THE HOUSING PROJECT

SALE PRICE PREDICTION

Submitted by:

Glenn Nigel Ebenezer

ACKNOWLEDGMENT

My sincere gratitude to the institute –Datatrained and Internship company-FlipRobo for giving me the opportunity to learn and evolve in the field of datscience.

INTRODUCTION

- Business Problem Framing

  The Problem Statement involved predicting the Sale Price of Housing Units spread across certain locality, and physical and land specifications existing over a timeframe of age, remodelling etc.

- Conceptual Background of the Domain Problem

  A fair idea of how real estate works, how prices vary with different attributes be it physical, geographical coordinates, additional features etc. I personally enjoyed working on this project as it had a lot to do with my present job occupation.

- Review of Literature

  A Study was done about the real estate pricing across the areas mentioned in the “neighbourhood” column. And additionally, how apart from the physical features listed as the dataset columns, other attributes such as

- size of kitchen, style of kitchen cabinetry,
- flooring,
- heating,
- cellar,
- earthquake resistant,
- tornado and hurricane probability and the housing units counter measures
- Housing Insurance and its cover
- Bedroom furnishing
- Power Backup –Diesel Generator
- Surveillance of property

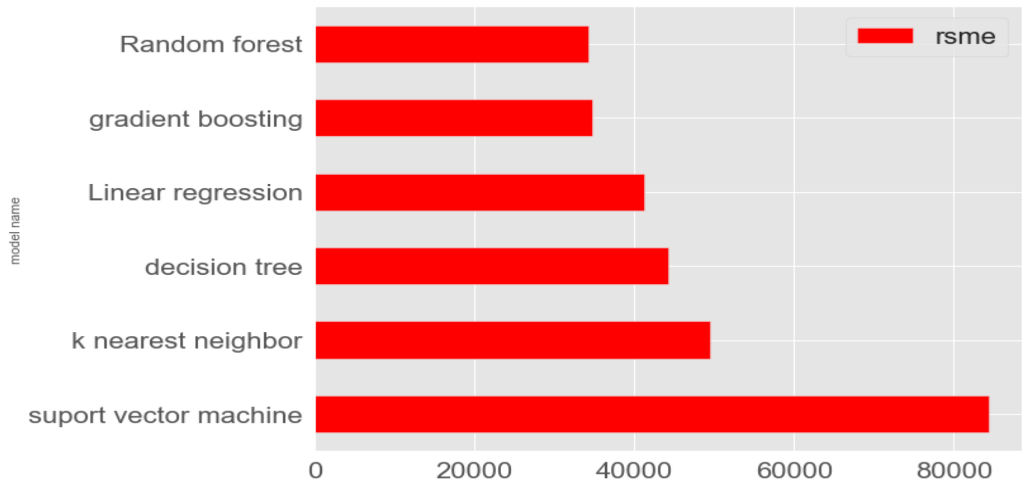
These above among others turned out to be key points determining the SalePrice, which unfortunately wernt in this dataset.
But a rough idea was surely acquired to understand the Dataset and its requirements better.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

The following steps were made based on our EDA:

a. Null values were replaced using mode and mean method. Some columns with more than 45% null values, were dropped. Hence all null values were treated.
b. Used Label and one hot encoding on numerical categorical data.
c. Using an imputer, fitting on the train dataset was applied on the test dataset to fill all the Null values. We couldn't afford to drop these values as the dataset wasn't large enough. Hence this was treated too.
d. The train and test data features were scaled using min max scalar by scikit.
e. We used a kit of models to predict, but got the best predictions from SVR machine model.



f. We then used Grid search CV to do a hyper parameter tuning to find the best parameters to fit and transform model and then again predicted using the same model but with best features.

g. Next we applied this model and predicted the Sale Price based on the test data. where we achieved a mean square log error of about 0.234…

- Data Sources and their formats

Out[314]:

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	Neighborhood	Condition1	Cc
0	337	20	RL	86.0	14157	Pave	NaN	IR1	HLS	AllPub	Corner	Gtl	StoneBr	Norm	
1	1018	120	RL	NaN	5814	Pave	NaN	IR1	Lvl	AllPub	CulDSac	Gtl	StoneBr	Norm	
2	929	20	RL	NaN	11838	Pave	NaN	Reg	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	
3	1148	70	RL	75.0	12000	Pave	NaN	Reg	Bnk	AllPub	Inside	Gtl	Crawfor	Norm	
4	1227	60	RL	86.0	14598	Pave	NaN	IR1	Lvl	AllPub	CulDSac	Gtl	Somerst	Feedr	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
287	83	20	RL	78.0	10206	Pave	NaN	Reg	Lvl	AllPub	Inside	Gtl	Somerst	Norm	
288	1048	20	RL	57.0	9245	Pave	NaN	IR2	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	
289	17	20	RL	NaN	11241	Pave	NaN	IR1	Lvl	AllPub	CulDSac	Gtl	NAmes	Norm	
290	523	50	RM	50.0	5000	Pave	NaN	Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Feedr	
291	1379	160	RM	21.0	1953	Pave	NaN	Reg	Lvl	AllPub	Inside	Gtl	BrDale	Norm	

292 rows × 80 columns

In [14]:

```
# display the content of the train data
train.sample(5)
```

Out[14]:

dDeckSF	OpenPorchSF	EnclosedPorch	3SsnPorch	ScreenPorch	PoolArea	PoolQC	Fence	MiscFeature	MiscVal	MoSold	YrSold	SaleType	SaleCondition	Sa
144	78	0	0	0	0	NaN	NaN	NaN	0	7	2009	New	Partial	
156	20	0	0	0	0	NaN	NaN	NaN	0	5	2008	WD	Normal	
0	0	0	0	0	0	NaN	MnPrv	NaN	0	4	2008	WD	Normal	
0	104	0	0	0	0	NaN	NaN	NaN	0	4	2010	WD	Normal	
0	0	0	0	0	0	NaN	NaN	NaN	0	6	2006	WD	Normal	

1. From the above displays, we can see that there is one less column in the test data which is the sale price i.e the target.
2. There are alot of missing values.
3. In both data there are fairly large number of categorical and numerical values.

