PROJECT REPORT

DATA SCIENCE AND MACHINE LEARNING IN CANADA

HOUSE PRICE PREDICTION

INSTRUCTOR: VAHID HADAVI

CLASS: AML 1104

TEAM MEMBERS: LEKSHMI CHANDRAN SHEELA
PATRICIA ADOLPH
POOJA SELBY

Contents

Introduction

The price of houses plays an important role when planning to buy a new home. The high discrepancy between the price of the house makes it more complicated and expensive for homebuyers to choose their dream home. Predicting the price of a house based on several factors can uphold the transparency among buyers and they can easily compare different house prices through this model. Variant number of factors can influence the price of a house like location, physical conditions like building type and quality, facilities available, etc. The main objective of this project is to predict the final price for a house in residential regions using machine learning algorithms.

The following steps are followed to predict the house price:

1.Data Description

The dataset for predicting the house price has been downloaded from the Kaggle dataset “House Prices”. The dataset contains 81 features describing every aspect of residential homes in the regions of Ames, Iowa with 1460 observations which will be analyzed to predict the sale price of a house in the regions.

The data used will be undergoing a combination of pre-processing steps to improve the accuracy of the prediction.

2.Exploratory Analysis

It is one of the important steps of performing early studies on data to spot anomalies, outliers, test hypothesis, and missing values. Exploratory analysis can help to discover patterns by finding interesting relations among the variables with the help of graphical representations and statistics.

The initial inferences that we get from our dataset is:

- There are 1460 instances of the data. Total number of attributes equals 81, of which 35 are numerical, 44 are categorical including Id and Salesprice.

- No duplicate values in the dataset.
- Exploring Categorical variables using count and count% in factors of categorical variables.
- There are 18 features with missing values that needs to be handled.

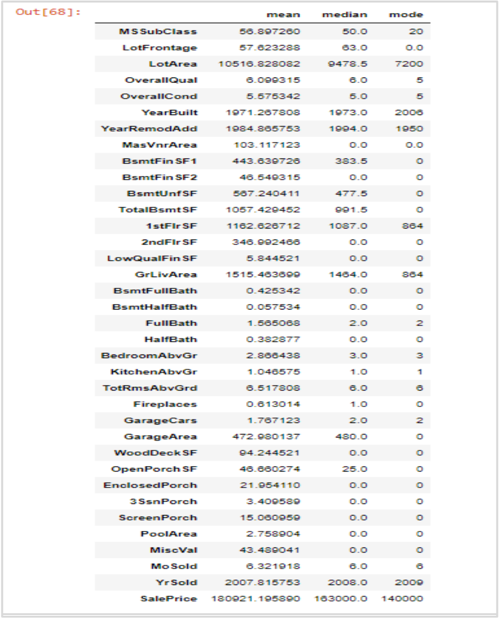
  LotFrontage, Alley, MasVnrType, MasVnrArea, BsmtQual, BsmtCond, BsmtExposure BsmtFinType, BsmtFinType, Electrical1FireplaceQu, GarageType, GarageYrBlt, Garage Finish, GarageQua, GarageCond, PoolQC, Fence, MiscFeature.

The following are the analysis carried out:

1. **Measure of central tendency**

   For describing certain features of data, some aspects of the data can be described quantitatively. So, it is necessary to summarize the dataset into a single value. This type of value that usually comes in the center (somewhere between the two extremes) and represents the entire data set is called as measure of central tendency. Measures of central tendency can be classified as:

   - Mean – It is a value that is obtained by dividing the sum of all the observations by the number of observations
   - Median - Median is the middle number of a set of values when sorted either ascending or descending.
   - Mode – Mode is the frequent number - A number which has the highest frequency in a set of a data is called mode.



