PREDICTIVE MODELLING TO OPTIMIZE ABC – FLOORING MASTERS BUSINESS

Utilizing the power of data mining to support business problems

GROUP - 23

PRACHI PARAG DESHKAR - 210237944

GAGAN SACHETH SHETTY - 210197062

TABLE OF CONTENTS

INTRODUCTION

The problem

ABC flooring masters are a company specializing in reflooring of houses. They are looking to optimize their business by understanding their customer base & the amount they pay while they buy their house so that premium segment of flooring options such as natural stone flooring, exotic hardwood flooring etc. & cheap set of options such as sheet vinyl, linoleum etc. can be offered based on their spend appetite & carpet size optimization for various ranges. For example, logical assumption is someone who spends 100,000 would find it affordable to spend 10% of the sale price on flooring which is 10,000 & find a suitable offering at this price range for respective size of their & optimize the size of the carpet for bulk.

Stakeholders and their necessity

- ABC – flooring masters – Investors and board

  Primary stakeholder – Need for optimizing flooring business
- Carpet manufacturing unit

  Create tooling based on right size for different types of flooring
- Sales and marketing department

  Pitch the right type of flooring based on customer type
- Customers

  Indirect stakeholders – To be satisfied with the options offered by ABC

EXPLORATORY DATA ANALYSIS

Type of the Dataset

It is a cross-sectional data as we observe a snapshot of various properties(subjects) at one period.

Dimensions of the Dataset

The dataset consists of 11 variables – (10 predictors, 1 target) and 1145 records

Variables-

The predictor variables are selected based on the flooring company.

Variables	Definitions	Types	Role
SalePrice	Selling Price of the property	Continuous	Target
LotArea	Lot size in square feet	Continuous	Predictor
OverallQual	Rates overall material and finish of the house	Ordinal	Predictor
BsmtFinSF1	Type 1 finished square feet	Continuous	Predictor
TotalBsmtSF	Total square feet of basement area	Continuous	Predictor
1stFlrSF	First Floor square feet	Continuous	Predictor
GrLivArea	Above grade (ground) living area square feet	Continuous	Predictor
KitchenAbvGr	Kitchens above grade	Ordinal	Predictor
OpenPorchSF	Open porch area in square feet	Continuous	Predictor
KitchenQual	Kitchen quality	Categorical	Predictor
YearBuilt	Original construction date	Ordinal	Predictor

Verbal Presentation

Considering record 7 (Property/house id= 7)

This property is of 10382 sqft which is old as it was built in 1973. It has 1 kitchen which is of typical/average quality and an open porch of 204 sqft. The above ground area is of 2090 sqft with a first floor of area 1107 sqft. The house has a basement of 1107 sqft with an 859 sqft of type 1 finished basement. This property was sold for $200000

Level of Data

Each record represents the variables & attributes related to a housing property.

Univariate Analysis

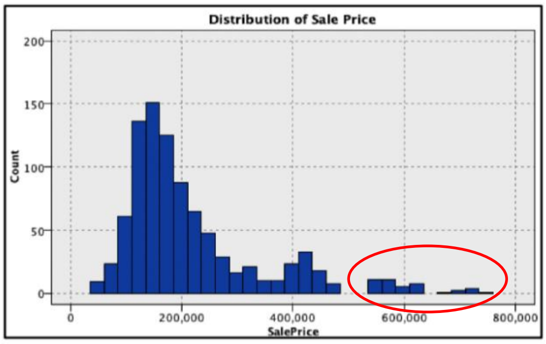
Continuous variables

1) SalePrice- Selling Price of the property

Interpretation-

By visualising the data using histogram we can spot the outliers. We also observe, the huge variation and moderate std. deviation. We see that the curve is right skewed and leptokurtic. We can verify these characteristics of the curve through the values of the statistical measures.

Statistical Measures	Values
Central Tendency	Mean = 217334.656 Median = 176000 Mode = 140000
Dispersion	Std Deviation = 128173.740 Range = 720100 Coeff of Variation= 58.975%
Skewness & Kurtosis	Skewness = 1.657 Kurtosis = 2.561
Visualization-	

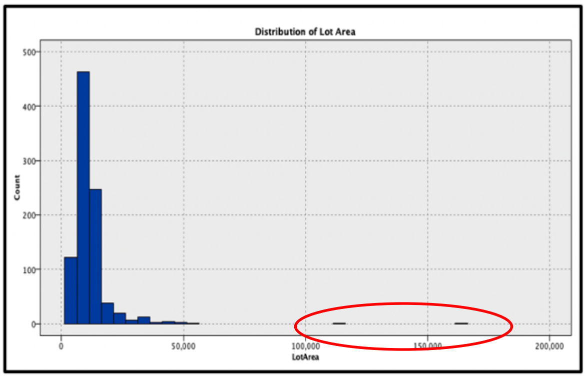


2) LotArea- Lot area in sqft.

Interpretation-

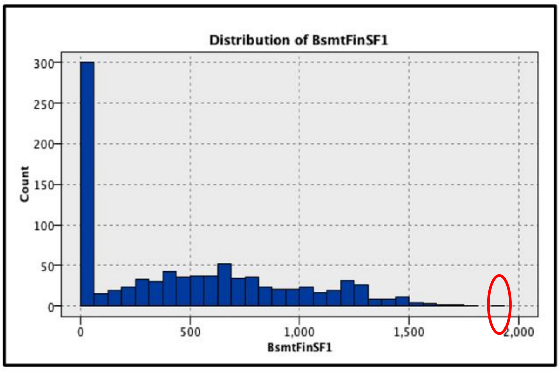
The outliers are at a significant distance and are prominent. The standard deviation is low, and the variation is high. The kurtosis and skewness values are high specifying that the curve is platykurtic and right-skewed which can be verified from the graph. Even though the range is high, majority of data lies between 0 to 50,000 sqft.

Statistical Measures	Values
Central Tendency	Mean = 11284.115 Median = 10005 Mode = 7200
Dispersion	Std Deviation = 8571.678 Range = 163360 Coeff of Variation= 75.96%
Skewness & Kurtosis	Skewness = 9.055 Kurtosis = 136.531
Visualization-	



Statistical Measures	Values
Central Tendency	Mean = 489.305 Median = 428 Mode = 0
Dispersion	Std Deviation = 459.845 Range = 1880 Coeff of Variation= 93.97%
Skewness & Kurtosis	Skewness = 0.567 Kurtosis = -0.745

Visualization-



3) BsmtFinSF1- Type 1 finished square feet

Interpretation-

We observe the significantly high frequency of no basements(mode) and minimal outliers. The spread of data is huge with high deviation from the mean and huge variation. The curve is right-skewed and leptokurtic.

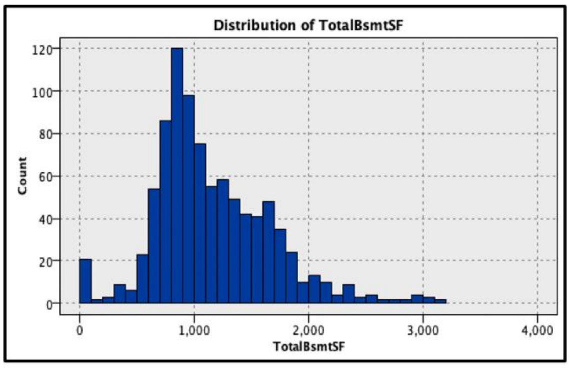
4) TotalBsmtSF- Total square feet of basement area

Interpretation-

We observe a spread from 0 to 3200 with no traces of outliers in the histogram. We observe moderate deviation and variation. The curve is right skewed and leptokurtic.

Statistical Measures	Values
Central Tendency	Mean = 1161.149 Median = 1051 Mode = 0
Dispersion	Std Deviation = 516.589 Range = 3200 Coeff of Variation= 44.49%
Skewness & Kurtosis	Skewness = 0.807 Kurtosis = 1.430

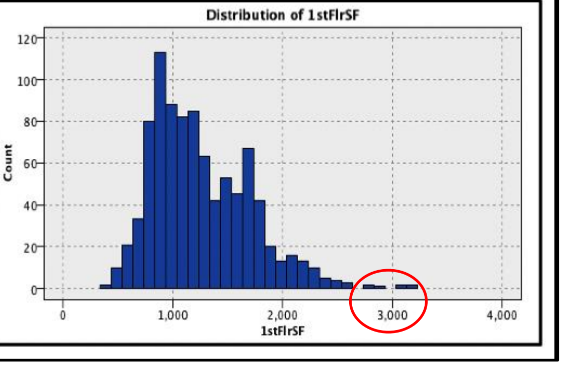
Visualization-



5) 1stFlrSF- First Floor square feet

Statistical Measures	Values
Central Tendency	Mean = 1258.359 Median = 1158 Mode = 864
Dispersion	Std Deviation = 454.731 Range = 2894 Coeff of Variation= 36.13%
Skewness & Kurtosis	Skewness = 0.914 Kurtosis = 0.942

Visualization-



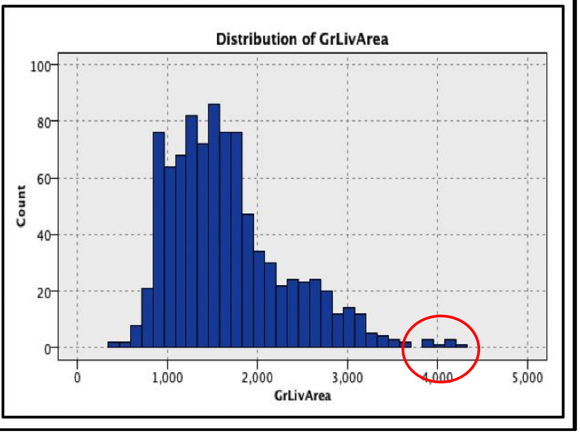
Interpretation-

The majority records lie between 334 to 2800 and we observe some number of outliers as well. The deviation from mean is low and so is the variation. The curve is right-skewed and leptokurtic.

6) GrLivArea- Above grade (ground) living area square feet

Statistical Measures	Values
Central Tendency	Mean = 1668.256 Median = 1552 Mode = 864
Dispersion	Std Deviation = 661.353 Range = 3982 Coeff of Variation= 39.64%
Skewness & Kurtosis	Skewness = 1.003 Kurtosis = 0.942

Visualization-

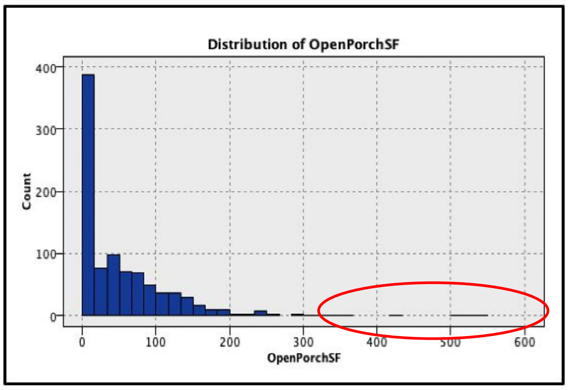


Interpretation-

The majority records lie between 334 to 3800 and we observe some number of outliers as well. We observe high deviation and low variation. The curve is right-skewed and leptokurtic.

