

HOUSING PRICE PROJECT REPORT





SUBMITTED BY:
SRIDHAR N





ACKNOWLEDGMENT

I wish to express my sincere thanks to the following person, without him I would not have been able to complete this project;

Dr Harsha Burri Reddy, whose insight and knowledge in this subject guided us throughout this project.





BUSINESS PROBLEM FRAMING



The real estate sector is an important industry with many stakeholders ranging from regulatory bodies to private companies and investors. Among these stakeholders, there is a high demand for a better understanding of the industry operational mechanism and driving factors. Today there is a large amount of data available on relevant statistics as well as on additional contextual factors, and it is natural to try to make use of these in order to improve our understanding of the industry





Motivations

Being extremely interested in everything having a relation with the Machine Learning, the independant project was a great occasion to give me the time to learn and confirm my interest for this field. The fact that we can make estimations, predictions and give the ability for machines to learn by themselves is both powerful and limitless in term of application possibilities. We can use Machine Learning in Finance, Medicine, almost everywhere. That's why I decided to conduct my project around the Machine Learning.







Idea



As a first experience, I wanted to make my project as much didactic as possible by approaching every different steps of the machine learning process and trying to understand them deeply. Known as “toy problem” defining the problems that are not immediate scientific interest but useful to illustrate and practice, I chose to take Real Estate Prediction as approach. The goal was to predict the price of a given apartment according to the market prices taking into account different “features” that will be developed in the following sections.



Basic Statistical Measures

With the type of my attributes in mind, I used basic statistical measures on my data such as the mean, the mode, the standard deviation, etc. For the rest of the rapport I will be focus only on the data coming from hongkonghomes to keep the reading easy to follow.

LotFrontage	MasVnrArea	GarageYrBlt	
count	954.00000	1161.000000	1104.000000
mean	70.98847	102.310078	1978.193841
std	24.82875	182.595606	24.890704
min	21.00000	0.000000	1900.000000
25%	60.00000	0.000000	1961.000000
50%	70.00000	0.000000	1980.000000
75%	80.00000	160.000000	2002.000000
max	313.00000	1600.000000	2010.000000



With the following result I was able to notify some interesting such as :



- There is no missing value.
- The mean for the number of bedrooms and number of bathrooms is not meaningful.
- Some estates don't have any bedroom but all have at least one bathroom.
- The Q1, Q2, Q3 (25%, 50%, 100%) are good indicators of the shape of the different attribute. The Q2 being the median that is not influenced by outliers unlike the mean.
- According to the max values, there are obviously outliers like for gross area where 75% of the data have a median about 1906, 10061 appears clearly to be an outlier as well as the max value for number of bedrooms.
- The standard deviation gives also an indication about what should be considered as outliers, but it is not a robust technique since the standard deviation use the mean to be computed.
- Finally, the minimum value of the price is 0 meaning there is at least a noisy tuple in the dataset.