Out[68]:

	mean	median	mode
MSSubClass	56.897260	50.0	20
LotFrontage	57.623288	63.0	0.0
LotArea	10516.828082	9478.5	7200
OverallQual	6.099315	6.0	5
OverallCond	5.575342	5.0	5
YearBuilt	1971.267808	1973.0	2006
YearRemodAdd	1984.865753	1994.0	1950
MasVnrArea	103.117123	0.0	0.0
BsmtFinSF1	443.639726	383.5	0
BsmtFinSF2	46.549315	0.0	0
BsmtUnfSF	567.240411	477.5	0
TotalBsmtSF	1057.429452	991.5	0
1stFlrSF	1162.626712	1087.0	864
2ndFlrSF	346.992466	0.0	0
LowQualFinSF	5.844521	0.0	0
GrLivArea	1515.463699	1464.0	864
BsmtFullBath	0.425342	0.0	0
BsmtHalfBath	0.057534	0.0	0
FullBath	1.565068	2.0	2
HalfBath	0.382877	0.0	0
BedroomAbvGr	2.866438	3.0	3
KitchenAbvGr	1.046575	1.0	1
TotRmsAbvGrd	6.517808	6.0	6
Fireplaces	0.613014	1.0	0
GarageCars	1.767123	2.0	2
GarageArea	472.980137	480.0	0
WoodDeckSF	94.244521	0.0	0
OpenPorchSF	46.660274	25.0	0
EnclosedPorch	21.954110	0.0	0
3SsnPorch	3.409589	0.0	0
ScreenPorch	15.060959	0.0	0
PoolArea	2.758904	0.0	0
MiscVal	43.489041	0.0	0
MoSold	6.321918	6.0	6
YrSold	2007.815753	2008.0	2009
SalePrice	180921.195890	163000.0	140000

Screenshot 1 : Mean, Median, Mode of each numerical features

2. Dispersion (Standard deviation and IQR)

Measure of dispersion is used for describing the spread of the data. The spread of data or variations around a central tendency can be described by a range of descriptive statics like variance, standard deviation, and interquartile range.

- Standard deviation – Standard deviation is the measure of the spread of data around the mean value.
- Skewness – It is defined as the degree of asymmetry observed in a probability distribution.
- Kurtosis – It is a measure to determine whether the data are heavily-tailed or lightly-tailed relative to a normal distribution. Data with high kurtosis are more likely to be outliers.
- Inter Quartile range – It is the amount of data spread at the middle of dataset (50%).

  IQR = Q3 – Q1, where Q3 is the third quartile and Q1 is the first quartile.

	types	count	missing count	missing %	min	Q1	mean	median	mode	Q3	max	IQR	IQR_lower	IQR_upper	stdev	skewness	kurtos
MSSubClass	int64	1460	0	0.0	20	20.00	56.897260	50.0	20	70.00	190	50.00	-55.000	145.000	42.300571	1.407657	1.5801
BsmtUnfSF	int64	1460	0	0.0	0	223.00	567.240411	477.5	0	808.00	2336	585.00	-654.500	1685.500	441.866955	0.920268	0.4749
TotalBsmtSF	int64	1460	0	0.0	0	795.75	1057.429452	991.5	0	1298.25	6110	502.50	42.000	2052.000	438.705324	1.524255	13.2504
1stFlrSF	int64	1460	0	0.0	334	882.00	1162.626712	1087.0	864	1391.25	4692	509.25	118.125	2155.125	386.587738	1.376757	5.7458
2ndFlrSF	int64	1460	0	0.0	0	0.00	346.992466	0.0	0	728.00	2065	728.00	-1092.000	1820.000	436.528436	0.813030	-0.5534
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
KitchenQual	object	1460	0	0.0	Ex	NaN	NaN	NaN	TA	NaN	TA	NaN	NaN	NaN	NaN	NaN	Na
Condition2	object	1460	0	0.0	Artery	NaN	NaN	NaN	Norm	NaN	RRNn	NaN	NaN	NaN	NaN	NaN	Na
Foundation	object	1460	0	0.0	BrkTil	NaN	NaN	NaN	PConc	NaN	Wood	NaN	NaN	NaN	NaN	NaN	Na
Electrical	object	1460	0	0.0	FuseA	NaN	NaN	NaN	SBrkr	NaN	SBrkr	NaN	NaN	NaN	NaN	NaN	Na
HeatingQC	object	1460	0	0.0	Ex	NaN	NaN	NaN	Ex	NaN	TA	NaN	NaN	NaN	NaN	NaN	Na

Screenshot 2: Calculating dispersion and IQR (Inter Quartile Range) for each feature.

