A picture containing text

Description automatically generated

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**

1. Introduction
2. Data Description
3. Exploratory Analysis
4. Data Pre-Processing
5. Model Building and Evaluation
6. Conclusion

**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.

A picture containing table

Description automatically generated

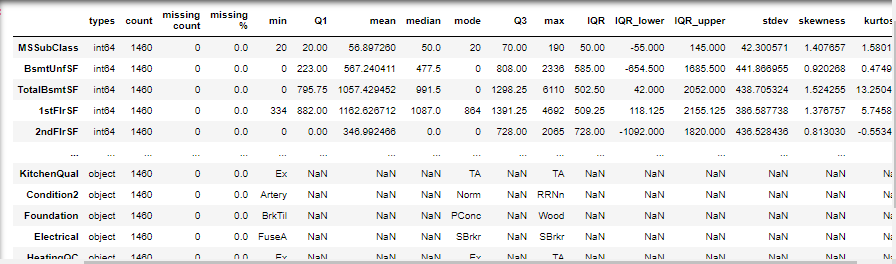
Screenshot 1 : Mean, Median, Mode of each numerical features

1. **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.



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.

1. **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.

1. **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:

1. **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.

Histogram

Description automatically generated Chart, line chart

Description automatically generated

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

Chart, histogram

Description automatically generated Chart, line chart

Description automatically generated

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.

1. **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**

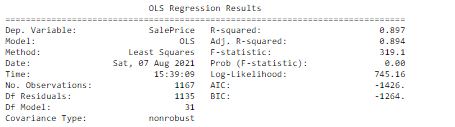
**Multiple linear regression using backward stepwise regression**

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:



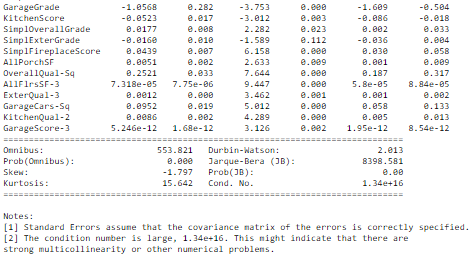
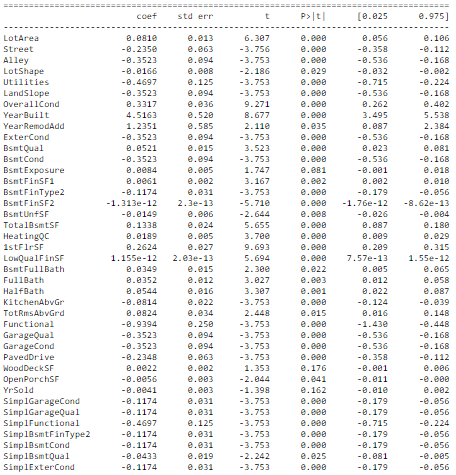


Figure 7 : Summary of multiple linear regression

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.

1. **Check for Homoscedasticity**

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

Chart, scatter chart

Description automatically generated

Screenshot 8: Residual vs fitted plot for homoscedasticity

1. **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.

Chart, histogram

Description automatically generated

Screenshot 9: Normality of error terms/residuals

1. **No autocorrelation of residuals**

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

Chart

Description automatically generated

Screenshot 10: Residual vs fitted values plot for autocorrelation check

1. **No Perfect Multicollinearity**

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

**5. Model Evaluation**

Modelhave 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.

Table

Description automatically generated

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**

Modelhave 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.

Table

Description automatically generated

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/>.