The dataset had a mixture of float, obj and int values.

Of these values, they had a mixture of categorical and numerical data.

Of the categorical data, there were some with int and float values which had to be encoded using label and one hot encoder.

- State the set of assumptions (if any) related to the problem under consideration

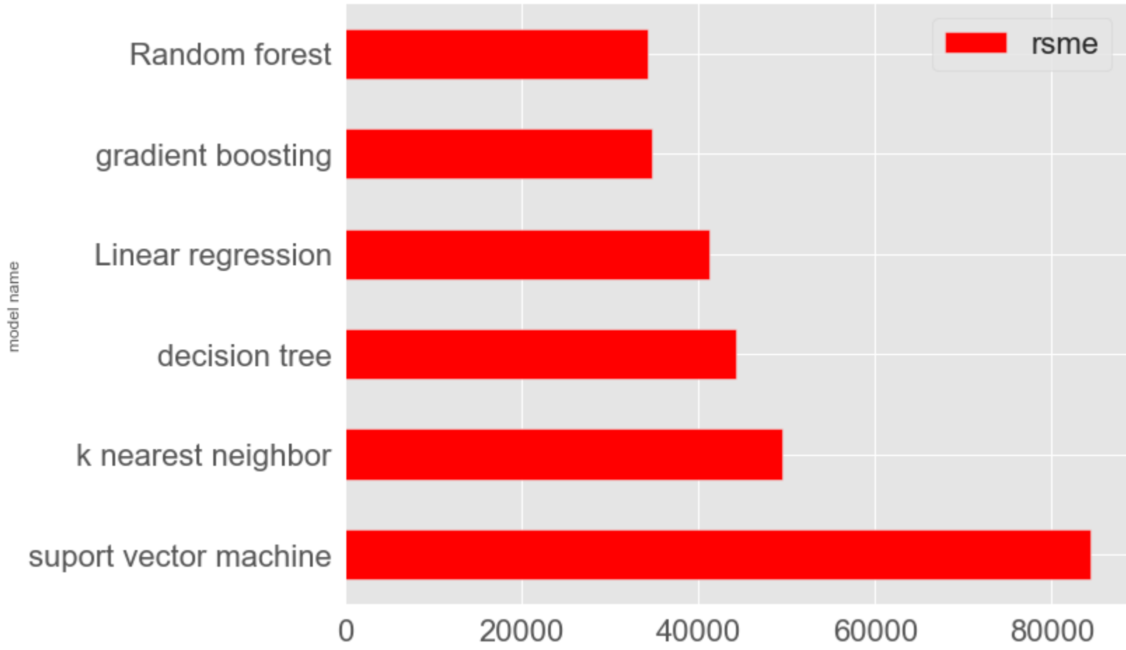
  Here we assumed looking at the EDA that features with more than 45% can be dropped and may not have a substantial relationship with the sale price.

- Hardware and Software Requirements and Tools Used
  The Problem statement was solved using Anaconda-Jupyter Notebook and the internet was used for research.

Model Development and Evaluation

- Identification of possible problem-solving approaches (methods)
  a. We did not start off with any assumption on the selection of the model to be used. We used Random forest, gradient Boosting, linear regression, decision tree, k nearest neighbour, and support vector regression machine models to predict the target label.
  b. We finally picked the best model based on the its outcome, which was Support Vector Regression Machine model.
  c. Then we used grid search cv method to find the best parameters and again ran the model with best parameters.
  d. We then used this model on the test data and achieved similar results.
  e. We evaluated the model using the root mean square error and finally got a value of rmse= 0.02378911039914229 on the test data.

- Testing of Identified Approaches (Algorithms)



- Run and Evaluate selected models

We used the following models to predict the target label.

a. Random forest
b. Gradient Boosting
c. Linear regression
d. Decision tree
e. KNearest neighbour
f. Support vector regression machine model

The snapshot of the codes used.