3. Data pre-processing

Real-world data are always unclean with missing values errors and outliers. Data preprocessing is the technique to convert unclean and biased data into a useful and efficient format. House-price dataset has been pre-processed using the following methods:

1. **Handling missing values**

Missing values occur when data is absent for any variables in the dataset. It is very important to handle missing values because missing values can significantly affect the conclusions derived from the dataset.

Handle missing values for features where median/mean/multiple imputation/regression-based imputation or most common value doesn't make sense in every scenario. The blind imputation will cause drastic distribution changes in the dataset. So, after understanding the data thoroughly we have found that some of the NA values are doesn’t means not available. For e.g.: The feature Alley in data description says NA means "no alley access" and feature BedroomAbvGr in data description NA most likely means 0 etc.

2. **Outlier Detection**

An outlier is a data that stands significantly different from other observations in the dataset. Our dataset has 1460 and removing outlier from 81 features can cause huge data loss. So, by using the IQR_upper and IOR_lower values obtained from quartiles and inter-quartile range, any outliers greater than IQR_upper are updated to be equal to IQR_upper and any outliers lesser than IQR_lower are updated to be equal to IQR_lower here. Hence, no outliers have been removed to prevent loss of data.

3. Variable Transformations

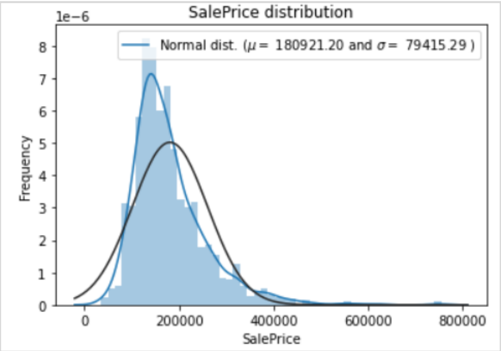
Variable transformation is the process to make the data more efficient, to make it work better for our model. Variable Transformations can be categorized into two:

i. Numerical Variable Transformation

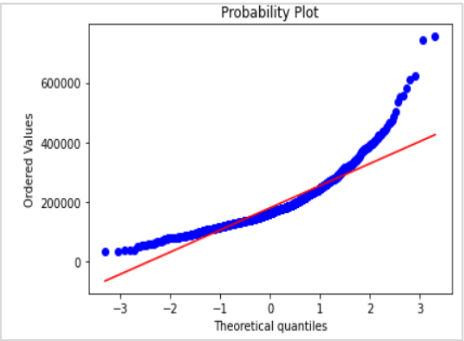
Numerical variable transformation transforms the skewed data distribution to a normal distribution.

- Plotting Distribution of target variable

  Distribution of the target variable 'sales price' is plotted using mu and sigma where mu is the sample mean and sigma is standard deviation.

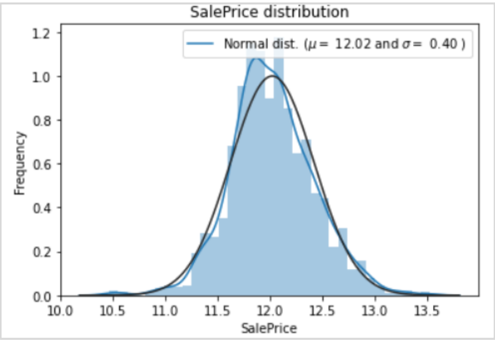


Screenshot 3: Histogram for sales price distribution

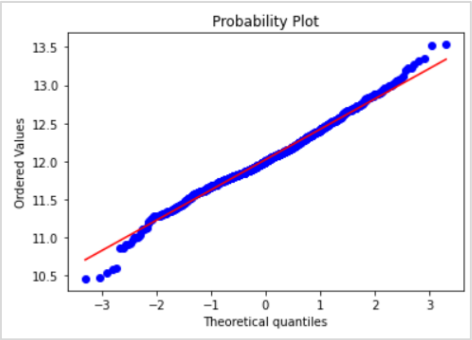


Screenshot 4: Probability plot

From the above 'histogram' and 'probability plot' we can find that the 'Sales Price' is not normal and follows a positively skewed distribution. So, we need to transform the variable using 'Log' to make the distribution to normal. After log transformation of target variable, it follows a normal distribution additionally we have transformed all the positively and negatively skewed distributions using log and exponential variable transformation techniques to get better prediction results



Screenshot 5: Normally distributed histogram



Screenshot 6: Probability plot

- Log Transform of the skewed numerical features

  We can observe that some of the variables are skewed distribution in the dataset. For example, 'LotArea' which is Positively skewed after transformation.