7) **OpenPorchSF-** Open porch area in square feet

Statistical Measures	Values
Central Tendency	Mean = 50.739 Median = 32 Mode = 0
Dispersion	Std Deviation = 66.700 Range = 547 Coeff of Variation= 44.49%
Skewness & Kurtosis	Skewness = 2.440 Kurtosis = 10.282
Visualization-	



Interpretation-

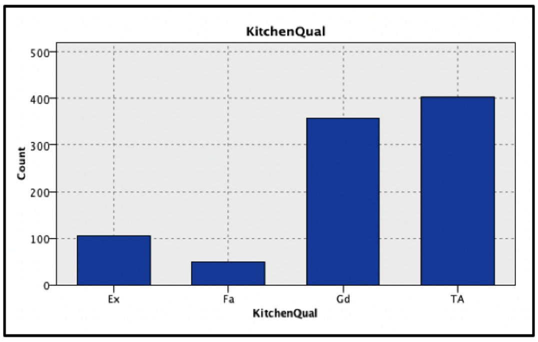
We observe huge traces of outliers. The frequency of 0 is high specifying that most of the properties don’t have an open porch.

The spread is huge with high deviation but low variation. The curve is highly right skewed and platykurtic.

Nominal variable

KitchenQual- Kitchen quality

Statistical Measure	Value
Mode	TA
Visualization-	

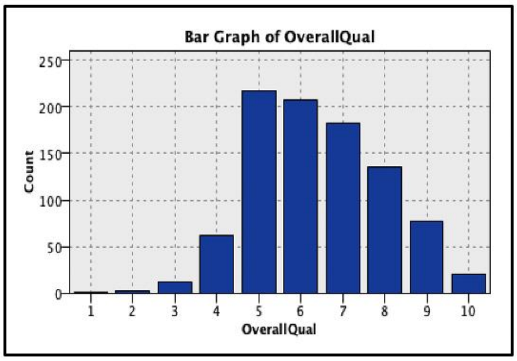


The mode is TA specifying that the frequency of Typical/Average quality kitchen are high followed by good (Gd) excellent (Ex) and Fa (Fair). There are no poor-quality kitchens.

Ordinal variables

1) OverallQual- Rates the overall material and finish of the house

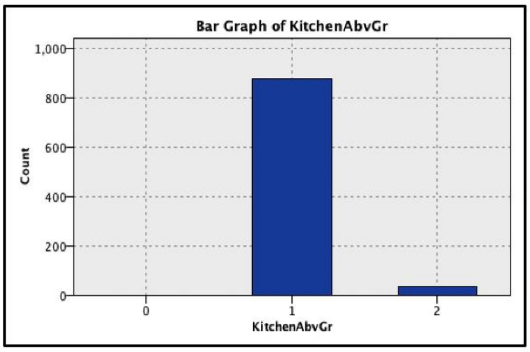
Statistical Measure	Value
Mode	5
Visualization-	



The mode is 5 specifying that the overall quality of most of the properties is average. There are a smaller number of poor-quality properties and a greater number of average and good quality properties.

2) KitchenAbvGr- Kitchens above grade

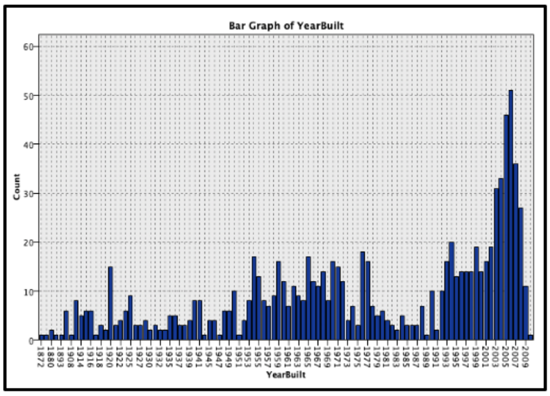
Statistical Measure	Value
Mode	1
Visualization-	



The mode is 1 specifying that the frequency of having one kitchen above ground is high whereas having no kitchen is absurd due to which there is no frequency of 0.

3) **YearBuilt**- Original construction date

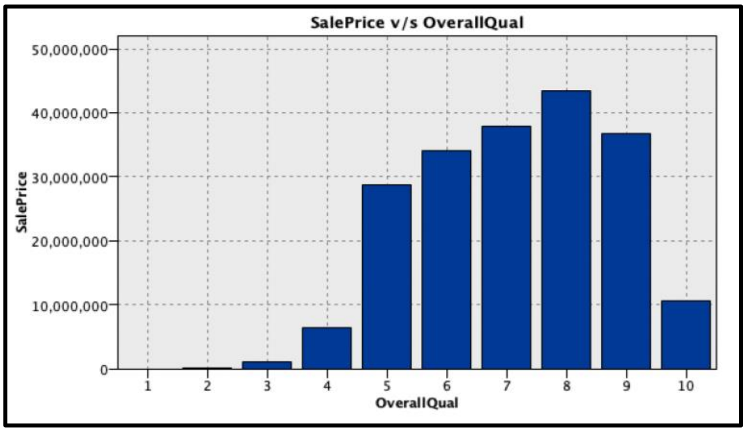
Statistical Measure	Value
Mode	2006
Visualization-	



We observe that maximum properties were built in the year 2006 and very less properties were built in the early years and there was a sudden increase from the year 1993.

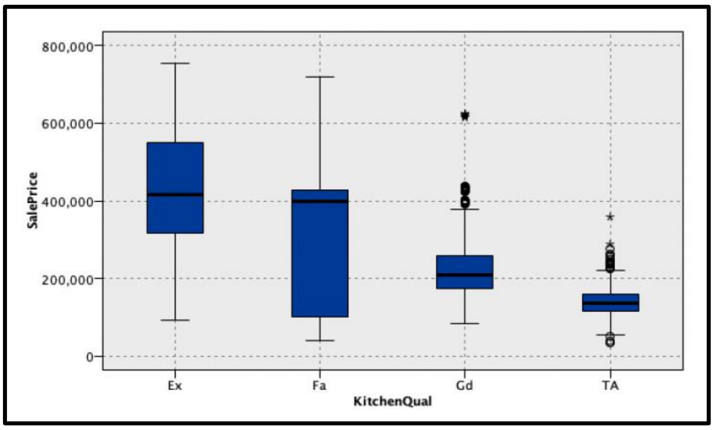
Bivariate Analysis (Target v/s Predictor)

1) SalePrice v/s OverallQual



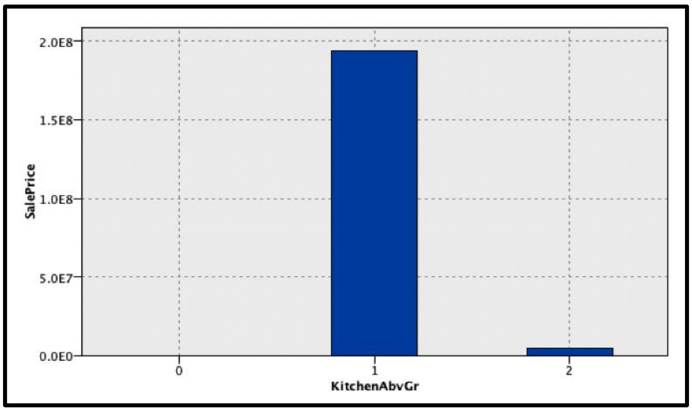
We can observe the relation between the quality rating and the properties with overall quality between 5-9 are expensive whereas the ones with low quality are cheaper. Although some of the cheap houses have high quality.

2) SalePrice v/s KitchenQual



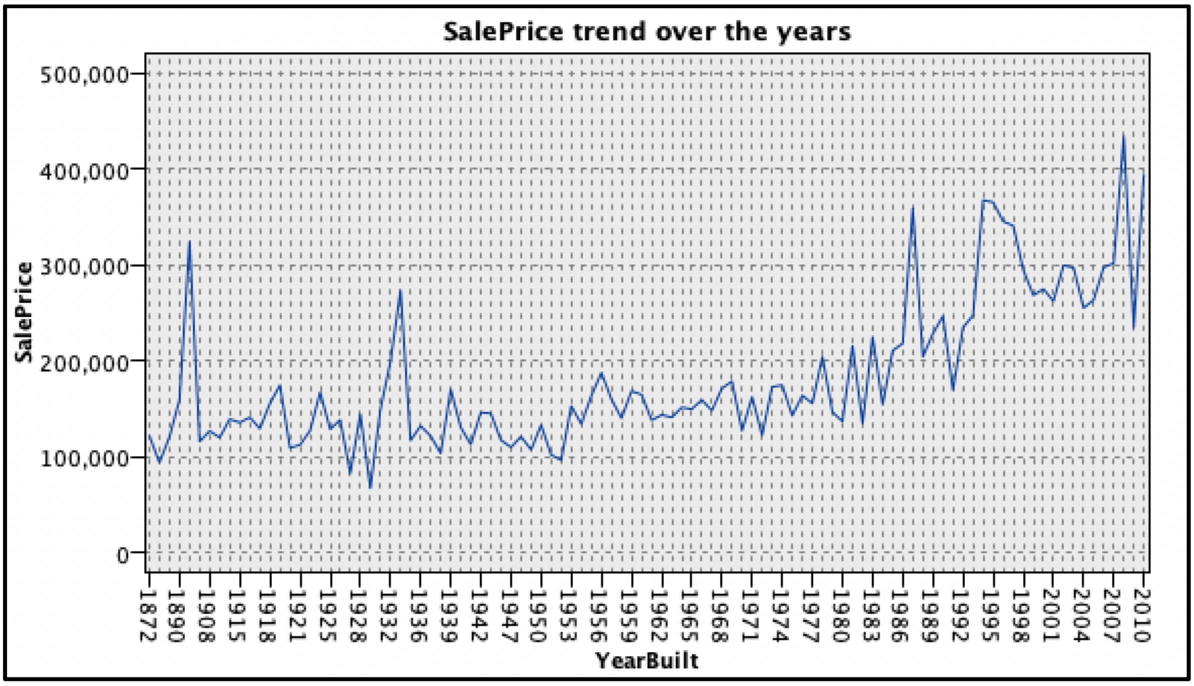
We can interpret outliers from the good and typical/average categories of KitchenQual with regards to the SalePrice. The spread of excellent and fair quality categories is high with regards to their SalePrice.