Out[314]:

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	LotConfig	LandSlope	Neighborhood	Condition1	Cc
0	337	20	RL	86.0	14157	Pave	NaN	IR1	HLS	AllPub	Corner	Gtl	StoneBr	Norm	
1	1018	120	RL	NaN	5814	Pave	NaN	IR1	Lvl	AllPub	CulDSac	Gtl	StoneBr	Norm	
2	929	20	RL	NaN	11838	Pave	NaN	Reg	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	
3	1148	70	RL	75.0	12000	Pave	NaN	Reg	Bnk	AllPub	Inside	Gtl	Crawfor	Norm	
4	1227	60	RL	86.0	14598	Pave	NaN	IR1	Lvl	AllPub	CulDSac	Gtl	Somerst	Feedr	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
287	83	20	RL	78.0	10206	Pave	NaN	Reg	Lvl	AllPub	Inside	Gtl	Somerst	Norm	
288	1048	20	RL	57.0	9245	Pave	NaN	IR2	Lvl	AllPub	Inside	Gtl	CollgCr	Norm	
289	17	20	RL	NaN	11241	Pave	NaN	IR1	Lvl	AllPub	CulDSac	Gtl	NAmes	Norm	
290	523	50	RM	50.0	5000	Pave	NaN	Reg	Lvl	AllPub	Corner	Gtl	BrkSide	Feedr	
291	1379	160	RM	21.0	1953	Pave	NaN	Reg	Lvl	AllPub	Inside	Gtl	BrDale	Norm	

292 rows × 80 columns

In [14]:

```
# display the content of the train data
train.sample(5)
```

Out[14]:

dDeckSF	OpenPorchSF	EnclosedPorch	3SsnPorch	ScreenPorch	PoolArea	PoolQC	Fence	MiscFeature	MiscVal	MoSold	YrSold	SaleType	SaleCondition	Sa
144	78	0	0	0	0	NaN	NaN	NaN	0	7	2009	New	Partial	
156	20	0	0	0	0	NaN	NaN	NaN	0	5	2008	WD	Normal	
0	0	0	0	0	0	NaN	MnPrv	NaN	0	4	2008	WD	Normal	
0	104	0	0	0	0	NaN	NaN	NaN	0	4	2010	WD	Normal	
0	0	0	0	0	0	NaN	NaN	NaN	0	6	2006	WD	Normal	

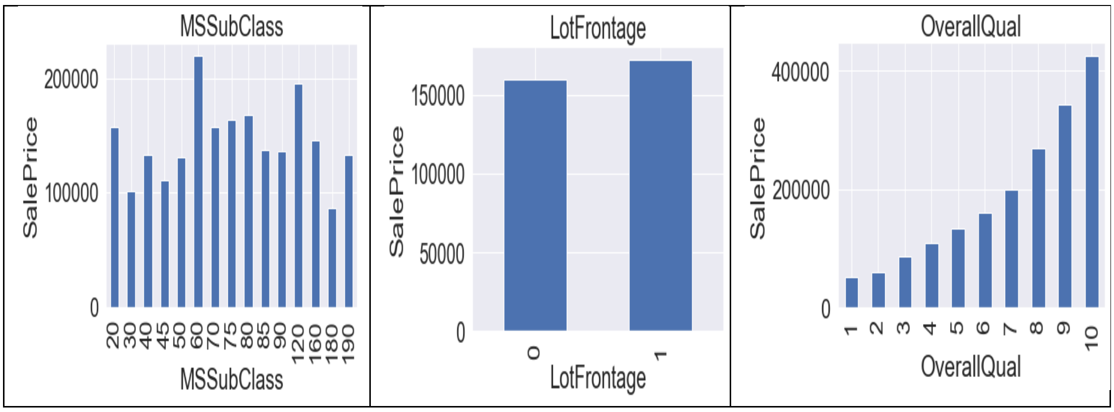
1. From the above displays, we can see that there is one less column in the test data which is the sale price i.e the target.
2. There are alot of missing values.
3. In both data there are fairly large number of categorical and numerical values.

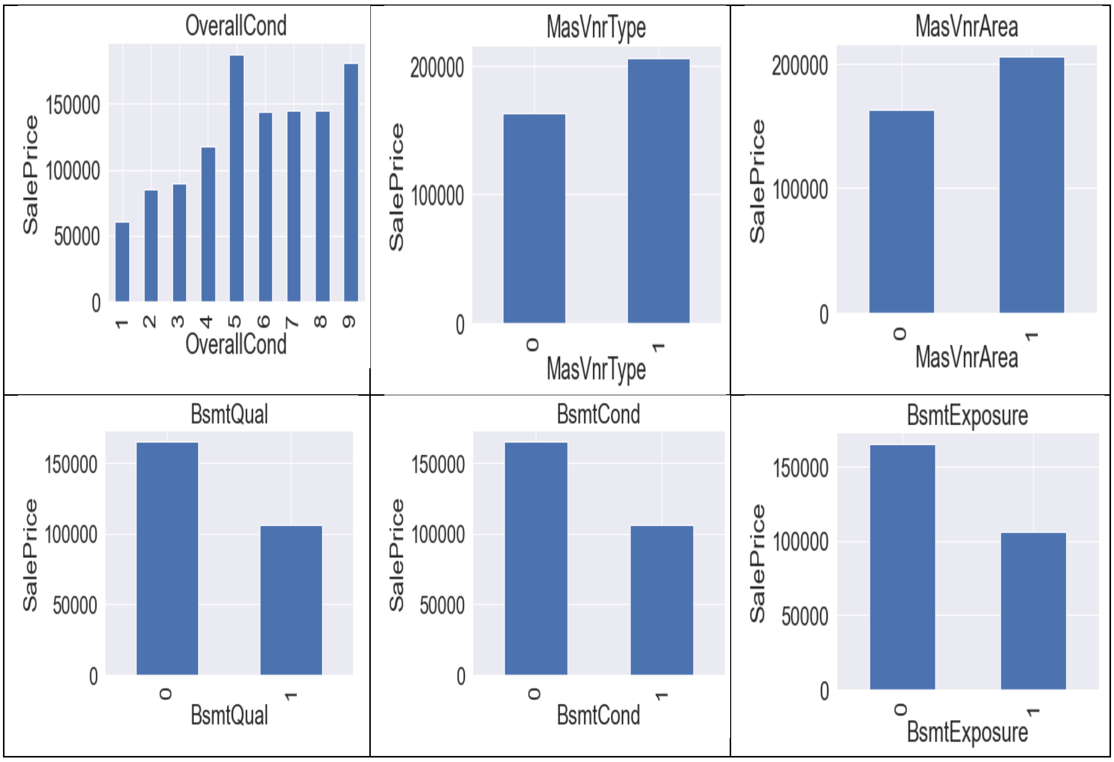
- Key Metrics for success in solving problem under consideration