- Transforming positively skewed independent variable to normal using Log transformation.27 positively skewed numerical features are subjected to log transform.

- Transforming negatively skewed distributed variables to normal using exponential transformation. There are zero skewed numerical features for log transform.

ii. Categorical Variable Transformations

Categorical variable transformation transforms categorical variable to numeric variable.

- Label Encoding

  It is an approach to convert the original data to integers.

  For e.g.: Variable Alley has features as Grvl, Pave, No and we are assigning ordered numbers as Grvl: 1, Pave: 2, No: 3.

- Simplification of existing features

  Some of the existing features having from 1 to 10 have been replaced with values. For e.g., Observations were marked from 1 to 10 which have been replaced as:1: 1, 2: 1, 3: 1, # bad, 4: 2, 5: 2, 6: 2, # average7: 3, 8: 3, 9: 3, 10: 3 # good

- Combination of existing features

  Some features in our dataset have been combined into one feature. e.g., OverallQual and OverallCond is combined as OverallGrade and added as a new feature in the dataset.

- Polynomials on the top 10 existing features

  Polynomials on the top ten important features that are relative to target variable ‘Sales Price’ are calculated.

  After variable transformation, our dataset contains 143 features with 119 numerical features and 24 categorical features.

4. Model Building

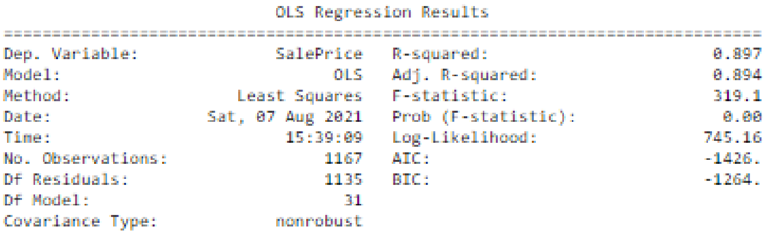
Multiple linear regression using backward stepwise regression

Multiple linear regression is a statistical technique using several explanatory variables to predict the outcome of a target variable. This method is best suited when having a small set of predicators (one or more) and when we do not know which independent variable creates the best prediction.

