Housing prices prediction using Regression

Ram Kakkar
Machine Learning
AI Technology and Systems
iamramkakkar@gmail.com
www.ai-techsystems.com

***Abstract*— Rising house prices, generally encourage consumer spending and lead to higher economic growth. A sharp drop in house prices adversely affects consumer confidence and lead to lower economic growth. Therefore House Prediction System must be build to help customer to arrange the right time amd right property to purchase the House. The principal idea of this article is to predict house pricing using Regression**

Keywords—Machine learning, Regression, Deep Learning

I. INTRODUCTION

Investment in Real Estate is profitable because the property value does not decline rapidly. In this project, we will develop and evaluate the performance and the predictive power of a model trained and tested on data collected from houses in Ames. Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. Once we get a good fit, we will use this model to predict the monetary value of a house located at the Ames. A model like this would be very valuable for a real state agent who could make use of the information provided in a daily basis.

II. METHODOLOGY

A. Data Collection

The dataset used in this project is open sourced dataset from Kaggle.com. It consist with 79 parameters that have the possibility of affecting the housing prices. Some parameters have Numerical value and some parameters have Categorical value. Some of the parameters are:

Parameters	Description	Datatype
SalePrice	the property's sale price in dollars	Numerical
MSSubClass	The building class	Categorical
MSZoning	The general zoning classification	Categorical
LotFrontage	Linear feet of street connected to property	Numerical
LotArea	Lot size in square feet	Numerical
Street	Type of road access	Categorical
Alley	Type of alley access	Categorical
LotShape	General shape of property	Categorical
LandContour	Flatness of the property	Categorical

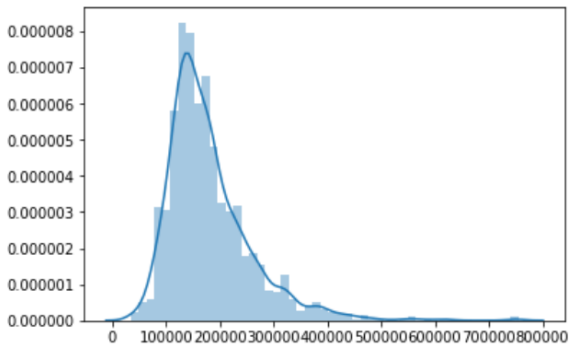
B. Data Preprocessing

Entire dataset is checked for NaN or Missing Values. Some NaN feature are replaced with “None” in which NaN has meaning and in some parameters NaN and Missing Values are replaced with most occurring value.After filling the NaN values and Missing data. The categorical data is encoded. Encoding refers to conversion of categorical values into numerical values. This was done using LabelEncoder of sklearn.preprocessing package. It was done both on train data and test data

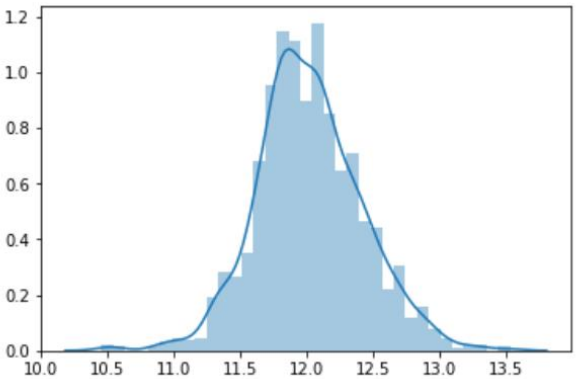
C. Data Analysis

Before applying any model to our dataset. We need to check the characteristics of our dataset. Thus we need to analyze our to dataset and study different parameters. We also need to find out the outliers present in our dataset. Outliers are experimental errors and they need to be excluded from the dataset else the will mislead our model and lower the accuracy of our model.

‘Sale Price’ Distribution is visualized and is found out that it is skewed right. Log transformation is necessary in order to maintain Normal distribution and therefore Log Transformation is applied to ‘Sale Price’

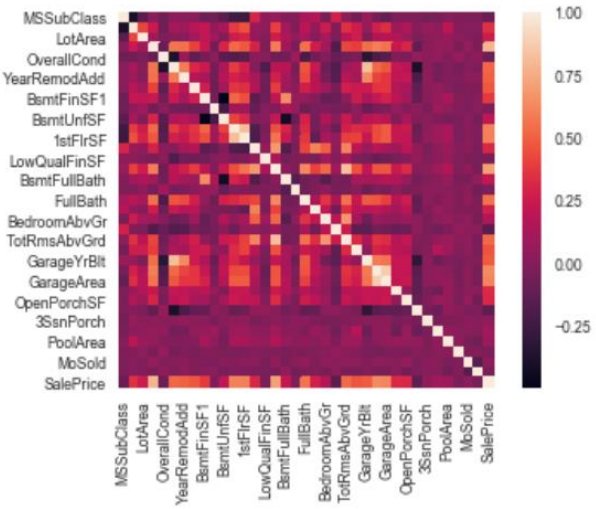


Sale Price without Log Transformation:



