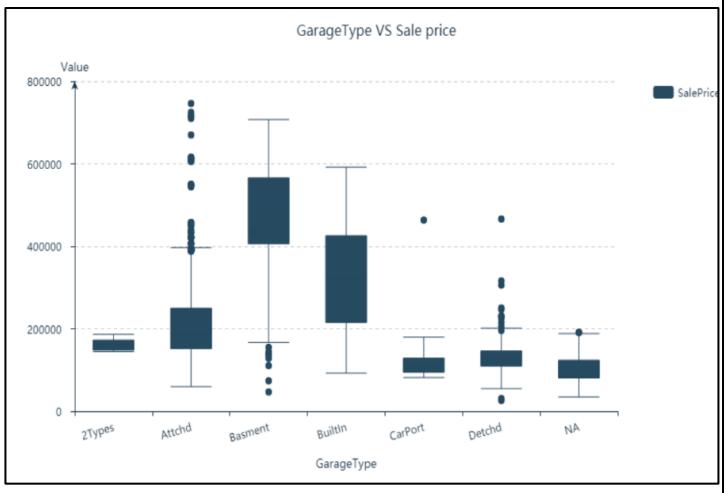


Figure 1.20 Box Plot for GarageType and SalePrice

10) GarageArea-Figure 1.21 reveals that values are not normally distributed, and there are three humps in the distribution, indicating that GarageArea has a multi model. We can observe from the scatterplot that the lot values for garage area are 0; these values relate to the NA garage type, indicating that the house does not have a garage. We can observe certain datapoints in the right bottom quadrant that are not following the trend since they offer a large parking size for a low price.