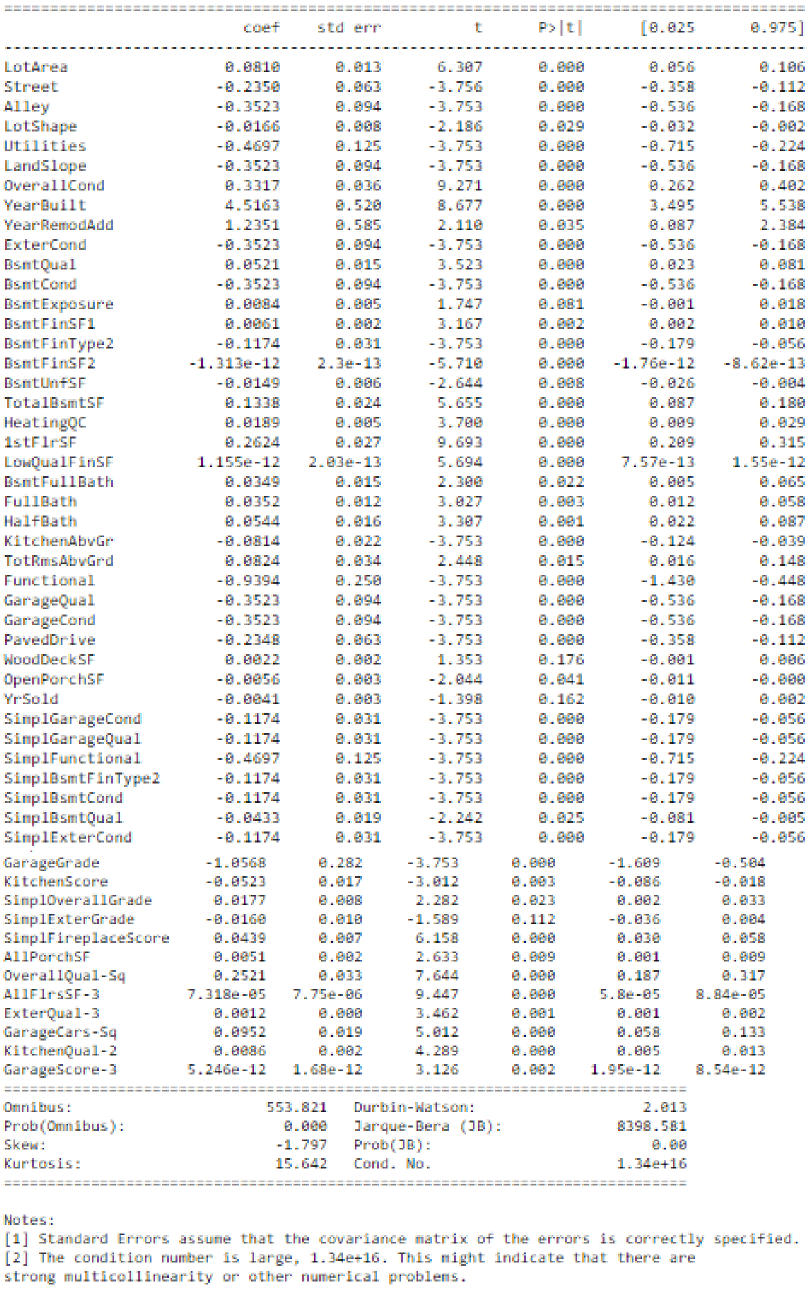
.**Backward selection** is the process in which all the independent variables are entered into the equation first and if they do not contribute to the regression equation each one is deleted one at a time. In other words, it is a reverse process, where we remove the variable that is statistically least significant. When the best model score is achieved and when there is no further improvement, we can stop using a stopping rule.

Training the 76 strongly corelated variables with our target variable sales price can make our model multicollinear and make our model inefficient. So, to avoid multicollinearity in the set of multiple regression variables we have removed features using the variance inflation factor (VIF). Even though the threshold is 10 we are still having multicollinearity in our model. After training the model using Multiple Regression, a model with overall accuracy of 0.88 is created.