3) SalePrice v/s KitchenAbvGr



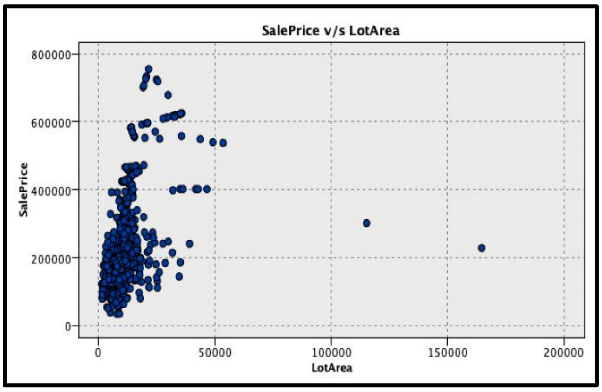
As the frequency of 1 kitchen above ground was high, we can see that their Sale price is also high.

4) SalePrice v/s YearBuilt.



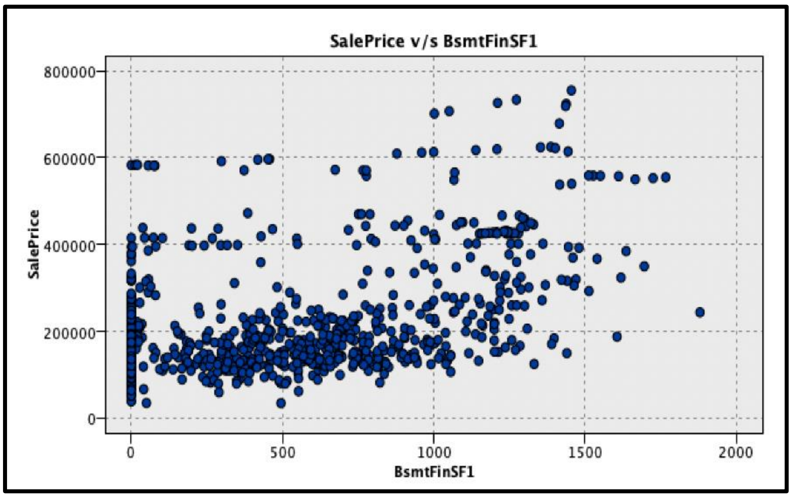
We fit a line to observe the variation in the prices over the years and observe non-uniform trend. We see that the prices had no significant increase between 1936-1980 and started increasing from the year 1986.

5) SalePrice v/s LotArea



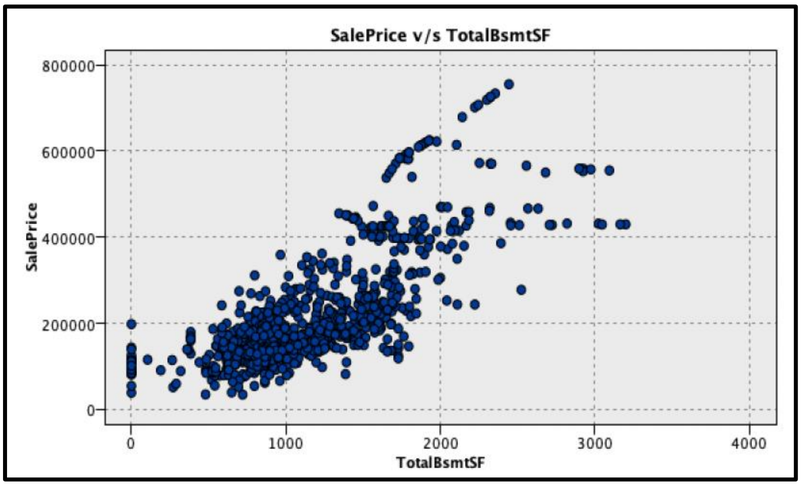
Majority of the properties have a lot area between 1300-60000 sqft & we can observe +ve correlation in that interval where the prices increase as lot area increases. We can see two outliers where the prices are low even when the area is huge.

6) SalePrice v/s BsmtFinSF1



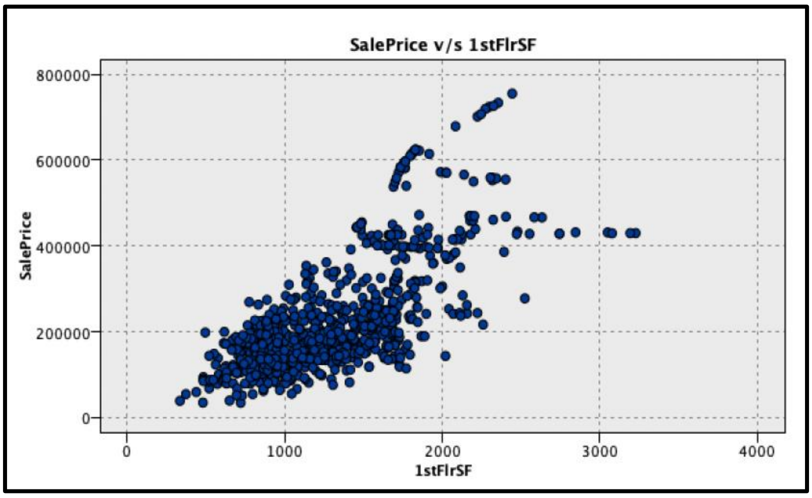
We can observe a low positive correlation as the even without a basement some properties have high prices but also as the basement finish increases the prices also increase.

7) SalePrice v/s TotalBsmtSF



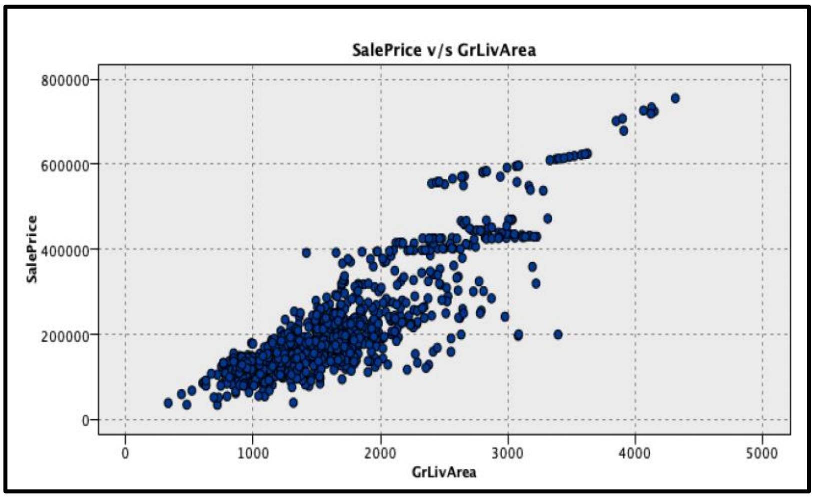
We can observe a prominent positive correlation between Total Basement Area and the SalePrice.

8) SalePrice v/s 1stFlrSF



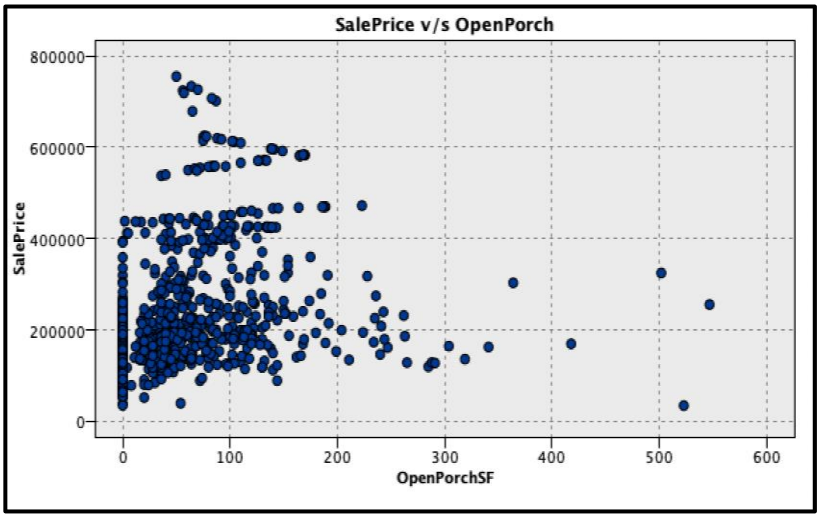
We can see an outward funnel pattern through which we can say that there exists positive correlation but there are some observations where the first-floor area is high, but the price is moderate.

9) SalePrice v/s GrLivArea



We can observe a prominent positive correlation between above ground living area and the SalePrice.

10) SalePrice v/s OpenPorchSF

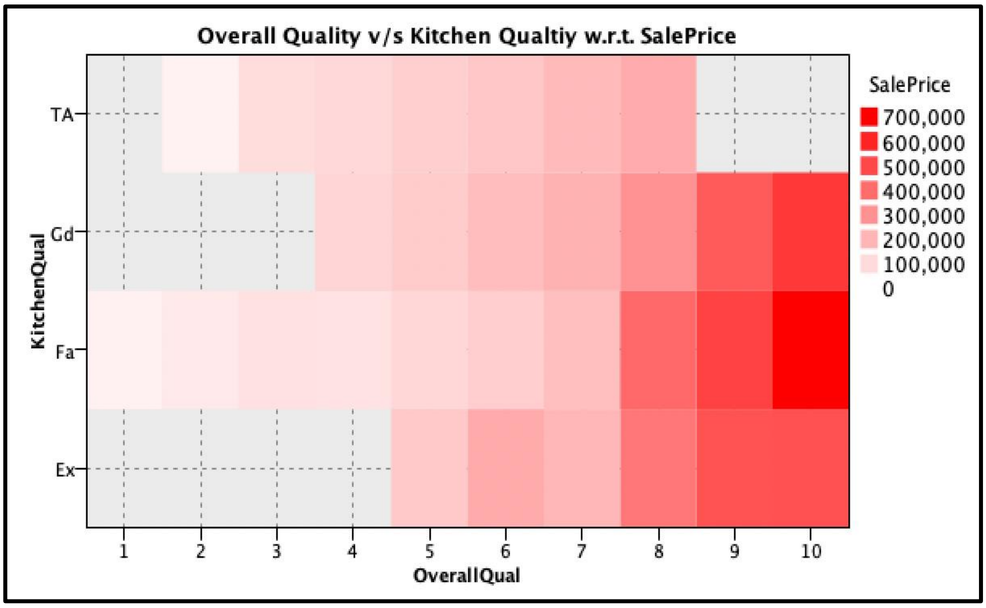


We observe a relatively low positive correlation between these two variables as even the prices of properties with no open porch are high and for some properties with high porch area the prices are low.