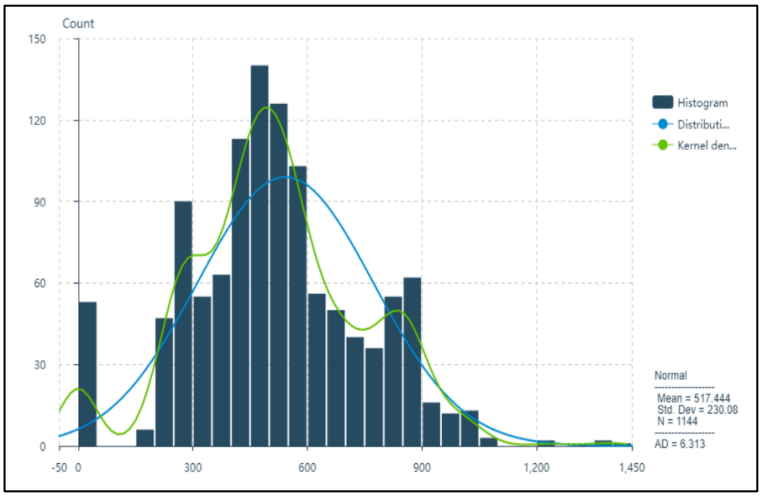


Figure 1.21 Histogram for Garage Area

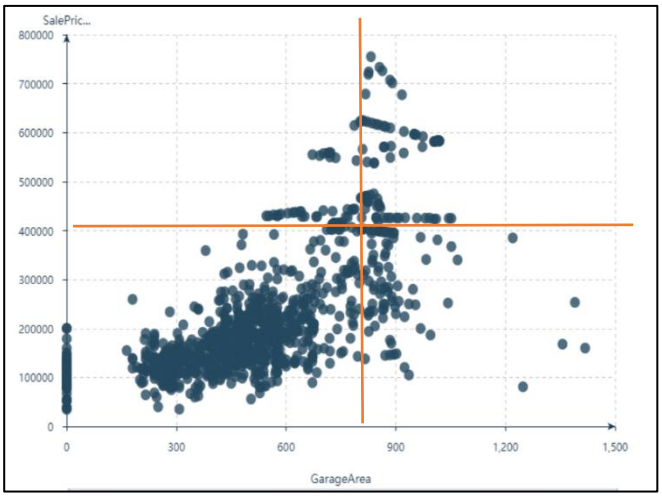


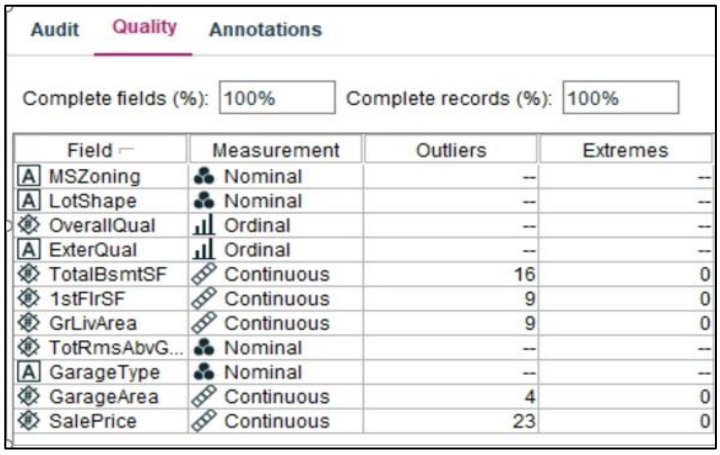
Figure 1.22 Scatterplot for Garage Area and SalePrice

Data Quality Assessment and Treatment

Outliers & Extremes

Outliers and extremes in the dataset must be dealt with because they can lead to erroneous results. However, if these values are corrected to reflect reality, it is preferable to maintain them in the dataset rather than eliminating them.

We used log transformation to produce normal distribution for continuous variables, which also manages the existence of outliers in the dataset.
For categorical variables we converted the variable to integer as the variable represent nominal or ordinal and values in dataset were in decimals.

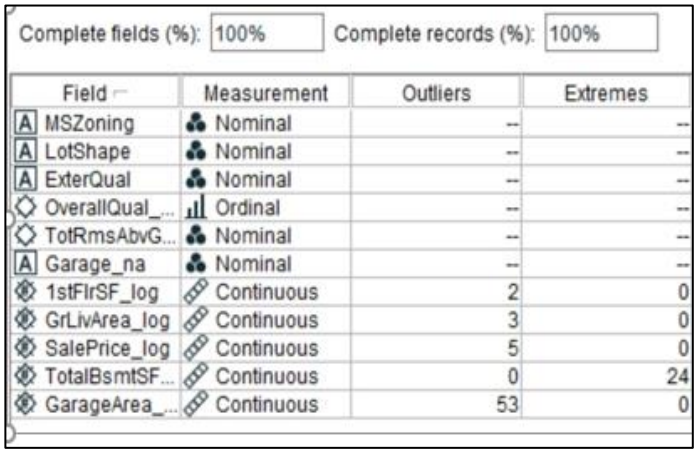


Audit | Quality | Annotations

Complete fields (%): 100% Complete records (%): 100%

Field	Measurement	Outliers	Extremes
MSZoning	Nominal	--	--
LotShape	Nominal	--	--
OverallQual	Ordinal	--	--
ExterQual	Ordinal	--	--
TotalBsmtSF	Continuous	16	0
1stFlrSF	Continuous	9	0
GrLivArea	Continuous	9	0
TotRmsAbvG...	Nominal	--	--
GarageType	Nominal	--	--
GarageArea	Continuous	4	0
SalePrice	Continuous	23	0

Figure 2.1 Outliers and Extremes before Log Transformation



Complete fields (%): 100% Complete records (%): 100%

Field	Measurement	Outliers	Extremes
MSZoning	Nominal	--	--
LotShape	Nominal	--	--
ExterQual	Nominal	--	--
OverallQual_...	Ordinal	--	--
TotRmsAbvG...	Nominal	--	--
Garage_na	Nominal	--	--
1stFlrSF_log	Continuous	2	0
GrLivArea_log	Continuous	3	0
SalePrice_log	Continuous	5	0
TotalBsmtSF...	Continuous	0	24
GarageArea_...	Continuous	53	0