The following is the summary of the multiple linear regression model:



```
                            OLS Regression Results
==============================================================================
Dep. Variable:              SalePrice   R-squared:                       0.897
Model:                            OLS   Adj. R-squared:                  0.894
Method:                 Least Squares   F-statistic:                     319.1
Date:                Sat, 07 Aug 2021   Prob (F-statistic):               0.00
Time:                        15:39:09   Log-Likelihood:                 745.16
No. Observations:                1167   AIC:                            -1426.
Df Residuals:                    1135   BIC:                            -1264.
Df Model:                          31
Covariance Type:            nonrobust
```



```
=======================================================================================
                            coef    std err          t      P>|t|     [0.025     0.975]
---------------------------------------------------------------------------------------
LotArea                   0.0810      0.013      6.307      0.000      0.056      0.106
Street                   -0.2350      0.063     -3.756      0.000     -0.358     -0.112
Alley                    -0.3523      0.094     -3.753      0.000     -0.536     -0.168
LotShape                 -0.0166      0.008     -2.186      0.029     -0.032     -0.002
Utilities                -0.4697      0.125     -3.753      0.000     -0.715     -0.224
LandSlope                -0.3523      0.094     -3.753      0.000     -0.536     -0.168
OverallCond               0.3317      0.036      9.271      0.000      0.262      0.402
YearBuilt                 4.5163      0.520      8.677      0.000      3.495      5.538
YearRemodAdd              1.2351      0.585      2.110      0.035      0.087      2.384
ExterCond                -0.3523      0.094     -3.753      0.000     -0.536     -0.168
BsmtQual                  0.0521      0.015      3.523      0.000      0.023      0.081
BsmtCond                 -0.3523      0.094     -3.753      0.000     -0.536     -0.168
BsmtExposure              0.0084      0.005      1.747      0.081     -0.001      0.018
BsmtFinSF1                0.0061      0.002      3.167      0.002      0.002      0.010
BsmtFinType2             -0.1174      0.031     -3.753      0.000     -0.179     -0.056
BsmtFinSF2            -1.313e-12    2.3e-13     -5.710      0.000  -1.76e-12  -8.62e-13
BsmtUnfSF                -0.0149      0.006     -2.644      0.008     -0.026     -0.004
TotalBsmtSF               0.1338      0.024      5.655      0.000      0.087      0.180
HeatingQC                 0.0189      0.005      3.700      0.000      0.009      0.029
1stFlrSF                  0.2624      0.027      9.693      0.000      0.209      0.315
LowQualFinSF           1.155e-12   2.03e-13      5.694      0.000   7.57e-13   1.55e-12
BsmtFullBath              0.0349      0.015      2.300      0.022      0.005      0.065
FullBath                  0.0352      0.012      3.027      0.003      0.012      0.058
HalfBath                  0.0544      0.016      3.307      0.001      0.022      0.087
KitchenAbvGr             -0.0814      0.022     -3.753      0.000     -0.124     -0.039
TotRmsAbvGrd              0.0824      0.034      2.448      0.015      0.016      0.148
Functional               -0.9394      0.250     -3.753      0.000     -1.430     -0.448
GarageQual               -0.3523      0.094     -3.753      0.000     -0.536     -0.168
GarageCond               -0.3523      0.094     -3.753      0.000     -0.536     -0.168
PavedDrive               -0.2348      0.063     -3.753      0.000     -0.358     -0.112