Sale price with Log TransformationAfter analysing the prediction value, Now the parameters is to be analysed. Correlation of each feature is calculated with 'Sale Price' to find out which feature affect the most. The correlation value is between -1 to +1. the feature with most positive correlation affects 'Sale Price' and the feature with the most negative correlation affects the 'Sale Price' most negatively. Correlation is found using corr() function

The correlation between each feature is can also be represented with a heatmap. heatmap is imported from a



seaborn package

SalePrice	1.0000
OverallQual	0.790982
GrLivArea	0.708624
GarageCars	0.640409
GarageArea	0.623431
TotalBsmtSF	0.613581
1stFlrSF	0.605852
FullBath	0.560664
TotRmsAbvGrd	0.533723
YearBuilt	0.522897
YearRemodAdd	0.507101
OverallCond	-0.077856
MSSubClass	-0.084284
EnclosedPorch	-0.128578
KitchenAbvGr	-0.13590

D. Training

After the data is cleaned, the model is to be selected and data is to be trained with the selected model. The dataset given and the predicted value is continuous. So regression technique is used. Regression is a technique used to model and analyze the relationship between variables and often time how they contribute and are related to produce a particular outcome together. A linear regression refers to regression model that is completely made up of liner variables. The training data was divided into training and testing so as to validate the model built. The accuracy is checked by calculating the R2 Score

Different types of regression models:

1. Linear Regression
2. Logistic Regression
3. Polynomial Regression
4. Stepwise Regression
5. Ridge Regression
6. Lasso Regression
7. Elastic Net Regression

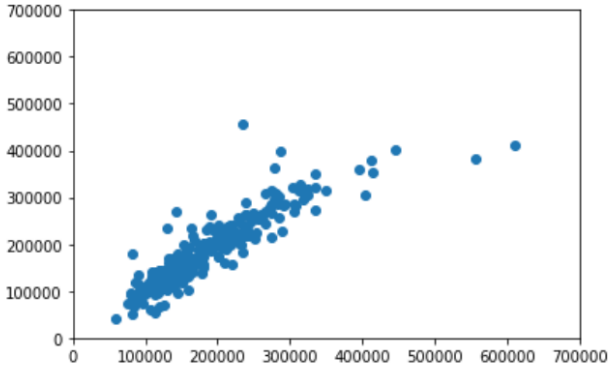
III. MODELS

A. Linear Regression

The multiple linear regression explains the relationship between one continuous dependent variable (y) and **two or** more independent variables($x1$,$x2$,$x3$…)**.** Multiple linear regression attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. Every value of the independent variable x is associated with a value of the dependent variable y.To view the fit of the model to the observed data, one may plot the computed regression line over the actual data points to evaluate the results.

Linear Regression model is imported form sklearn.lineaar_model and the model is fit by providing train data. The plot is drawn between the actual value and and the Predicted value from our model.

- There Must be Linear Relationship between independent and dependent variable
- Multiple Regression Suffers from multidisciplinary, auto correlation etc
- Linear Regression is very sensitive to outliers. It can terribly affect the regression line and evantually forcasted values
- Multicollinearity can increase the variance of the



cofficient estimates

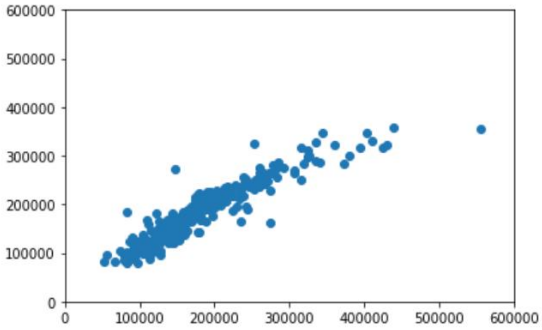
B. ***Ridge Regression***

It is the technique used when data suffers from multicollinearity that is independent variables are highly correlated. In multi col linearity even though the lease square estimates are unbiased, their variance are large which deviates the observed value from the true value. By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors

Above we saw the Linear eqaution but it also contains an error term. In Linear equation the prediction errors can be decomposed into two sub components. First is due to the biased and seceond is due to the variance

Ridge regression solves the multicollinearity problem through shrinkage parameter (lambda)

1. The assumtion of this regression is same as least squared regression except normality is not to be asumed
2. It shrinks the value of coefficients but doesn't reaches zero which suggests no feature selection feature
3. This is a regularization method



IV. CONCLUSION

I have performed Data Analysis of housing Dataset and preprocessed it for the modelling Successfully modelled 2 Regression algorithms and Calculated their r2 score and found out that Linear Regression is best suitable for this dataset with a r2 score of 0.81

V. REFERENCES

1. https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/
2. https://towardsdatascience.com/5-types-of-regression-and-their-properties-c5e1fa12d55e
3. https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data