Multivariate Analysis

Heatmap of Kitchen Quality and Overall Quality with regards to the Sale Price

As we saw that the kitchen quality and overall quality influences the Price of the property, we construct a heatmap where the tints of colours represent the range of prices. We can see that there is a positive correlation between kitchen quality and overall quality. Hence, when the kitchen quality is high, the overall quality is also high and so is the Sale Price of the property.



Data quality assessment and treatment

Outliers and extremes

Outliers: are the values of a variable which are distant from the majority (crowd) of the values. Outliers are generally calculated based on IQR, z-score (using std.dev) etc.

Extremes: The values which are more severe than the outliers, which are too distant from the majority mostly towards the end of range.

These quality issues maybe due to real life scenario or ease of analysis

Treatment: Log transformation, Nullify and imputing with C&R tree algorithm

Missing values

The values in a column which do not consist of any values, which may be due to error in data collection or more reason, this issue may be due to real life scenario

Treatment: Impute using C&R tree algorithm

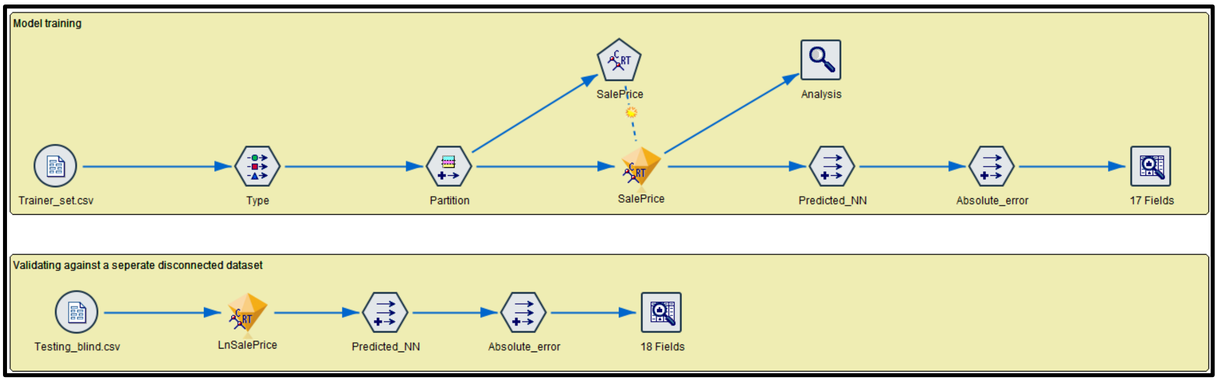
Anomalies

These are the cases which are far from the majority of the cases, which may negatively impact the performance of analysis.

Treatment: Identify using anomaly node and discard anomalous values

Handling data leakage

1) Validation dataset is created at the initial stage before modelling without the presence of target variable, so that validation test is conducted in a non-biased environment with no possibility of data leakage.



2) No external information was provided to the model training flow, to ensure non - biased learning

MODELLING

Predictive modelling formulation

We want to predict the Sale price of the houses using the predictor variables with the ML method of supervised learning, which will discover the relationship between the predictor variables and the Sale price (Target) and will further be able to predict the sale price for any new house.

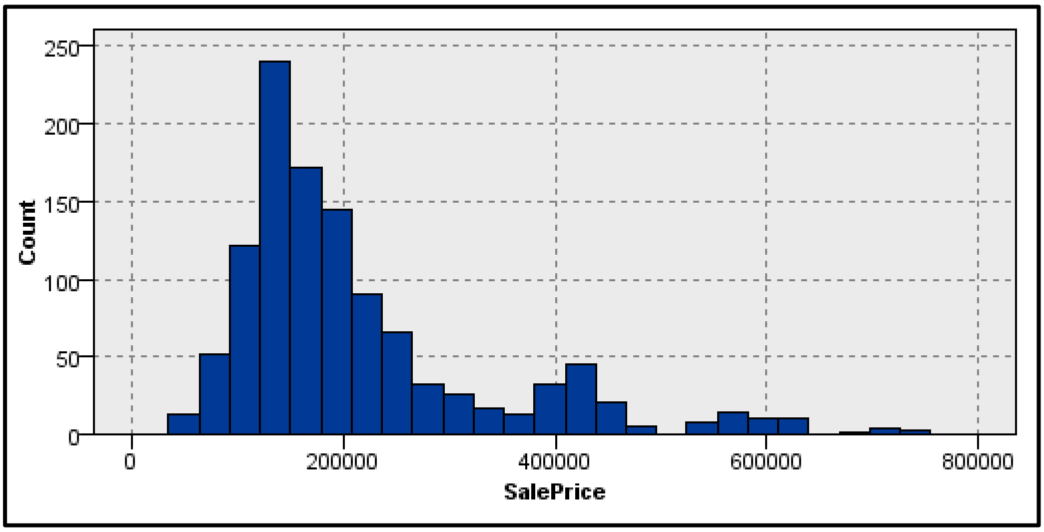
Type of the problem

As our target variable-SalePrice is continuous we have a regression problem.

Target variable assessment and treatment

Target variable: Sale price

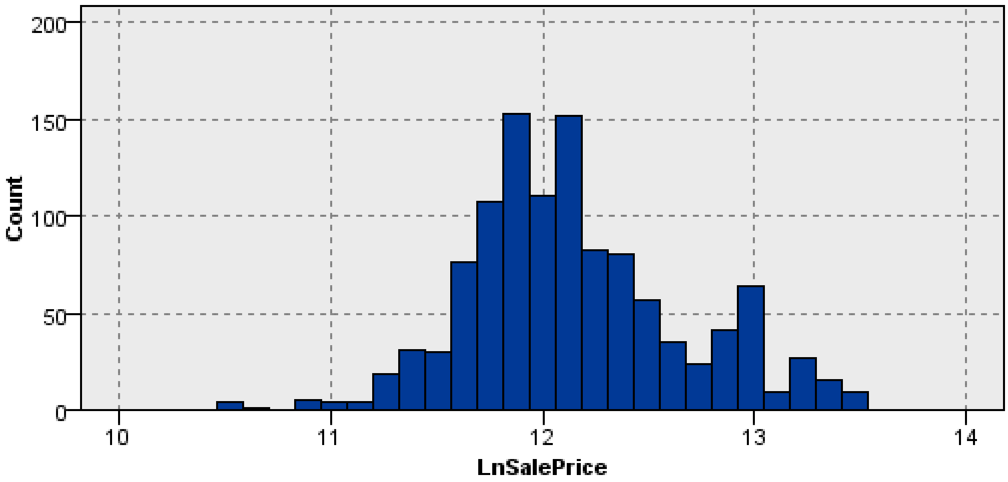
Assessment: Histogram to understand the spread of the data



Interpretation: Right skewed histogram, with large range of values. This is not ideal since there is an absence of symmetry in the distribution and there is an indication of outliers at the higher end of the sale price. Performance may not be good for this model as symmetry is preferred and outliers may distort the model understanding

Treatment:

1. Log transform – Performing this transformation normalizes skewed distributions. Also, the outliers may also be nullified when found in natural log. Distribution is as below



This is a good & symmetric distribution, which is better for predictive performance, and we can also see the range has been reduced which is good since this shows some of the outliers were handled.

2) Outlier management for target variable in training set: This step is essential for training model as this will help the model to understand the data better. Steps followed are

1. Identify outliers post log transformation

2. Nullify outliers
3. Impute null using C&R tree

Need for partitioning data

We partition our data into training and testing data

1) For accurate evaluation of our model.
2) To understand and avoid overfitting.

Performance metric

For the regression problem we can use MAE, MSE, RMSE, execution time, R-squared (goodness of fit).

Here we are selecting:

Execution Time**:** For a model which is constantly updating, we consider whether it is time efficient as it can cost our business.

Difference between MAE of train and test**:** For understanding if our model is overfitting.

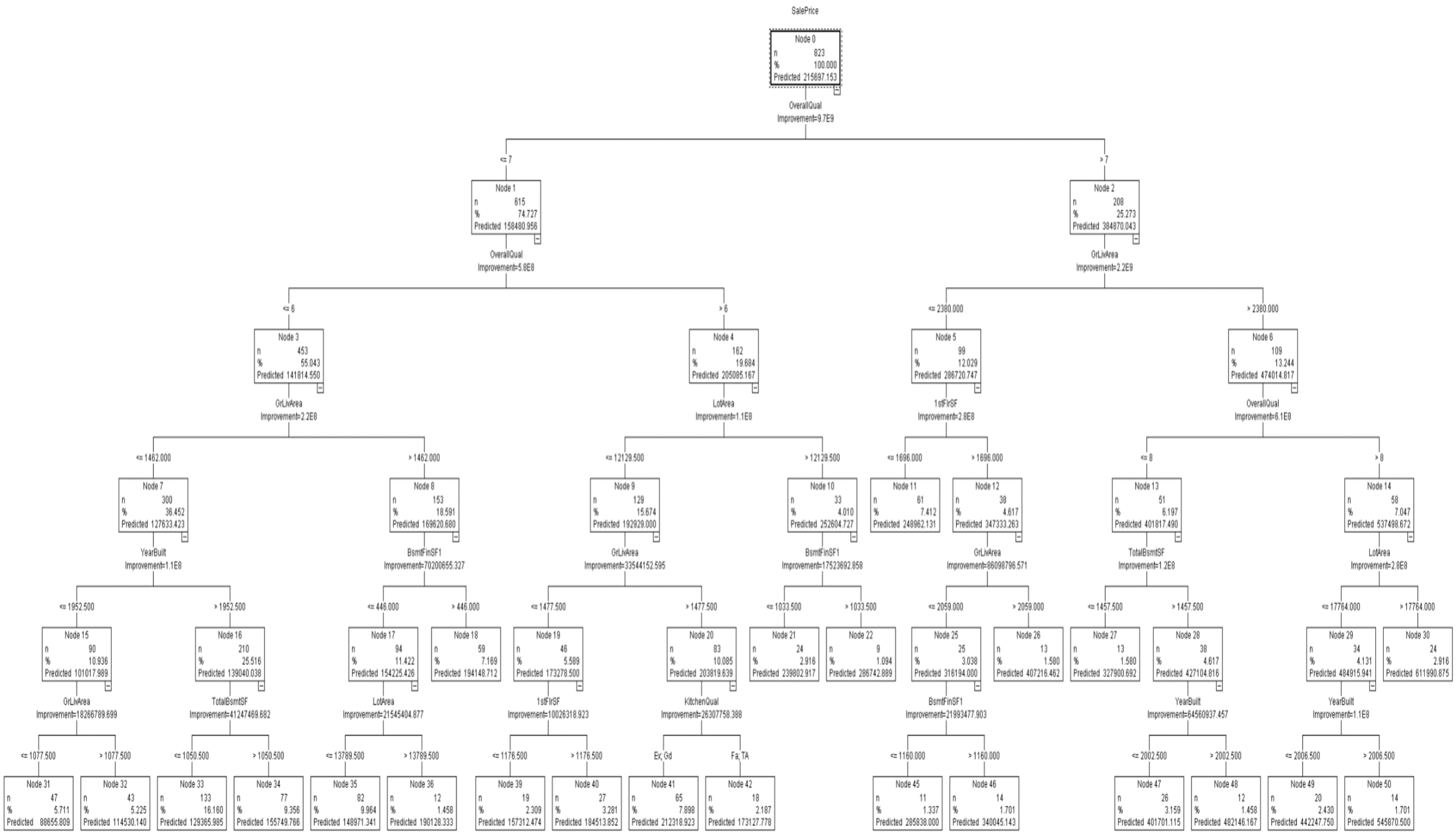
Mean Absolute Error**:** The average of the absolute difference between predicted value and the actual value

Metric selection**:** MAE is a metric which treats any amount of error alike, unlike RMSE which penalizes large error more as there is squaring, for the considered business case which estimates probable spend appetite based on predictors we don't want a perfectionist model. The impact due to this huge error is not significant to the company, where they can simply offer a different carpet type and a certain carpet type can also satisfy a range of sale prices.