WoodDeckSF                0.0022      0.002      1.353      0.176     -0.001      0.006
OpenPorchSF              -0.0056      0.003     -2.044      0.041     -0.011     -0.000
YrSold                   -0.0041      0.003     -1.398      0.162     -0.010      0.002
SimplGarageCond          -0.1174      0.031     -3.753      0.000     -0.179     -0.056
SimplGarageQual          -0.1174      0.031     -3.753      0.000     -0.179     -0.056
SimplFunctional          -0.4697      0.125     -3.753      0.000     -0.715     -0.224
SimplBsmtFinType2        -0.1174      0.031     -3.753      0.000     -0.179     -0.056
SimplBsmtCond            -0.1174      0.031     -3.753      0.000     -0.179     -0.056
SimplBsmtQual            -0.0433      0.019     -2.242      0.025     -0.081     -0.005
SimplExterCond           -0.1174      0.031     -3.753      0.000     -0.179     -0.056
GarageGrade              -1.0568      0.282     -3.753      0.000     -1.609     -0.504
KitchenScore             -0.0523      0.017     -3.012      0.003     -0.086     -0.018
SimplOverallGrade         0.0177      0.008      2.282      0.023      0.002      0.033
SimplExterGrade          -0.0160      0.010     -1.589      0.112     -0.036      0.004
SimplFireplaceScore       0.0439      0.007      6.158      0.000      0.030      0.058
AllPorchSF                0.0051      0.002      2.633      0.009      0.001      0.009
OverallQual-Sq            0.2521      0.033      7.644      0.000      0.187      0.317
AllFlrsSF-3            7.318e-05   7.75e-06      9.447      0.000    5.8e-05   8.84e-05
ExterQual-3               0.0012      0.000      3.462      0.001      0.001      0.002
GarageCars-Sq             0.0952      0.019      5.012      0.000      0.058      0.133
KitchenQual-2             0.0086      0.002      4.289      0.000      0.005      0.013
GarageScore-3          5.246e-12   1.68e-12      3.126      0.002   1.95e-12   8.54e-12
==============================================================================
Omnibus:                      553.821   Durbin-Watson:                   2.013
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             8398.581
Skew:                          -1.797   Prob(JB):                         0.00
Kurtosis:                      15.642   Cond. No.                     1.34e+16
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.34e+16. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Figure 7 : Summary of multiple linear regression

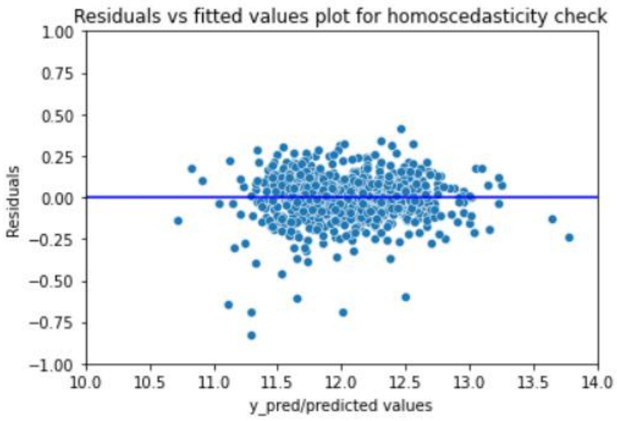
Validating assumptions of our model using 5 methods:

1. **Mean of Residuals**

   Mean of residual for our model is -6.837494984175303e-16 which is very close to zero.

2. **Check for Homoscedasticity**

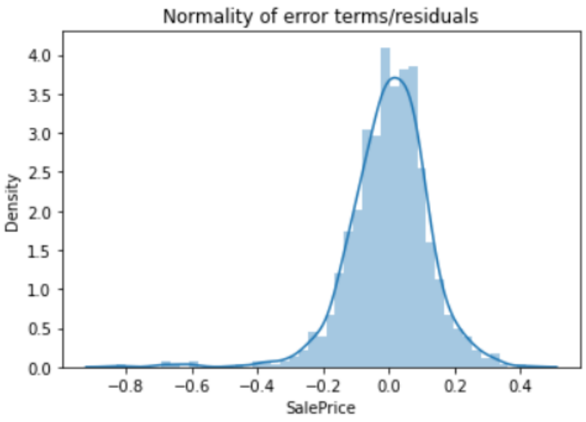
   Since there is no pattern observed from the graph, we can assume that no homoscedasticity is present in our model.



Screenshot 8: Residual vs fitted plot for homoscedasticity

3. **Check for Normality of error terms/residuals**

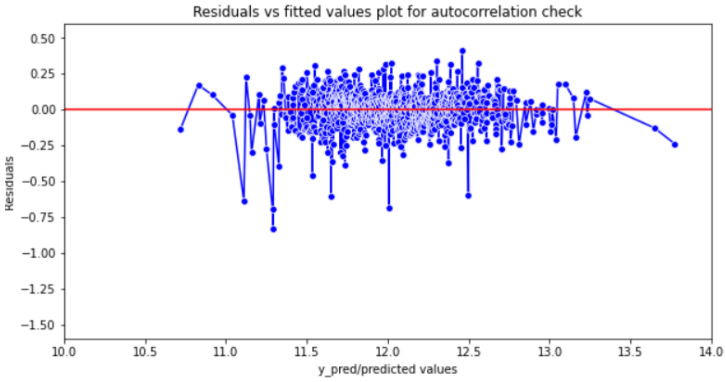
   We can observe that, the residual distribution follows a normal distribution with mean 0 and a small constant variance.



Screenshot 9: Normality of error terms/residuals

4. **No autocorrelation of residuals**

There is no autocorrelation in the model does not follow any pattern.



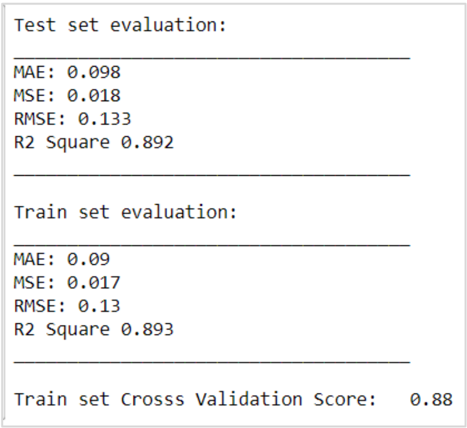
Screenshot 10: Residual vs fitted values plot for autocorrelation check

5. **No Perfect Multicollinearity**

We have checked the VIF factor for each variable and considered VIF < 10 value as cut-off.

5. Model Evaluation

Model have been evaluated to determine the goodness of fit between model and our data using Adjusted R-squared, mean absolute error, Root Mean Squared error, and Root Mean Squared Error.