  Used the RMSE or the Root mean square Error as the metric for this model.

- Visualizations

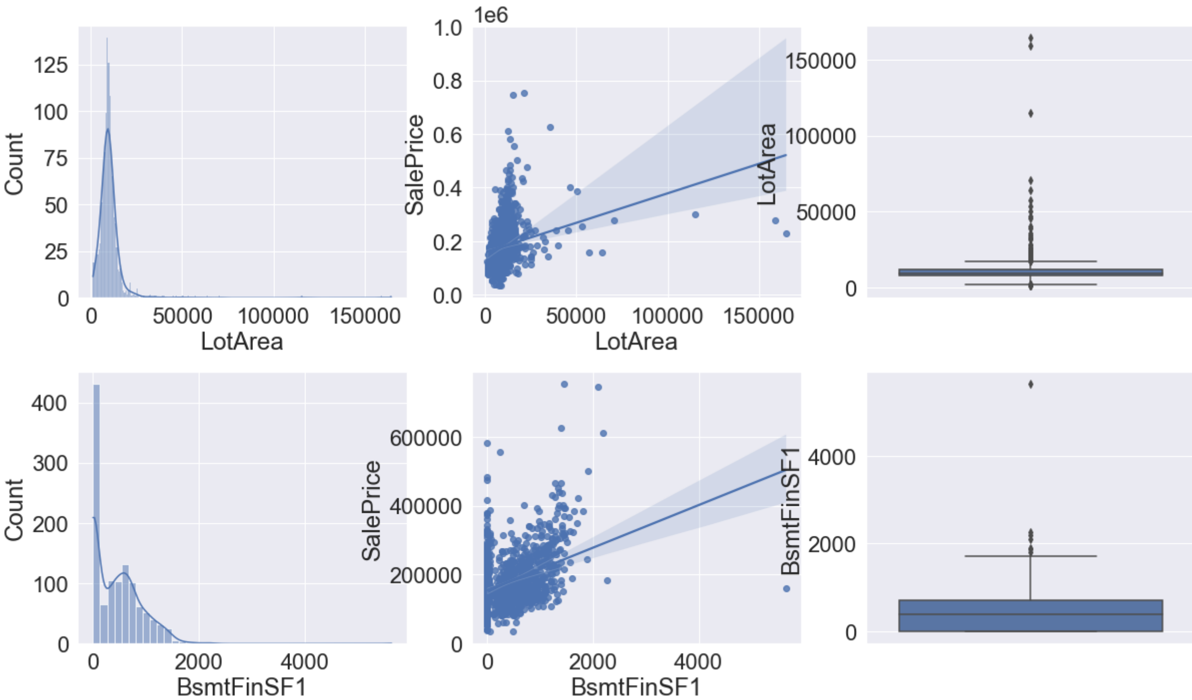
  Categorical data columns:

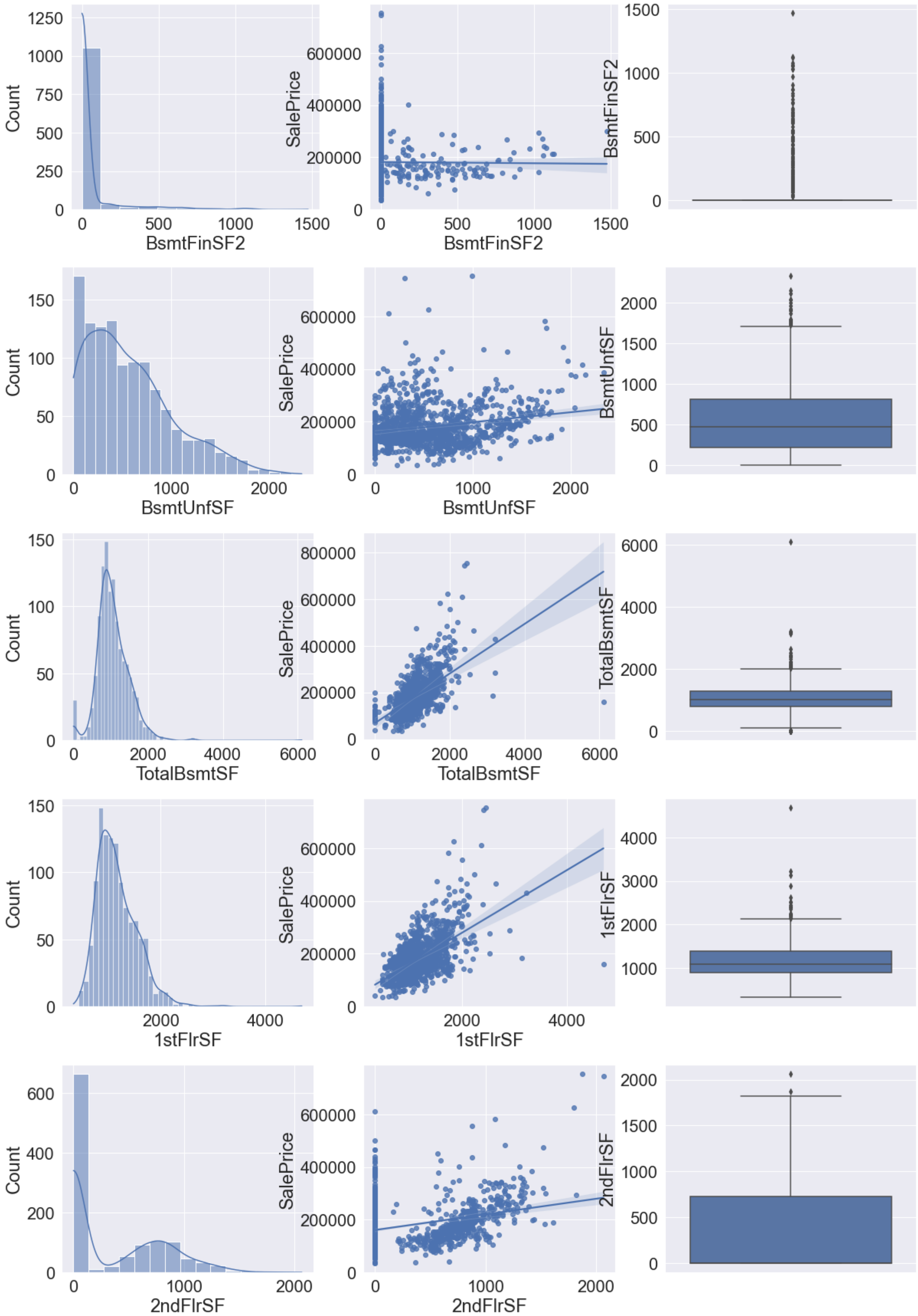


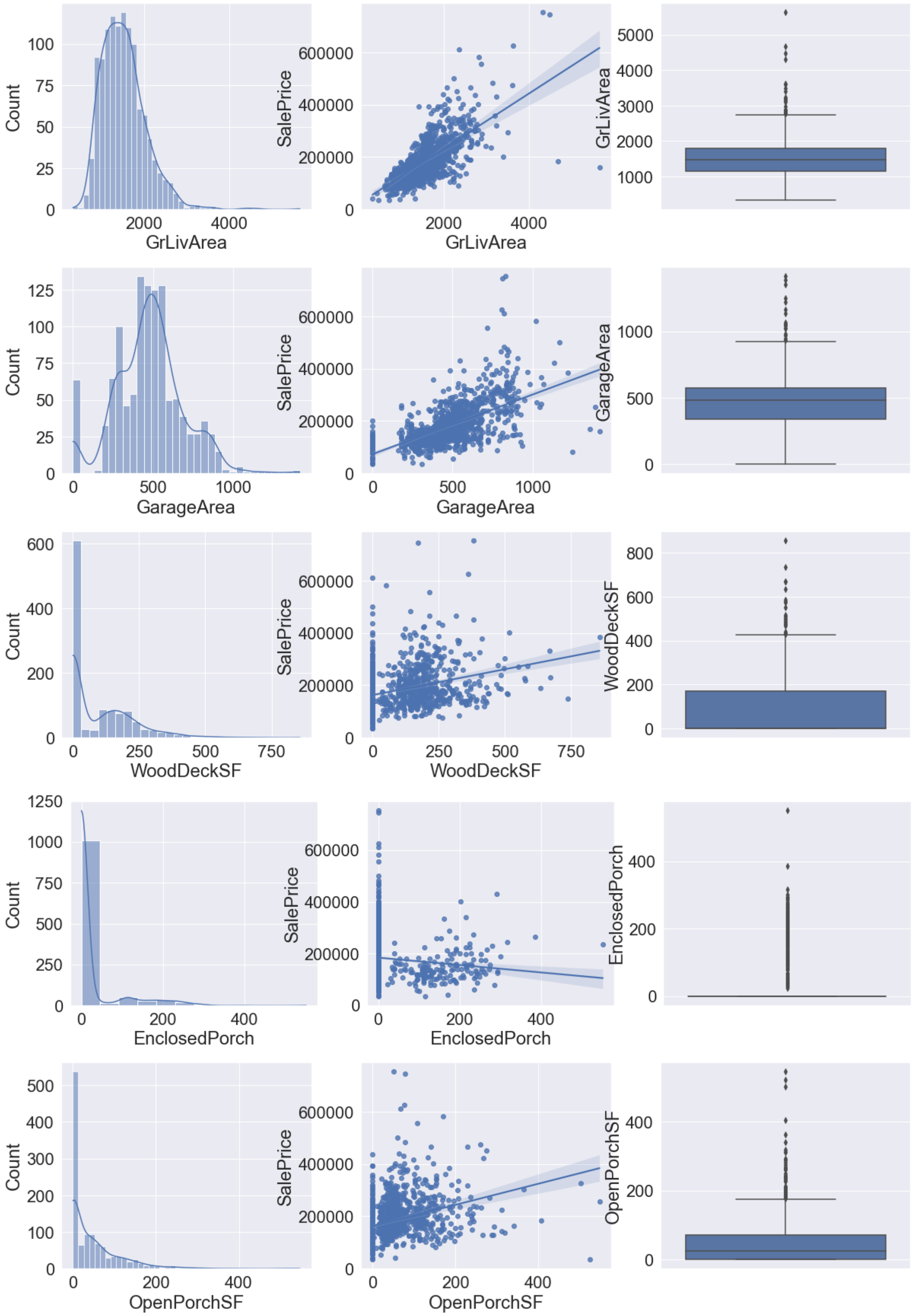


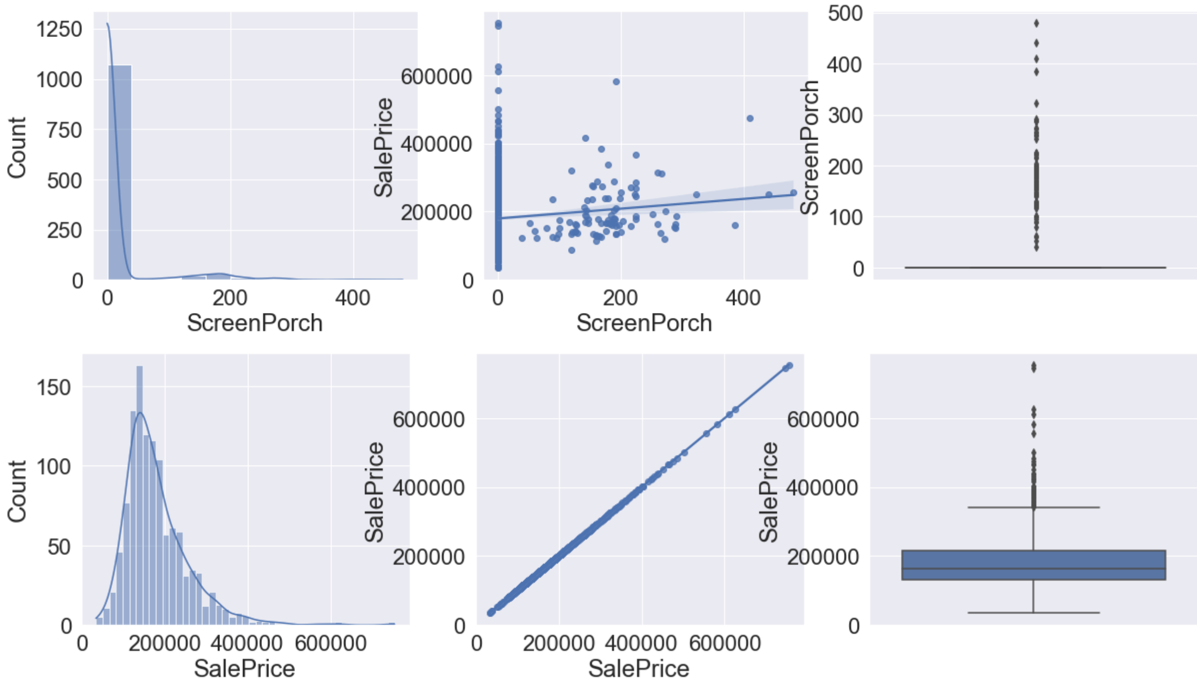
and so on………..

Numerical data columns visualisation:

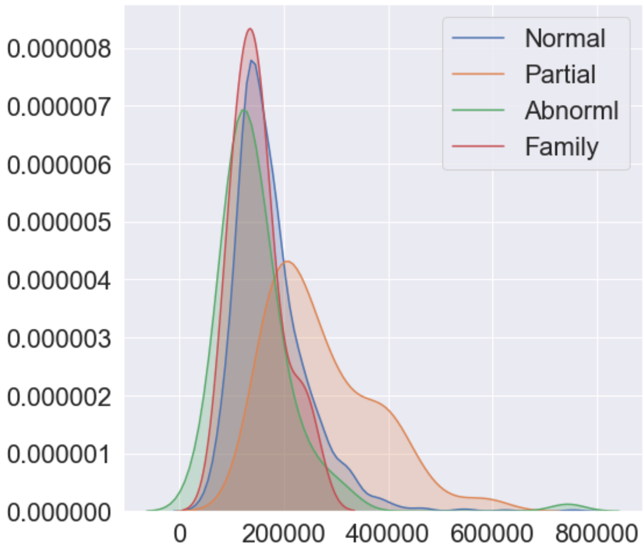



Count
1250
1000
750
500
250
0
0
500
1000
1500
BsmtFinSF2
SalePrice
600000
400000
200000
0
0
500
1000
1500
BsmtFinSF2
BsmtFinSF2
1500
1000
500
0
Count
150
100
50
0
0
1000
2000
BsmtUnfSF
SalePrice
600000
400000
200000
0
0
1000
2000
BsmtUnfSF
BsmtUnfSF
2000
1500
1000
500
0
Count
150
100
50
0
0
2000
4000
6000
TotalBsmtSF
SalePrice
800000
600000
400000
200000
0
0
2000
4000
6000
TotalBsmtSF
TotalBsmtSF
6000
4000
2000
0
Count
150
100
50
0
2000
4000
1stFlrSF
SalePrice
600000
400000
200000
0
2000
4000
1stFlrSF
1stFlrSF
4000
3000
2000
1000
Count
600
400
200
0
0
1000
2000
2ndFlrSF
SalePrice
600000
400000
200000
0
0
1000
2000
2ndFlrSF
2ndFlrSF
2000
1500
1000
500
0