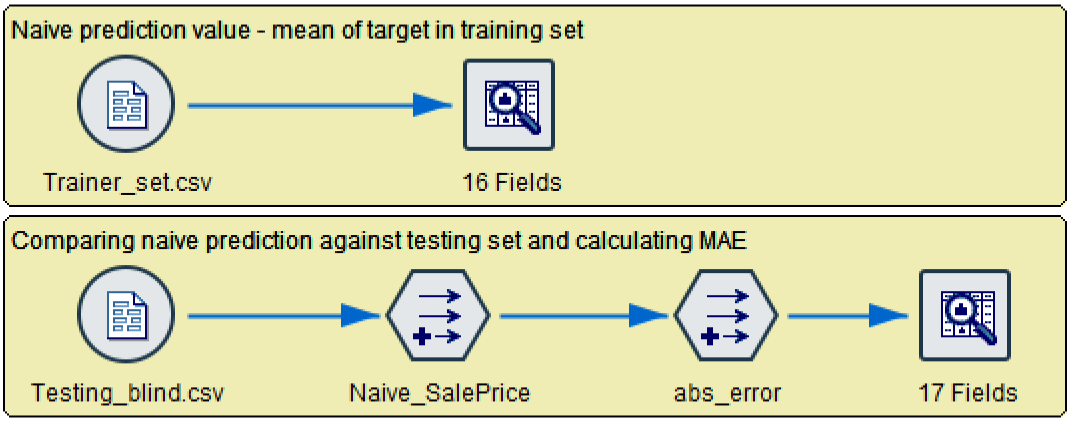
Using predictive modelling for data exploration

Using Decision tree (Regression tree) we visually represent the decisions using branches, nodes and leaves to find the region with higher purity of target value.

We have some stakeholders who are not techno savvy and a decision tree will benefit them for interpretation. In this case we can see the root node (target variable) branched out with regards to the Overall Quality split at the rating value of 7 and have divided the data in 72% and 27%, further living room area can be seen to split the dataset with the value 2380 sq. ft Hence, we can follow the tree using the values of the predictor variables and predict the sale price by observing the value in that particular node.


SalePrice
Node 0
n 823
% 100.000
Predicted 215697.153
OverallQual
Improvement=9.7E9
<= 7
> 7
Node 1
n 615
% 74.727
Predicted 158480.956
Node 2
n 208
% 25.273
Predicted 384870.043
OverallQual
Improvement=5.8E8
GrLivArea
Improvement=2.2E9
<= 6
> 6
<= 2380.000
> 2380.000
Node 3
n 453
% 55.043
Predicted 141814.550
Node 4
n 162
% 19.684
Predicted 205085.167
Node 5
n 99
% 12.029
Predicted 286720.747
Node 6
n 109
% 13.244
Predicted 474014.817
GrLivArea
Improvement=2.2E8
LotArea
Improvement=1.1E8
1stFlrSF
Improvement=2.8E8
OverallQual
Improvement=6.1E8
<= 1462.000
> 1462.000
<= 12129.500
> 12129.500
<= 1696.000
> 1696.000
<= 8
> 8
Node 7
n 300
% 36.452
Predicted 127633.423
Node 8
n 153
% 18.591
Predicted 169620.680
Node 9
n 129
% 15.674
Predicted 192929.000
Node 10
n 33
% 4.010
Predicted 252604.727
Node 11
n 61
% 7.412
Predicted 248962.131
Node 12
n 38
% 4.617
Predicted 347333.263
Node 13
n 51
% 6.197
Predicted 401817.490
Node 14
n 58
% 7.047
Predicted 537498.672
YearBuilt
Improvement=1.1E8
BsmtFinSF1
Improvement=70200655.327
GrLivArea
Improvement=33544152.595
BsmtFinSF1
Improvement=17523692.858
GrLivArea
Improvement=86098796.571
TotalBsmtSF
Improvement=1.2E8
LotArea
Improvement=2.8E8
<= 1952.500
> 1952.500
<= 446.000
> 446.000
<= 1477.500
> 1477.500
<= 1033.500
> 1033.500
<= 2059.000
> 2059.000
<= 1457.500
> 1457.500
<= 17764.000
> 17764.000
Node 15
n 90
% 10.936
Predicted 101017.989
Node 16
n 210
% 25.516
Predicted 139040.038
Node 17
n 94
% 11.422
Predicted 154225.426
Node 18
n 59
% 7.169
Predicted 194148.712
Node 19
n 46
% 5.589
Predicted 173278.500
Node 20
n 83
% 10.085
Predicted 203819.639
Node 21
n 24
% 2.916
Predicted 239802.917
Node 22
n 9
% 1.094
Predicted 286742.889
Node 25
n 25
% 3.038
Predicted 316194.000
Node 26
n 13
% 1.580
Predicted 407216.462
Node 27
n 13
% 1.580
Predicted 327900.692
Node 28
n 38
% 4.617
Predicted 427104.816
Node 29
n 34
% 4.131
Predicted 484915.941
Node 30
n 24
% 2.916
Predicted 611990.875
GrLivArea
Improvement=18266789.699
TotalBsmtSF
Improvement=41247469.682
LotArea
Improvement=21545404.877
1stFlrSF
Improvement=10026318.923
KitchenQual
Improvement=26307758.388
BsmtFinSF1
Improvement=21993477.903
YearBuilt
Improvement=64560937.457
YearBuilt
Improvement=1.1E8
<= 1077.500
> 1077.500
<= 1050.500
> 1050.500
<= 13789.500
> 13789.500
<= 1176.500
> 1176.500
Ex; Gd
Fa; TA
<= 1160.000
> 1160.000
<= 2002.500
> 2002.500
<= 2006.500
> 2006.500
Node 31
n 47
% 5.711
Predicted 88655.809
Node 32
n 43
% 5.225
Predicted 114530.140
Node 33
n 133
% 16.160
Predicted 129365.985
Node 34
n 77
% 9.356
Predicted 155749.766
Node 35
n 82
% 9.964
Predicted 148971.341
Node 36
n 12
% 1.458
Predicted 190128.333
Node 39
n 19
% 2.309
Predicted 157312.474
Node 40
n 27
% 3.281
Predicted 184513.852
Node 41
n 65
% 7.898
Predicted 212318.923
Node 42
n 18
% 2.187
Predicted 173127.778
Node 45
n 11
% 1.337
Predicted 285838.000
Node 46
n 14
% 1.701
Predicted 340045.143
Node 47
n 26
% 3.159
Predicted 401701.115
Node 48
n 12
% 1.458
Predicted 482146.167
Node 49
n 20
% 2.430
Predicted 442247.750
Node 50
n 14
% 1.701
Predicted 545870.500

Baseline model

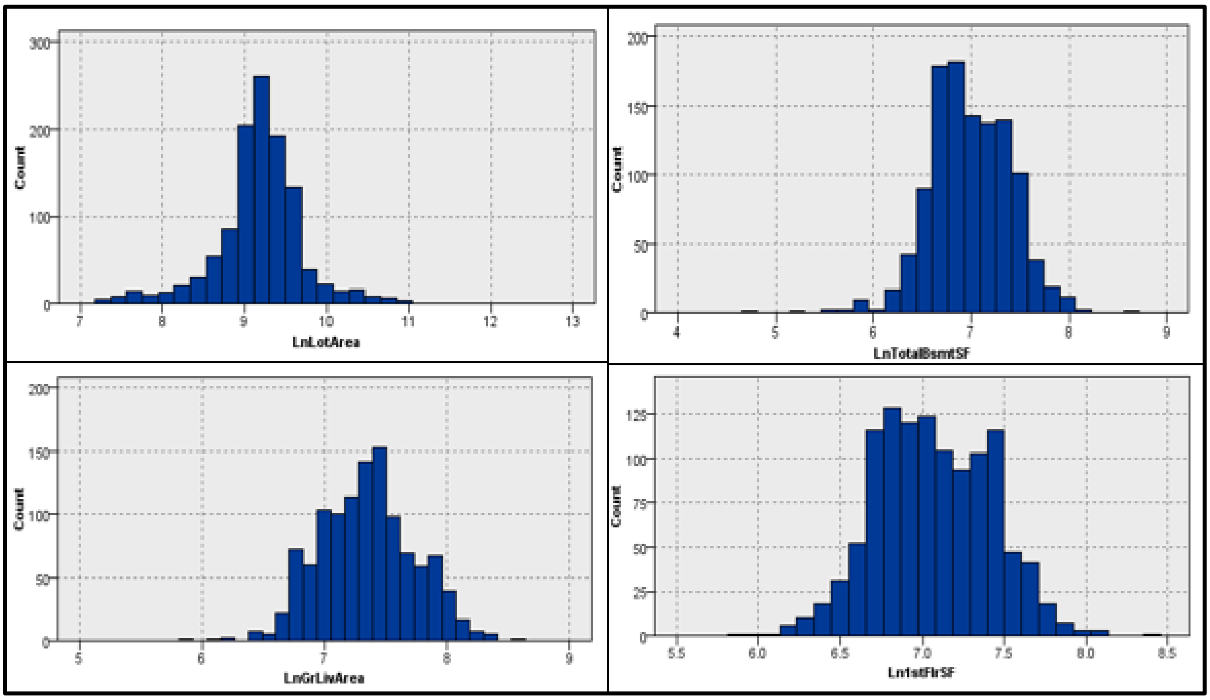


Naïve model is considered as baseline for this project. As explained above since this is a classification problem. The Naïve model prediction for the sale price is equal to the mean of the target variable column i.e., 217252.844**.** Applying this value in the testing dataset and finding the MAE value would yield the result of **90296.869**.