```
Test set evaluation:
_____________________________________
MAE: 0.098
MSE: 0.018
RMSE: 0.133
R2 Square 0.892
_____________________________________

Train set evaluation:
_____________________________________
MAE: 0.09
MSE: 0.017
RMSE: 0.13
R2 Square 0.893
_____________________________________

Train set Crosss Validation Score:   0.88
```

Screenshot 11: Model Evaluation using Multiple Linear Regression

From the above assumptions testing and evaluation metrics of Multiple Linear Regression using Stepwise Back Propagation Algorithm, we have observed that the house price prediction model using Stepwise Linear Regression satisfies all the statistical significance and assumptions except 'multicollinearity'. The large number of independent features of house price prediction causes high correlation between the independent variables. This leads to multicollinearity. So, it is hard to interpret the house price prediction model and creates an overfitting problem.

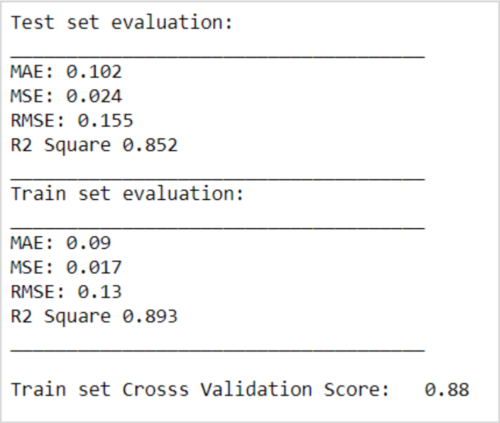
RIDGE REGRESSION

Ridge regression model is used to analyse the data that have multicollinearity. When data have multicollinearity, there will be unbiased least-squares with large variances, and this can result in making predicted values to be far away from the actual values. By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors. Therefore, the model will be more reliable. Since the model build with Multiple Linear Regression using

'Stepwise Back Propagation Algorithm' contains multicollinearity we choose to build the final model using 'Ridge Regression'. Formulae to calculate regression equation is Y=XB+e where, y is the dependent variable, x is the independent variable, B is the regression coefficients and e stands for error.

Model Evaluation

Model have been evaluated to determine the goodness of fit between model and our data using Adjusted R-squared, mean absolute error, Root Mean Squared error, and Root Mean Squared Error. The model build using Ridge Regression has an accuracy of 0.88. Thus avoids over fitting.



```
Test set evaluation:
_____________________________________
MAE: 0.102
MSE: 0.024
RMSE: 0.155
R2 Square 0.852
_____________________________________
Train set evaluation:
_____________________________________
MAE: 0.09
MSE: 0.017
RMSE: 0.13
R2 Square 0.893
_____________________________________

Train set Crosss Validation Score:   0.88
```

Screenshot 12: model evaluation using Ridge Regression

Conclusion

The analysis and prediction of house price dataset can help to resolve the discrepancy of house price in residential regions. The prediction results from out model will fulfill the opportunity not only to predict the house prices but also helps for foresee the future demands for house and a region. The insights can also guide Real Estate Investment teams to take better and right investments. In this project, we have used multiple linear regression using backward elimination and Ridge regression for predicting house sales prices. After model evaluation for each model, it is evident that Ridge Regressor is more efficient since it avoids multicollinearity with an accuracy of 0.88.

References

House Price Prediction | Kaggle https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data?select=train.csv

Forecasting: Principles and Practice (2nd ed). 12.9 Dealing with missing values and outliers. (n.d.). https://otexts.com/fpp2/missing-outliers.html.

Measures of central tendency and variability. ABCTE Prepare to Teach Workshops. (n.d.). https://www.americanboard.org/ptk/measures-of-central-tendency-and-variability/.

Selection process for multiple regression. Complete Dissertation by statistics solutions. (n.d.). https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/selection-process-for-multiple-regression/?__cf_chl_jschl_tk__=pmd_74389c44f71e14addbe64fdabc32043ac6a94f12-1628193864-0-gqNtZGzNAk2jcnBszQm6

DEI, M. (2019, December 28). *Catalog of Variable Transformations To Make Your Model Work Better*. BOOK "DATA ANALYSIS TECHNIQUES TO WIN KAGGLE." https://towardsdatascience.com/catalog-of-variable-transformations-to-make-your-model-works-better-7b506bf80b97.

Ashok, P. (2020, October 15). *What is Ridge Regression?* Great Learning. https://www.mygreatlearning.com/blog/what-is-ridge-regression/.