Figure 2.2 Outliers and Extremes after Log Transformation

After taking a log for continuous, we can observe that there are now 53 outliers for Garage area and 24 for TotalBsmtSF in the above figure. The number of outliers was reduced after log transformation for all continuous variables except TotalBsmntSF and GarageArea because the dataset comprised 53 records with "0" garage area and 24 entries with "0" TotalBsmntSF. After conducting data analysis, we discovered that 0 represents the data's actuality, as the residences lacked a garage and a basement. Modeler transformed these 0 values into null values after log transformation (log0=undefined).

We had three options for dealing with null values resulting from log transformation: 1) impute using algorithm, which we couldn't do because there was no garage or basement, so imputing these values implied that we were going against reality 2) Remove the null values, however if we do, we won't be able to train our model for residences that don't have a garage or basement. 3) putting a 0 in these null variables to notify the model that these are the genuine values for no garage and no basement.

So, we filled null values with 0 for TotalBsmntSF and GarageArea after log transformation.

Missing Values- we don't have any value missing or blank in our dataset, we only have 0 value for garage area and basement area, however they represent reality and cannot be removed.

Anomaly- Anomaly detection is basically used to identify unusual cases in data. We have 1 anomaly in our dataset where we have irregular lot shape which excellent exterior and overall quality with large basement, 1st floor and ground living area, but the price of the house is low, this is very unusual case. Therefore, we handed it by removing it from our dataset

MSZoning	LotShape	OverallQual	ExterQual	TotalBsmtSF	1stFlrSF	GrLivArea	TotRmsAbvGrd	GarageType	GarageArea	SalePrice
RL	IR3	10	Ex	6110	4692	5642	12	Attchd	1418	160000

Figure 2.3 Anomaly in dataset

Data Leakage

When we use the target variable to fill in null values, we get data leakage. There was no data leakage in our situation because we didn't utilise any algorithm to compute missing values using the target variable (sale price).

MODELLING

Predictive modelling formulation

Different modelling techniques were used to estimate the property's sale price and compare it to the actual price, allowing us to quantify the difference between the projected and real sale price. The sale price can be predicted using multiple models and the predicting power of ten independent factors.

Type of the problem

The problem is regression problem since the target variable is numerical. Different algorithm used for predicting sale price are:

- Decision Tree
- Linear Regression
- Random Forest
- Neural Network

Beside this we also used Feature Engineering to check any improvement in the model's prediction power.

Target variable assessment and treatment

The target variable SalePrice is not normally distributed and is right skewed, as shown in the graph. We used the log transformation of the variable to try to normalise the data because the target variable is right skewed. This plot also demonstrates multi-modeling because there are 3,4 high peaks.

Log transformation has somewhat abled to normally distribute SalePrice however two model still can be seen even after transformation.