Our best model is performing 20200.882 which is almost 78% improvement from the baseline.

Feature engineering

Log transformation **-** As described in descriptive analytics section, all the variables are found to be skewed. Hence, log transformation can help obtain normal distribution for the continuous variables. There was around 10% improvement in the model experienced due to this



Imputation of missing values using algorithm

This step as explained above is done post nullification of outliers, using algorithm ensures imputation of the closest value possible. The model seemed to improve by around 5%.

Converting kitchen quality from nominal to ordinal

This step involved converting the ordinal values ranging from poor to excellent with numerical values, 0 – 5. Slight improvement of around 2% was found with implementation of this step.

Converting year built to age of the house

This step involved subtracting the max year of the column with all the values which would give us the age of the house, this seemed a logical choice since age of the house can have an impact on the sale price and the hunch was found to be right, with an improvement of around 7-8% in the model.

Division of log - lot area with log – liv area

This step involved taking ratio of log – lot area with log – liv area as area of living area is a proportion of lot area. Small improvement over existing of around 2-3% was found.

Division of log – total basement area with log – type 1 finished basement area

This step involved taking ratio of log – total basement area with log – type 1 finished basement area of living area is a proportion of lot area. Considerable improvement over existing of around 10% was found.

Models summary and evaluation

Neural network - the best performing model

Intro to model

Neural network consists of input layer, one or more hidden layer and an output layer. The inputs are taken in with weights and bias, after transforming through activation function (ReLu, Sigmoid etc.) optimum values for weights are found and passed to output.

Evaluation Summary table

Model	Hyperparameters	MAE - Train	MAE - Test	Time taken	MAE Training	MAE Testing blind	Diff - MAE (Train - Test)	Remarks
Neural network	Default	0.078	0.103	< 1 second	17614.645	20200.88	2586.237	No signs of overfitting, good model
Neural network	Boosting	0.041	0.103	3 seconds	11011.76	18108.75	7096.990	Heavy overfitting as expected with boosting due to emphasis on accuracy in training, also time intense
Neural network	Bagging	0.052	0.091	2 seconds	12137.85	19495.27	7357.415	Heavy overfitting not as much as boosting, increasing component models might help stabilize
Neural network	Radial Basis Function	0.159	0.176	< 1 seconds	39961.26	40935.44	974.185	MAE value too high, no signs of overfitting
Neural network	Hidden layer 1: 15 units Hidden layer 2: 15 units	0.085	0.107	< 1 seconds	19139.14	21483.81	2344.670	Very good model performance against blind training set, similar to default model
Neural network	Number of component models for Bagging: 50	0.05	0.088	5 seconds	11828.26	17796.29	5968.030	Increasing component model helped stabilize the bagging model reflected by decrease in difference of MAE
Neural network	Number of component models for Bagging: 75	0.05	0.09	15 Seconds	11774.95	17575.19	5800.236	Increasing component model helped stabilize the bagging model reflected by decrease in difference of MAE, but time intense
Neural network	Number of component models for Boosting: 50	0.037	0.099	10 Seconds	10617.4	18348.57	7731.165	Increasing component model did not help stabilizing boosting model, this is not a good model for our scenario

The best model for our problem is highlighted above

Decision tree - the highly interpretable model

Intro to model

Decision tree creates branches based on conditions from various variables to lead a decision from root to leaf and end up with a value for regression problems. The constructed tree can be viewed to understand how the branches are split up and which variables are the key decision makers higher up in the branch

Evaluation Summary table

Model	Hyperparameters	MAE - Train	MAE - Test	Time taken	MAE Training	MAE Testing blind	Diff - MAE (Train - Test)	Remarks
Decision tree	Default	0.126	0.146	<1 second	25869.26	31396.49	5527.234	There is slight overfitting, and MAE is relatively high
Decision tree	Boosting	0.089	0.115	5 seconds	19542.13	23876.15	4334.019	Time intense but very good improvement in MAE values
Decision tree	Bagging	0.102	0.112	5 seconds	20854.99	23876.15	3021.161	each other showing low possibility of overfitting
Decision tree	Pruning turned off	0.126	0.146	<1 second	25788.11	31196.12	5408.013	High value of MAE
Decision tree	Max tree depth - 15	0.119	0.145	2 second	25105.81	30735.49	5629.682	High value of MAE
Decision tree	Max tree depth - 15, Prune off	0.116	0.141	2 second	24743.14	30774.88	6031.734	High value of MAE, with signs of overfitting, as expected with prune off
Decision tree	Stopping rule - 5% Parent, 3% Child	0.146	0.159	1 Second	32895.02	36629.66	3734.638	Highest value of MAE

Random forest - The worst performing model for dataset

Intro to model

Random forest is an ensemble method that utilizes multiple decision tree and ultimately aggregates from all trees based on central tendency measures. The samples are bootstrapped amongst the trees, which involves sample replication.

Evaluation Summary table

Model	Hyperparameters	MAE - Train	MAE - Test	Time taken	MAE Training	MAE Testing blind	Diff - MAE (Train - Test)	Remarks
Random forest	Default	0.04	0.105	5 Seconds	11519.65	40974.53	29454.882	Extreme overfitting found
Random forest	Number of trees - 20	0.033	0.104	5 Seconds	10135.07	44715.61	34580.546	Extreme overfitting found
Random forest	Max tree depth 15	0.03	0.107	2 Second	10822.2	46113.3	35291.096	Extreme overfitting found
Random forest	Use out of bag samples to estimate gen accuracy Extremely randomized tree	0.038	0.099	5 Seconds	10938.76	42917.1	31978.340	Extreme overfitting found
Random forest	Use out of bag samples to estimate gen accuracy Extremely randomized tree Hyper parameter optimization	0.019	0.092	240 second	7012.857	45267.89	38255.030	Extreme overfitting found

K-Nearest Neighbours (KNN) – The lazy learner

Intro to model

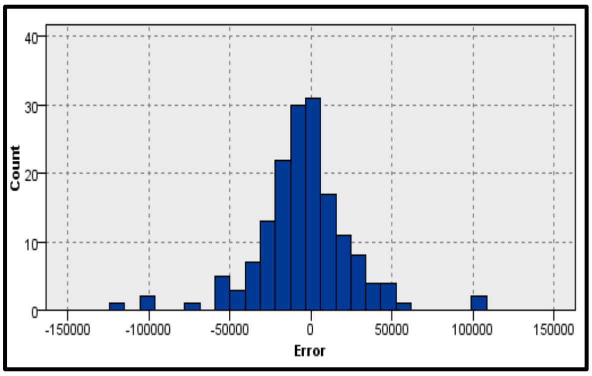
KNN method understands the behaviour of a data point based on its neighbours. This method memorizes the value based on the neighbour and spends effort finding appropriate neighbour in the entire set.

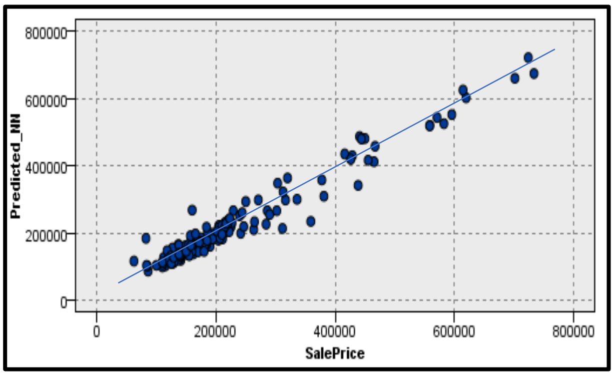
Evaluation Summary table

Model	Hyperparameters	MAE - Train	MAE - Test	Time taken	MAE Training	MAE Testing blind	Diff - MAE (Train - Test)	Remarks
K - Nearest Neighbors (KNN)	Default	0.09	0.141	2 seconds	22607.68	28868.69	6261.010	Not the best MAE, not very far from the best model
K - Nearest Neighbors (KNN)	Objective - Accuracy	0.123	0.184	2 seconds	22383.73	30034.8	7651.070	Slight overfitting found as expected
K - Nearest Neighbors (KNN)	Objective - Speed	0.072	0.141	1 Second	18764.34	26999.1	8234.753	Heavy overfitting found as a compromise for faster execution
K - Nearest Neighbors (KNN)	Automatic K selection, Min - 3, Max - 10	0.09	0.141	1 Second	22607.68	28868.69	6261.010	No change from default
K - Nearest Neighbors (KNN)	Automatic K selection, Min - 3, Max - 10 City - block metric	0.08	0.139	1 Second	20559.02	26935.88	6376.855	Performance improvement with change in calculation metric
K - Nearest Neighbors (KNN)	Automatic K selection, Min - 3, Max - 10 City - block metric Weight importance	0.081	0.137	1 Second	20627.53	26761.76	6134.226	Typical performance, no huge improvement with weight importance
K - Nearest Neighbors (KNN)	Automatic K selection, Min - 3, Max - 10 Euclidean metric Weight importance Cross validations - K = 20	0.085	0.135	1 Second	21288.12	28305.92	7017.794	Heavy overfitting found

Error cost analysis

To understand overestimation vs underestimation, we evaluate using histogram and scatterplot of predicted vs actual





We can see that error is mostly equally divided around 0 with slightly more density toward negative side showing slight underestimation.

With scatter we can see the outliers are relative less, the ones that are found seem to be above indicating some overestimation. We can say that both the errors are relatively very less in the model.

For the business scenario of the carpet floor company we can say that underestimation is a major pain where showing cheaper variant of flooring for someone who are potentially high paying client can lead to loss.

Relation of error values with predictors

Impact of predictor treatment with MAE

Data quality issue	**Treatment method**	**Impact on MAE**
Outliers	Nullify and impute with C&R tree algorithm	24224.864 – Non treated 18472.618 – Treated
Missing values	Impute with C & R tree algorithm	18472.618 – Non treated 16872.852 – Treated
Distribution	Log transformation	24224.864 – Non treated 18472.618 – Treated

Impact of predictor Transformation with MAE

Data quality issue	Treatment method	Impact on MAE
Distribution	Log transformation	18472.618 – Non treated 16872.852 – Treated

Impact of predictor selection with MAE

Feature importance based on model	MAE in Neural network
All features	23,283
C & R tree top 10	22,267
Neural network top 10	21,175

EXECUTIVE SUMMARY

The project aimed to create a predictive modelling to predict the sales price of the housing dataset provided, to help improve and optimize business for ABC flooring masters by understanding the impact of variables mostly related to carpet area with respect to sales price so that customer spend pattern can be better understood to target available varieties and sizes.