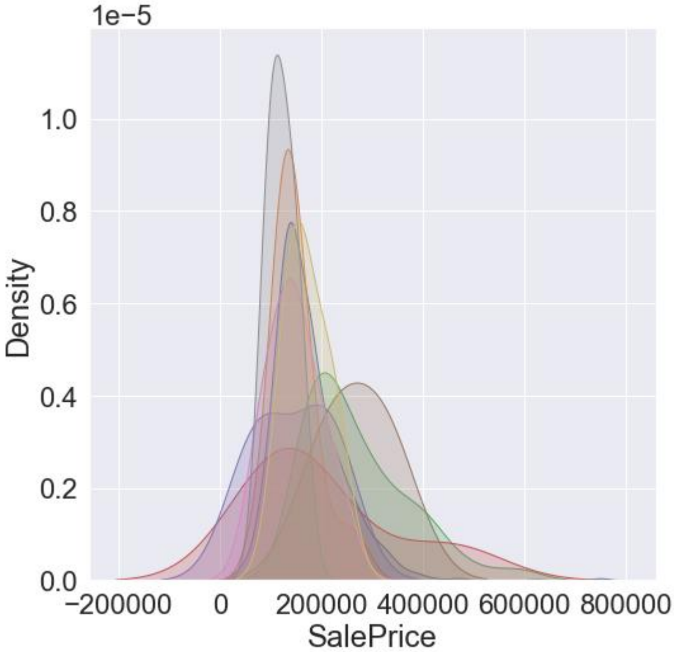

Count
GrLivArea
0 25 50 75 100
2000 4000
SalePrice
0 200000 400000 600000
GrLivArea
1000 2000 3000 4000 5000
Count
GarageArea
0 25 50 75 100 125
0 500 1000
SalePrice
GarageArea
0 500 1000
Count
WoodDeckSF
0 200 400 600
0 250 500 750
SalePrice
WoodDeckSF
0 200 400 600 800
Count
EnclosedPorch
0 250 500 750 1000 1250
0 200 400
SalePrice
EnclosedPorch
0 200 400
Count
OpenPorchSF
0 100 200 300 400 500
0 200 400
SalePrice
OpenPorchSF
0 200 400



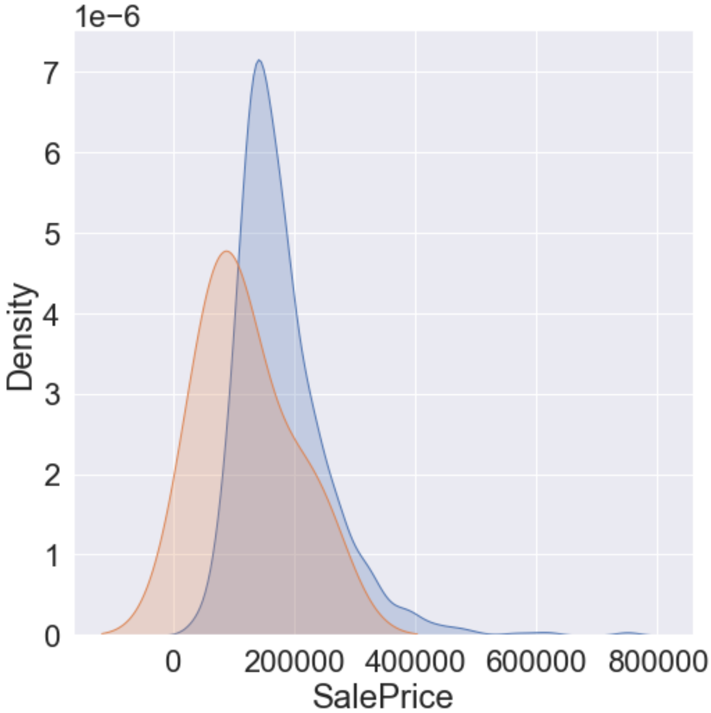
Analysing categorical data columns with more than 5 unique values.
Saleprice VS SaleCondition:



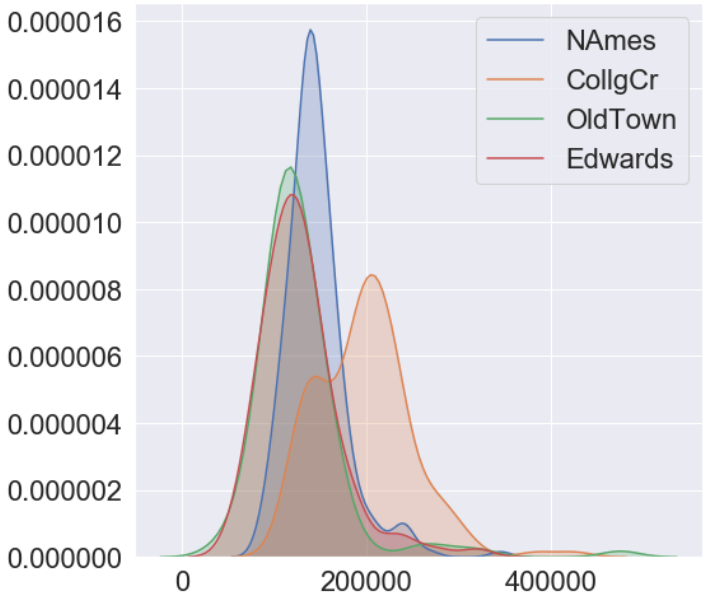
Saleprice VS Saletype:



SalePrice VS Street



SalePrice VS Neighborhood:

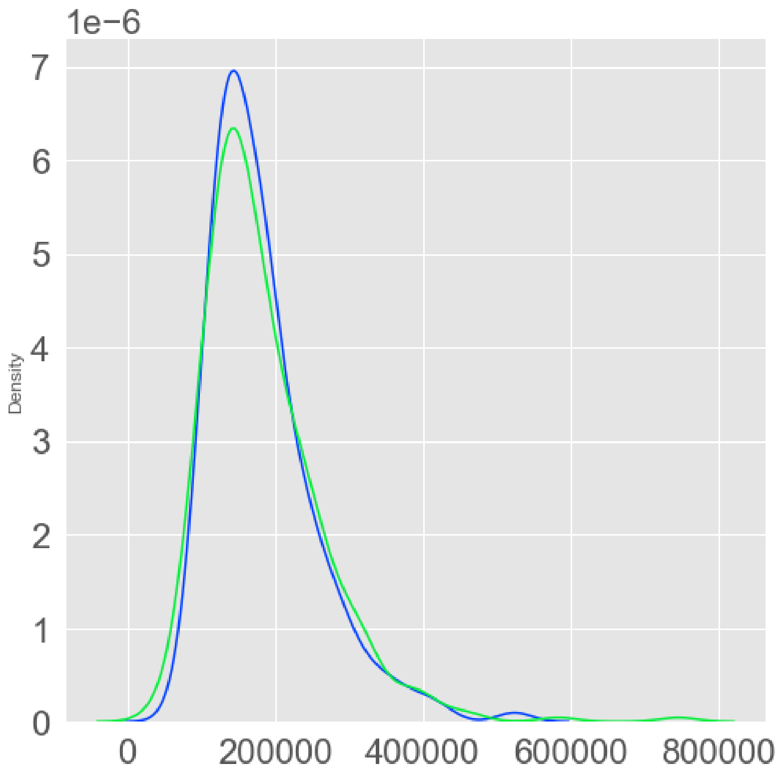


Other Plots like correlation plots are too big to be legible on this document, hence they have been truncated.

- Interpretation of the Results

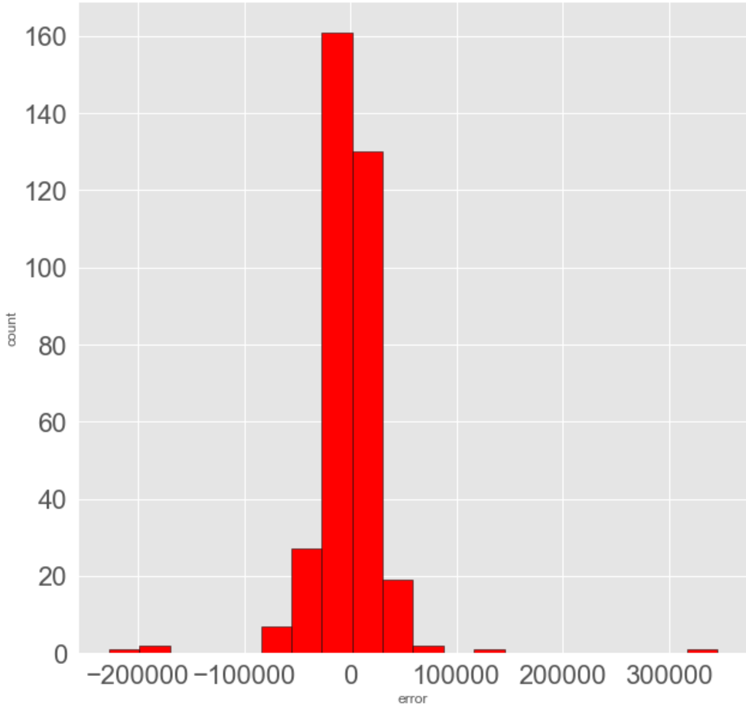
We successfully built our model and predicted the Sale Price in the test data. We used rmse to evaluate the model and found that it has an error of about 0.234….

Here are the plots for better visualisation.



We see that the test data is more or less in line with the train data plot, so we can assume that it maybe accurate.

Let us loos at the error plot



CONCLUSION

- Key Findings and Conclusions of the Study

  There were a lot of issues with the dataset, had to be cleaned with a lot of preprocessing techniques.

- Learning Outcomes of the Study in respect of Data Science & Limitations of this work and Scope for Future Work

  The results showed us that machine learning is applicable to our problem, with the final model able to the predict the SalePrice.

  Move over, as the number of trees increases, the amount of over fitting increases. Both the test and training error decrease as the number of trees increase but the training error decreases more rapidly.

  We also saw that hyperparamter tuning was able to improve the performance of the model although at a considerable cost in terms of time invested. This is a good reminder that proper feature engineering and gathering more data has a much larger pay-off than fine-tuning the model. We also observed the trade-off in run-time versus accuracy, which is one of many considerations we have to take into account when designing machine learning models.