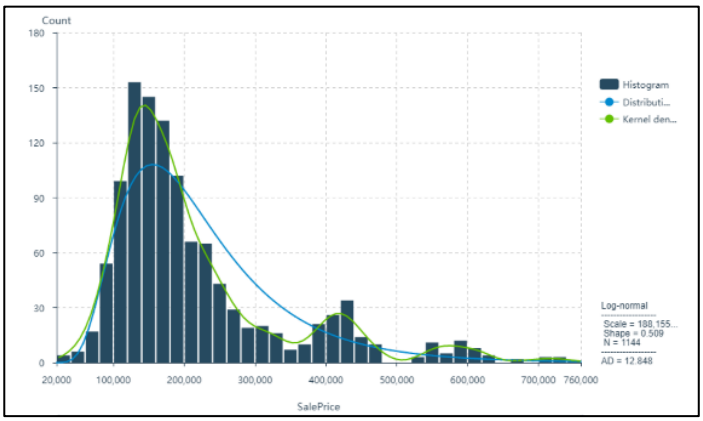


Figure 4.1 Histogram for SalePrice

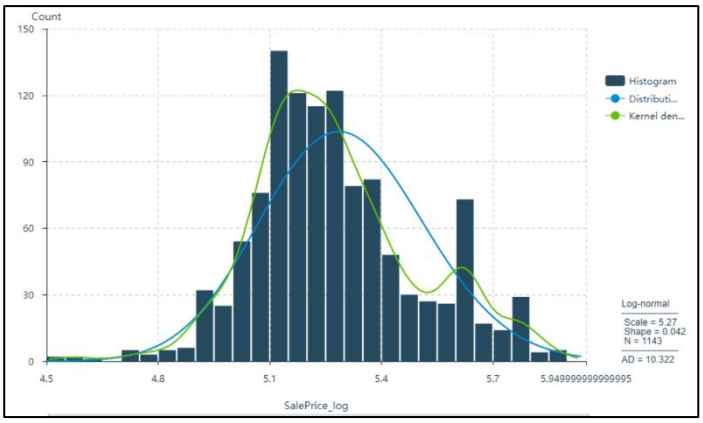
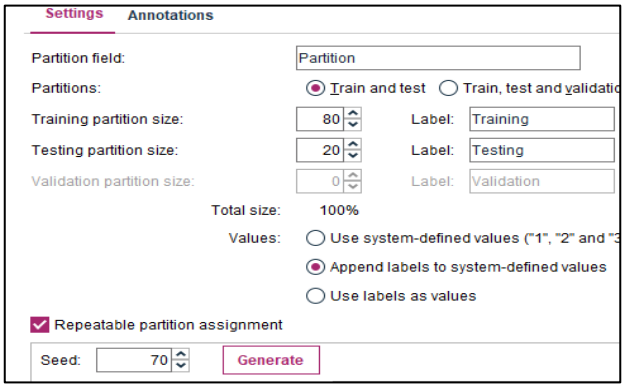


Figure 3.1.1 Histogram for SalePrice after log transformation

Partitioning Requirement

In our dataset, the data is partitioned into training and testing with an 80:20 ratio. Because we need to train our model using the relationship between predictors and target variable based on past data, we must partition the dataset into training and testing. Testing data

will evaluate model performance on unseen data whether model has learned or memorized the process.



SalePrice_log	TotalBsmtSF_log_trfd	GarageArea_log_trfd	Partition
5.319	2.932	2.739	1_Training
5.259	3.101	2.663	2_Testing
5.349	2.964	2.784	1_Training
5.398	3.059	2.922	1_Training
5.155	2.901	2.681	1_Training
5.487	3.227	2.803	2_Testing
5.301	3.044	2.685	2_Testing
5.114	2.979	2.670	1_Training
5.072	2.996	2.312	1_Training
5.112	3.017	2.584	2_Testing
5.538	3.070	2.867	1_Training

Figure 3.2 Partition of dataset in training and testing

Performance Metrics Used

Because we're dealing with a regression problem, performance measurements like R^2 be used to check for goodness of fit, as well as error metrics like RMSE or MAE (mean absolute error) to evaluate the difference between actual and projected values.

We picked MAE over RMSE to quantify error since we have a large number of outliers in our dataset that cannot be ignored because they represent reality; therefore, MAE is a better alternative than RMSE.

Baseline Model

Baseline model can be described as the simplest model; we use it as a benchmark to compare the performance of other models. This necessitates calculating the baseline model's MAE.

The MAE was derived using the Mean of the Actual Value as the anticipated value for the Baseline model.

MAE for baseline model is: 92191.99613

MAE on testing set using neural network is: 21368.64 (76% improvement from baseline).

Predictive Modelling

C&R Tree

The relevance of features may be seen in the fig, which is based on a decision tree. OverallQual is the most important factor, whereas MSZoning is the least important. The best MAE obtained after tweaking the hyperparameters is 27875.638 for training data and 31111.35 for testing data. There is a large discrepancy in training and testing error, indicating overfitting of the data.