The steps involved include

1. Understanding the data central tendency, spread with an exploratory analysis
2. Based on the understanding, treating the errors and quality issues in data to make the process of data analysis better
3. Deriving new features based on existing features to improve the performance of the predictive analysis
4. Trying various available predictive options and comparing them with various performance metrics such as error and execution time to obtain the best model
5. Testing the performance of the prediction model against a separate dataset which simulates the real-life scenario by not having the sale price variable.

Finally, neural network was found to be the best predictive model. This was selected based on execution time, which was found to take less than 1 second, and a training mean absolute error of 17,615 with a variation of 2,600. This means for any value predicted sales price there is a possibility of error ranging from -20,200 to +20,200.

The next steps that can be taken include

- Collect more data to enhance model training performance
- Include more predictors
- Use more feature engineering to break the limitations imposed by model tuning

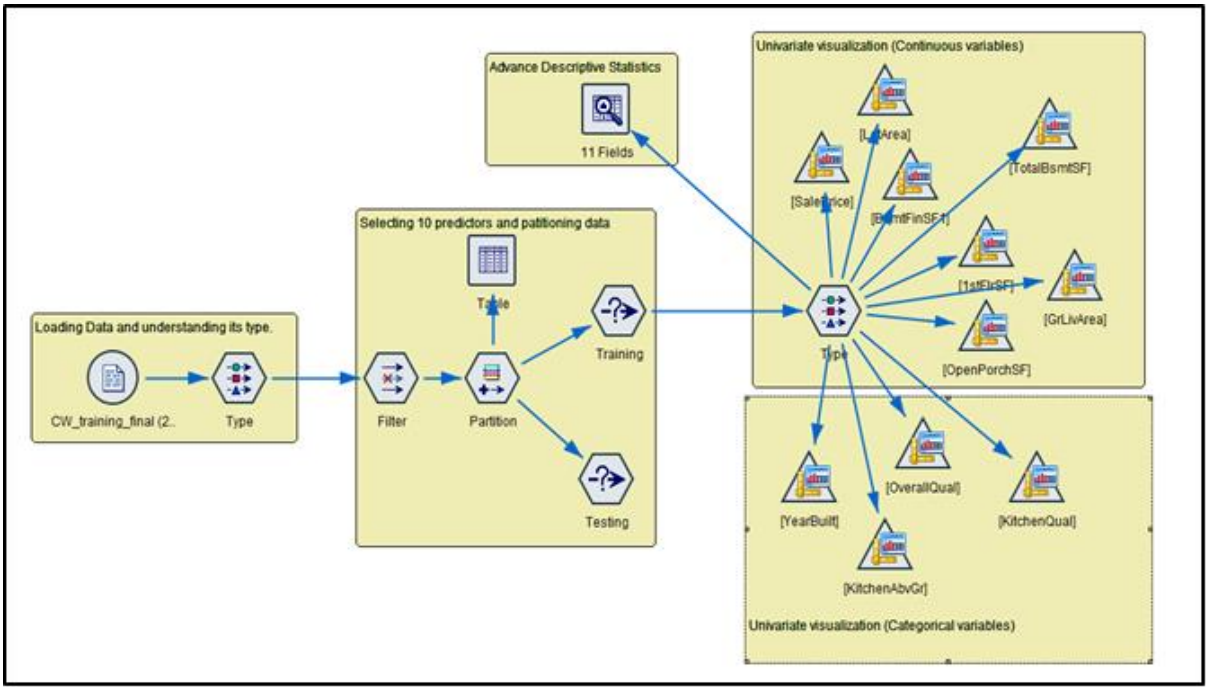
New ideas for relevant projects

- Perform predictive modelling on area to optimize flooring size better
- Understand predictive performance based on specific area in a house like kitchen, basement, living area etc.

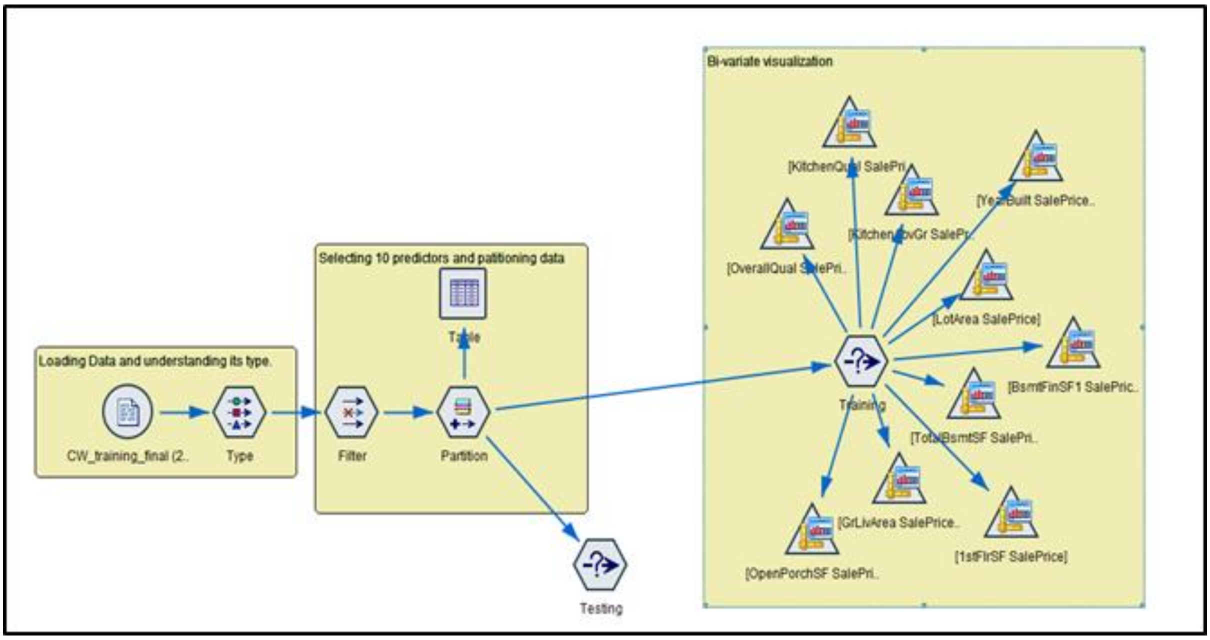
APPENDIX

EDA - Appendix

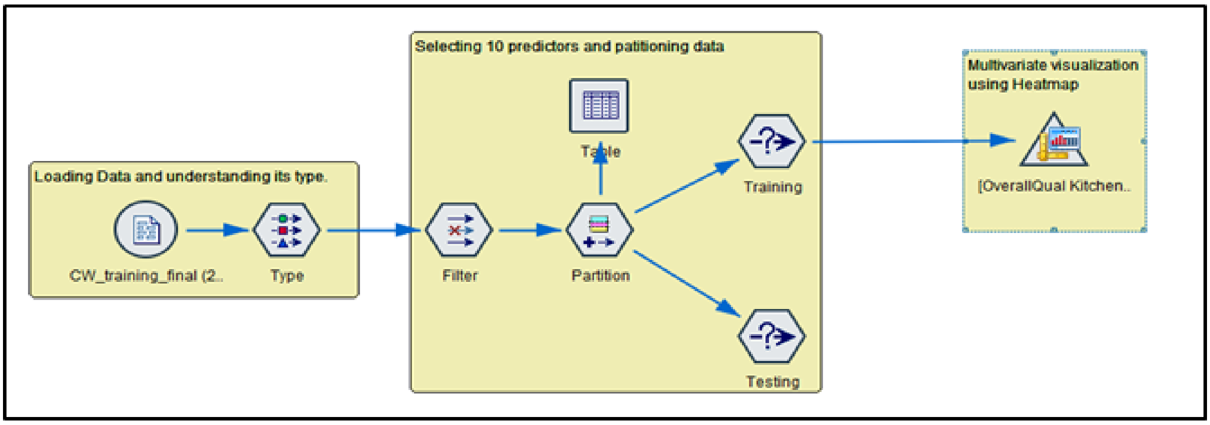
Appendix 1- Univariate Analysis



Appendix 2 – Bivariate Analysis

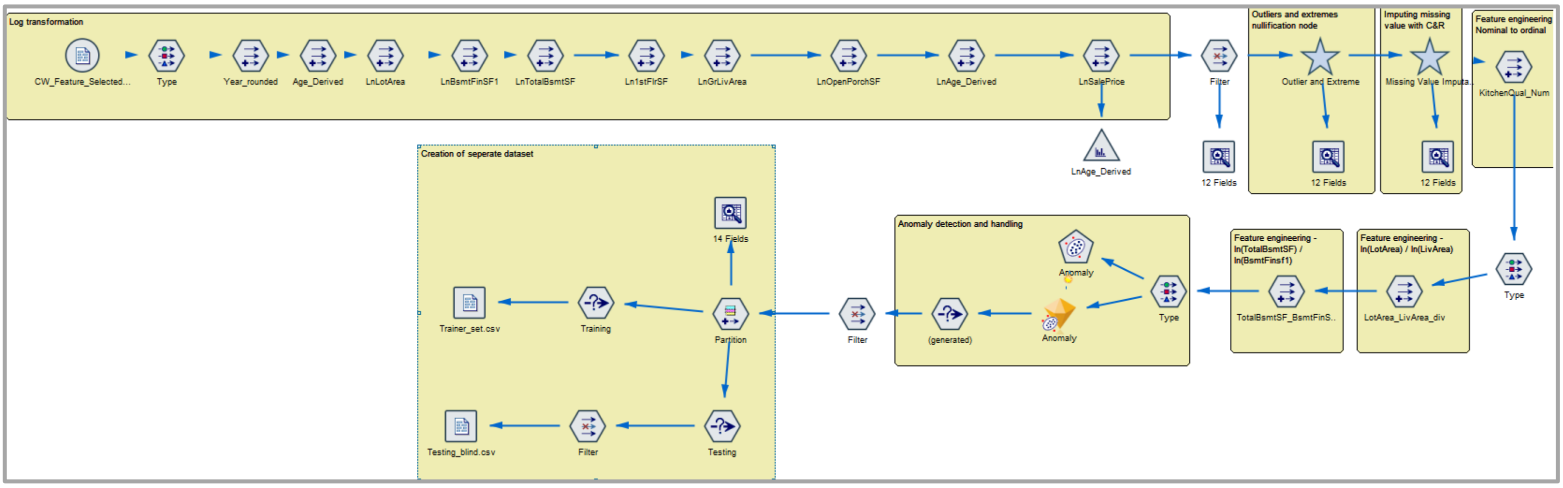


Appendix 3 – Multivariate Analysis

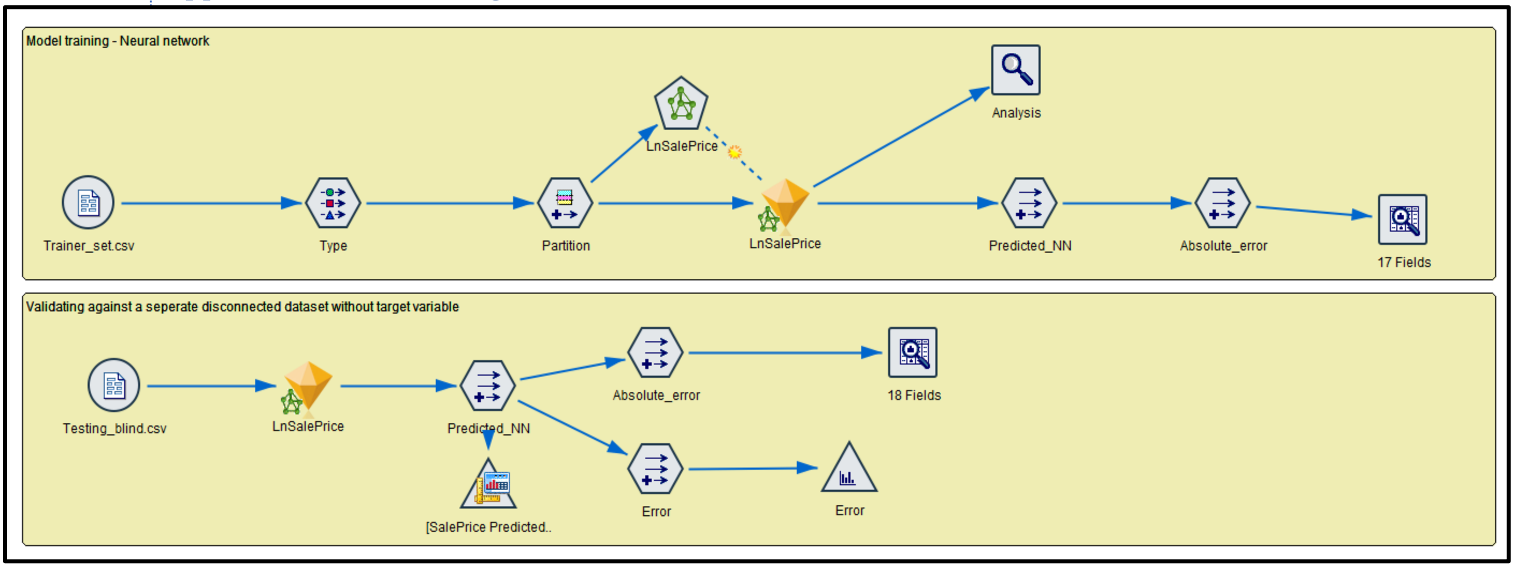


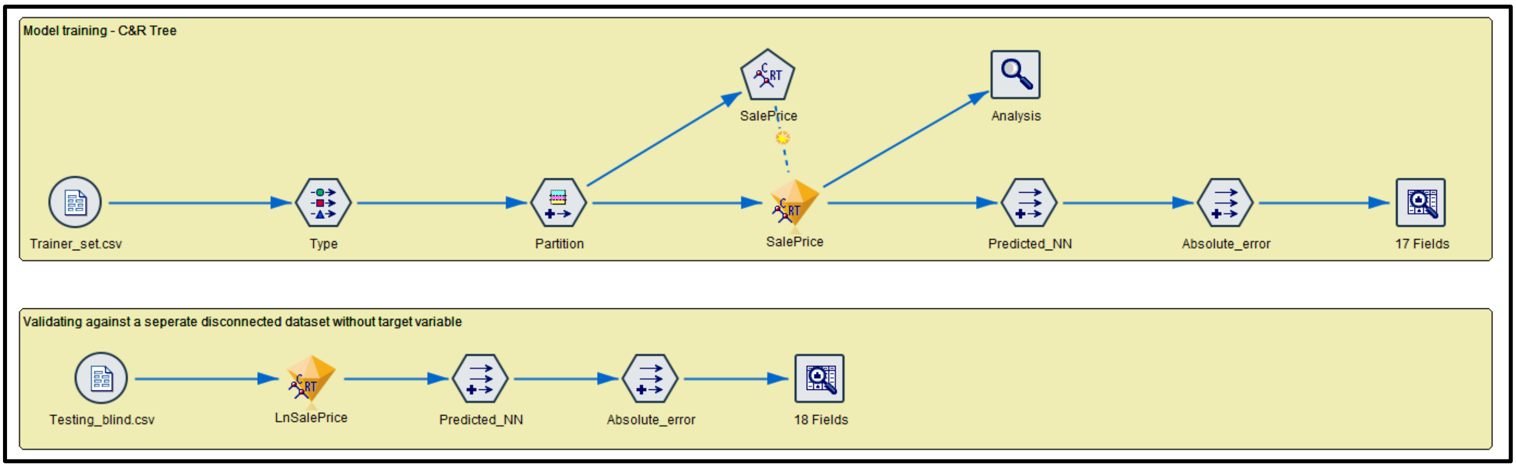
Modeling – Appendix

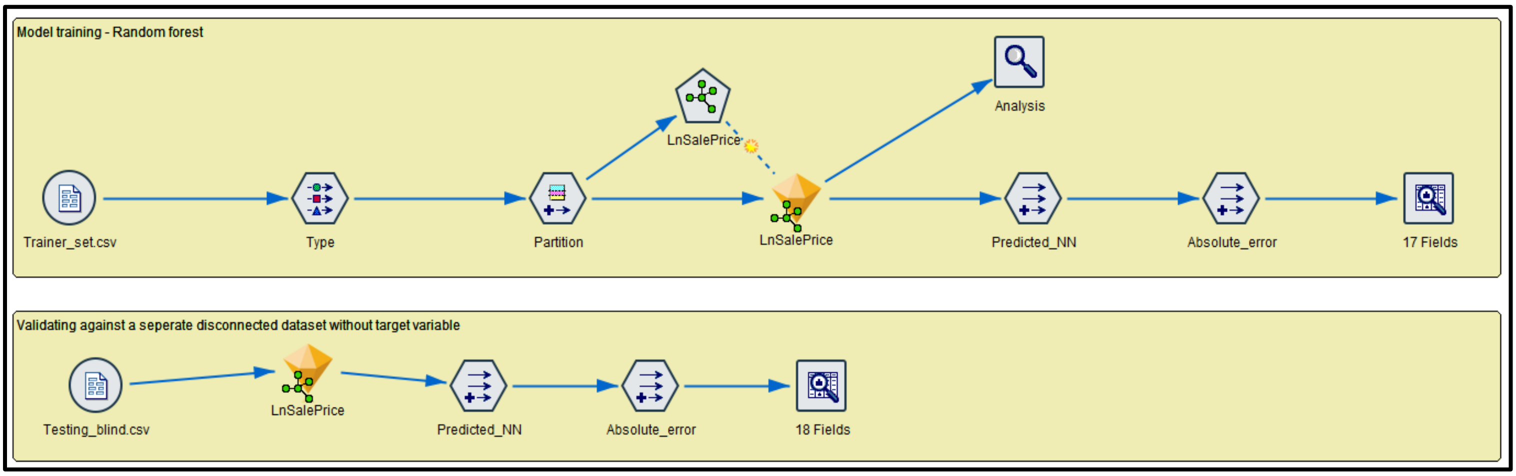
Appendix 1 – Data Pre-processing



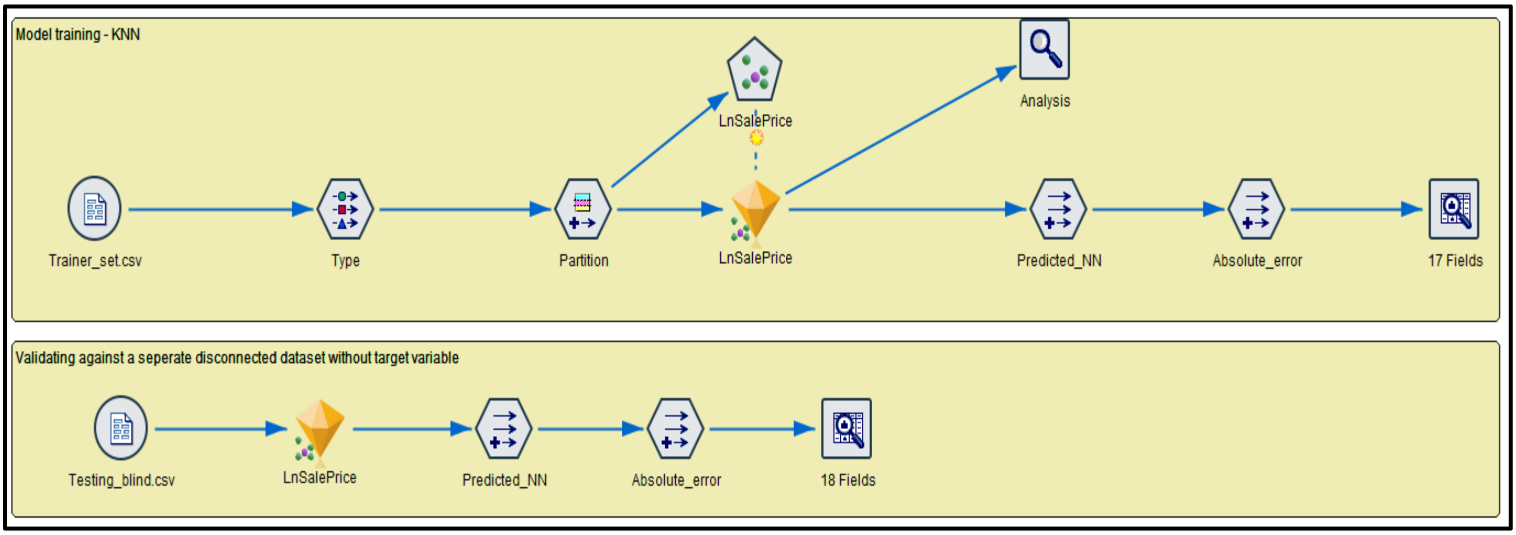
Appendix 2 – Modelling







Model training - Random forest
Analysis
LnSalePrice
Trainer_set.csv
Type
Partition
LnSalePrice
Predicted_NN
Absolute_error
17 Fields
Validating against a seperate disconnected dataset without target variable
LnSalePrice
Testing_blind.csv
Predicted_NN
Absolute_error
18 Fields



Model training - KNN
Analysis
LnSalePrice
Trainer_set.csv
Type
Partition
LnSalePrice
Predicted_NN
Absolute_error
17 Fields
Validating against a seperate disconnected dataset without target variable
Testing_blind.csv
LnSalePrice
Predicted_NN
Absolute_error
18 Fields