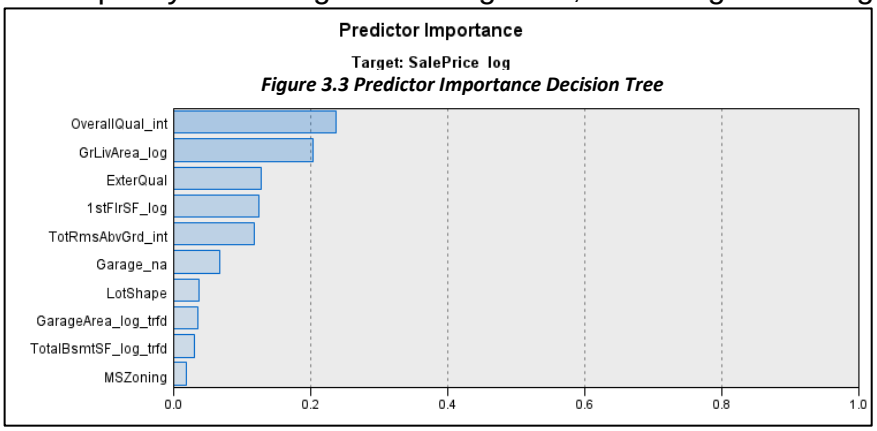


Figure 3.3 Predictor Importance Decision Tree

Linear Regression

Another model used to calculate the MAE linear regression model. The MAE obtained from linear regression on training and testing dataset is 30153.89 and 30254.42 respectively. Very less difference between actual and predicted.

Random Forest

The prediction importance based on the random forest approach may be shown in fig. 3.4. MSZoning is the most important variable, which is surprising given that it was the least important in Decision tree models.

The aim of random forest is to mix numerous decision trees to create a single, potentially powerful model. On the training and testing datasets, the best MAE is 30859.49 and 33971.32, respectively.

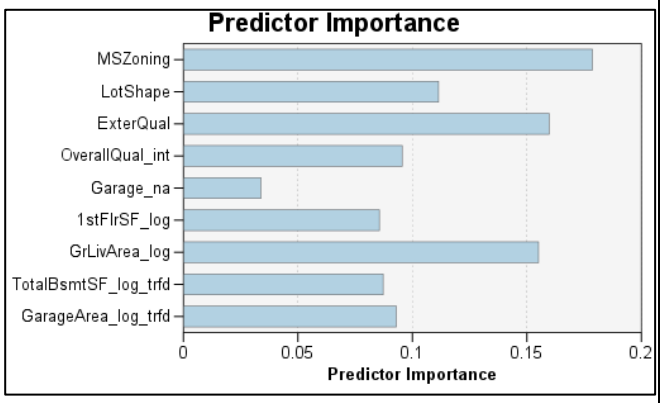


Figure 3.4 Predictor Importance by Random Forest

Neural Network

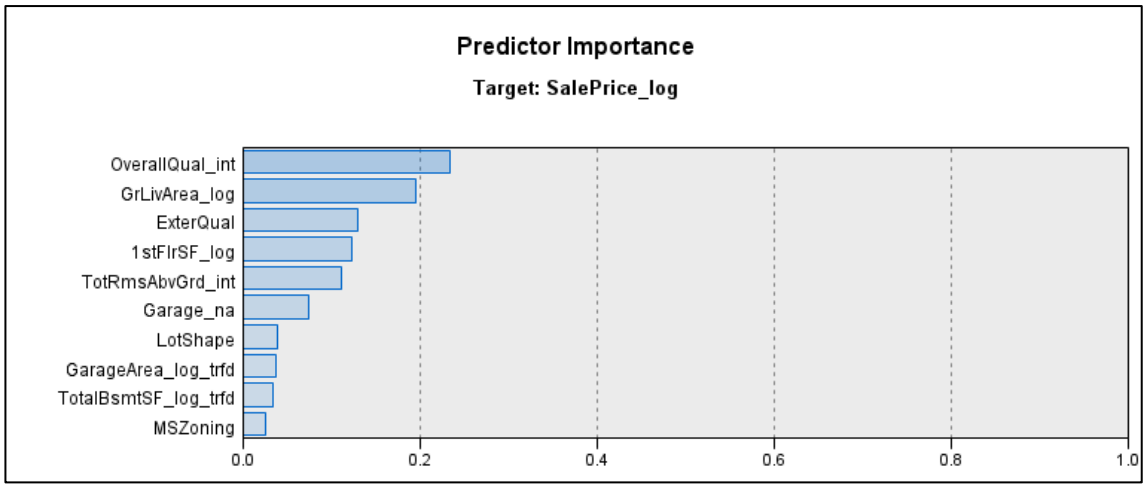


Figure 3.5 Predictor Importance by Neural Network

The feature importance generated by the Neural Network is shown in Figure 3.5. The approach is employed on the regression model since it outperforms many other methods, and the MAE achieved on the testing and training sets is lower than that of the other models.

The MAE for training and testing dataset is 18991.6 and 21368.64 respectively which is best performance so far which makes it best model for assessment.

Feature Engineering

To improve the model's prediction performance, we developed new features from existing features and utilised them in our feature to see if we could improve prediction values and MAE. Two new features were created in our model from the existing features:

- We did reclassification i.e., reclassified the categories for the field LotShape on best model neural network. Here we have 4 categories IR1(slightly irregular), IR2(Moderately irregular), IR3(irregular) and Reg(regular). We put together the values for IR1, IR2, IR3 as one category Irregular and other category we have is regular. MAE is 19135.87 on training and 21142.51 on testing. FE of reclassification aided in the reduction of neural network testing error.

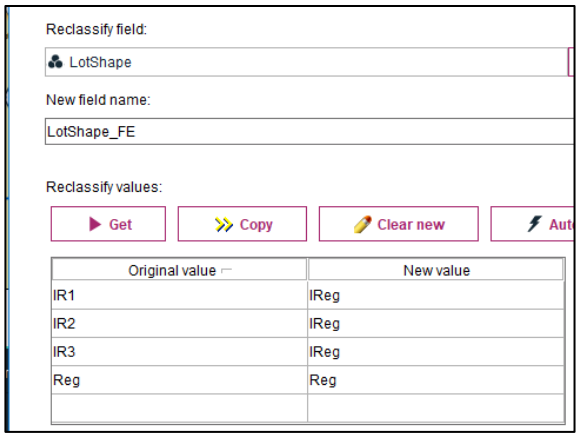


Figure 3.6 Reclassification of field LotShape

- we used clustering to build 10 clusters for the TotalBsmtSF predictor. The MAE for training is 20456.8, while the MAE for testing is 22435.32, indicating that the testing error is larger than the training error in the original model. As a result, clustering had no effect on performance.

Hyperparameters Tuning: The top parameters on which models worked well are listed below. A neural network with two hidden layer 1 nodes and seven hidden layer 2 nodes with bagging resulted in a low MAE on both training and testing. Even after multiple changes in the number of models and tree depth, random forest performed the poorest. **These are best hyperparameters, complete file is attached for the reference.**

S.No	Model	Parameters					Training	Testing
1	Decision Tree	Bagging	custom:10+no pruning	stoppin rule:2:1	impurity measure	gini	0.45	0.53
					overfit prevention set	30		
					random seed	681644031		
2	Linear Regression	default	method:Enter	mode:simple			0.61	0.063
3	Random Forest	number of models	300				0.62	0.67
		sample size	1					
		max tree depth	15					
		min child node	7					
4	Neural Network	Bagging	Multilayer perceptron	1st layer:2	2nd layer: 7		0.42	0.47

Method Used

Because we're utilising SalePrice to forecast the property's price, it's a supervised learning method.

Methods			
Decision Tree	Linear Regression	Random Forest	Neural Network
It may be used for classification and regression analysis, with all predictors being fed into a decision tree model in which datasets are dispersed into feature space and impurity is measured using trees and their depths	With the help of the OLS model, all of the predictors are fitted in such a way that the difference between actual and projected SalePrice is minimised, and the unknown SalePrice may be determined by projecting it on the line of best fit	Random Forest calculates the average performance of many decision trees by iterating all possible parameter combinations	To reduce the error between actual and anticipated sales, target values are evaluated by modifying weights for all predictors under hidden layers. The sigmoid function and backpropagation methods are used to calculate the price

Error Cost Analysis

Since we have a regression problem cost error could be overestimated or underestimated based on the actual and the predicted values of the SalePrice.

Case 1: **Overestimate**- If the property predicted price is higher than the actual price, we can say that our model is overestimated. If the real estate agent sells the property at the overestimated price, the property seller will benefit financially, while the property buyer will lose money because he or she paid more than the actual cost of the property.

It's also possible that an overpriced house will not sell, in which case the owner or property seller will lose money.

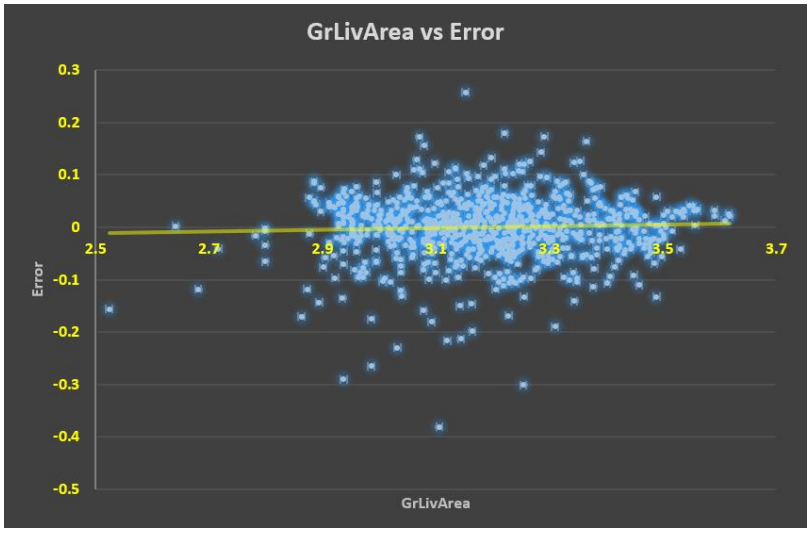
Case 2: **Underestimate-** If the predicted price is lower than the actual price, we can say that the model is underestimating the property. This will allow the property to sell quickly because the price is lower than the actual cost, but the property seller will lose money because the price is underestimated, and the property will not sell for the desired price. Whereas, in the instance of a property buyer, he or she would feel pleased to have made a good deal by purchasing a house at an undervalued price.

Overestimate can be worst-Buyer.

Underestimate can be worst- Seller, Real estate agent.

Error Values vs Predictors.

As can be seen in the diagram, the error term has a zero mean and a constant variance. However, there are a few errors that are somewhat off the mean, i.e., below the trendline, but overall, we can state that our model is fine for prediction.



Note: All values are in log.

Conclusion

Apart from given features of the property, sale price also depends on other factors such as nearest to hospitals, schools, malls, stock prices etc. Using predictive modelling we can get an estimation of the price but can't expect accurate results. With the help of our best model, we can expect difference of 21,368 in property price of new house excluding other environmental factors either selling or buying.

Recommendation- Because low-quality data might lead to erroneous outcomes, data must be gathered accurately and carefully also quantity of data may increase performance, and we can predict better outcomes. Our model can be used by a variety of stakeholders for a variety of purposes, such as estimating new house prices or understanding essential elements in a property that may affect the price.

Appendix

[Calculation.xlsx]

[HyperParameter tuning.xlsx]