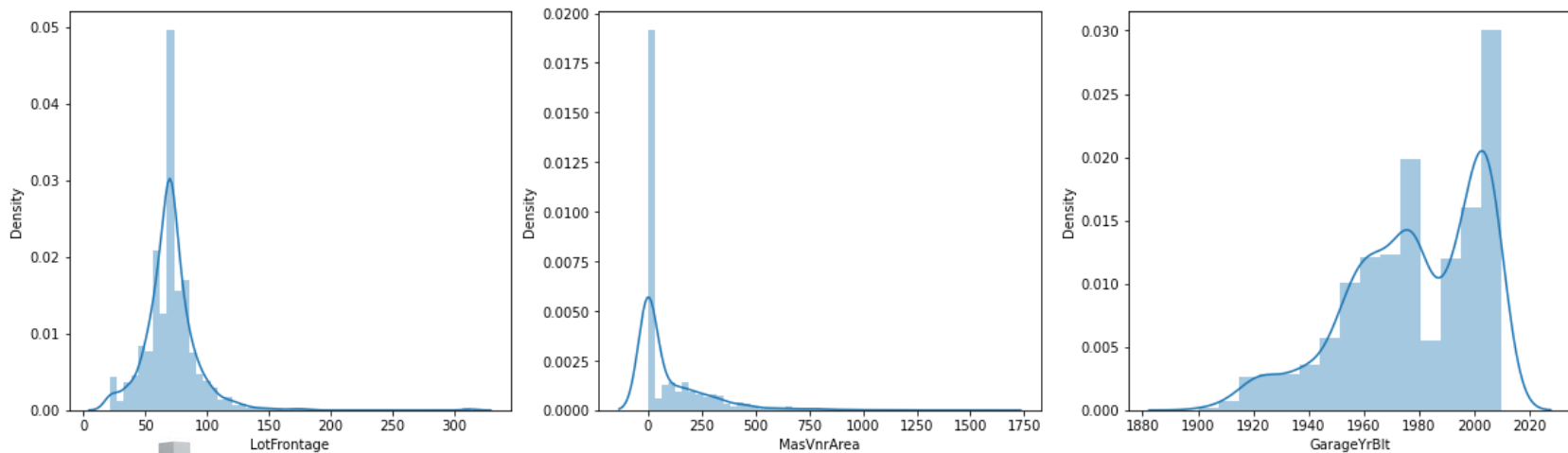




Visualizing The Data



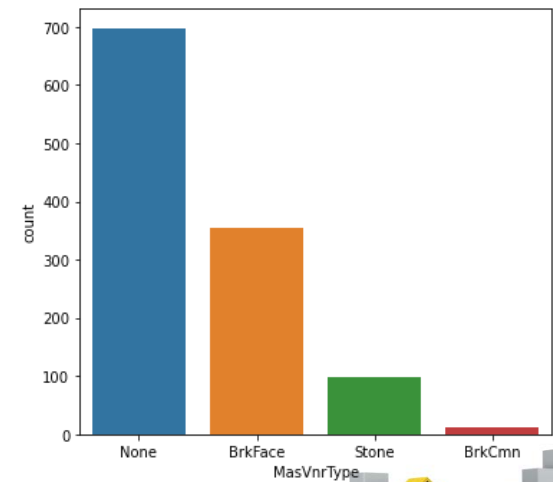
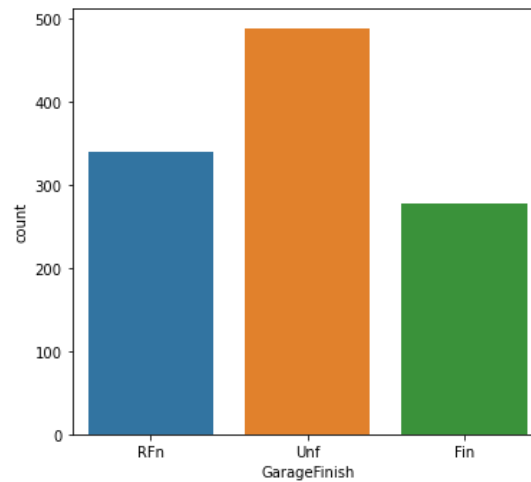
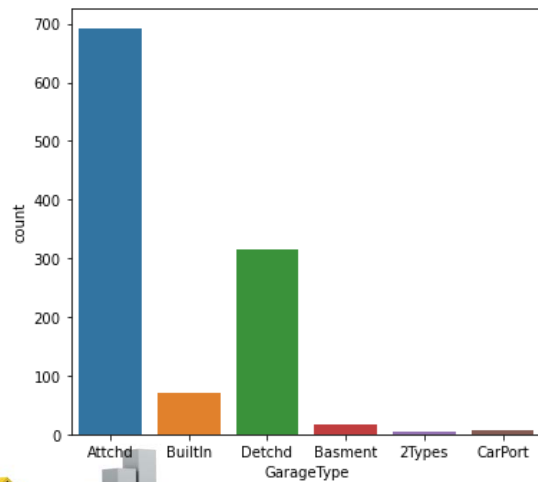
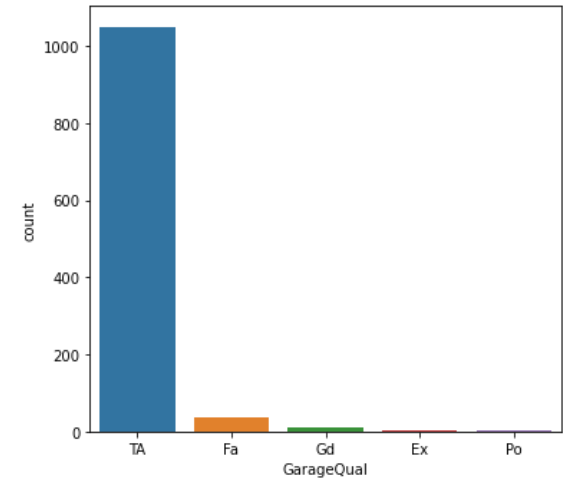
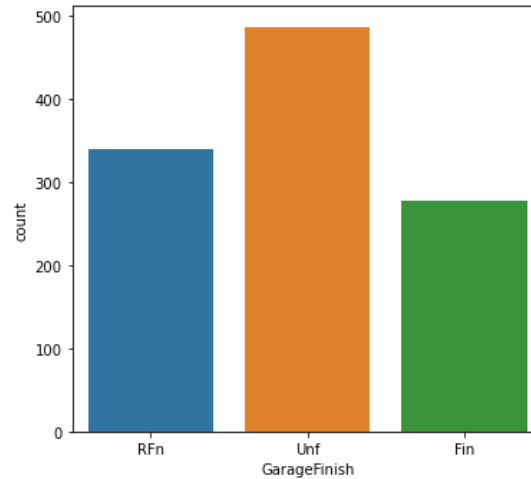
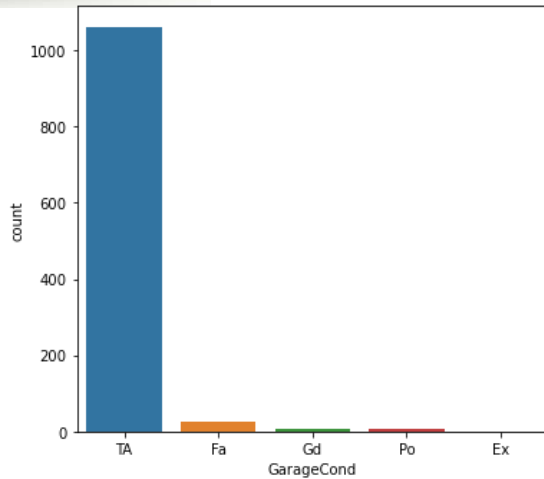
Next, I tried to figure out what kind of chart I could use and it turns out that each one has its own advantage depending on what we want to visualize. Let's start with histogram. A histogram is graphical representation of the distribution of a continuous data. It estimates the probability distribution. It is composed of bins, representing a range of values, on the x axis there are the values, on the y axis there are the frequencies of the bins. The histograms gave me a better intuition of how my variables were distributed.



: On the left, the distribution of the variable gross area; On the right the distribution of the variable saleable area.



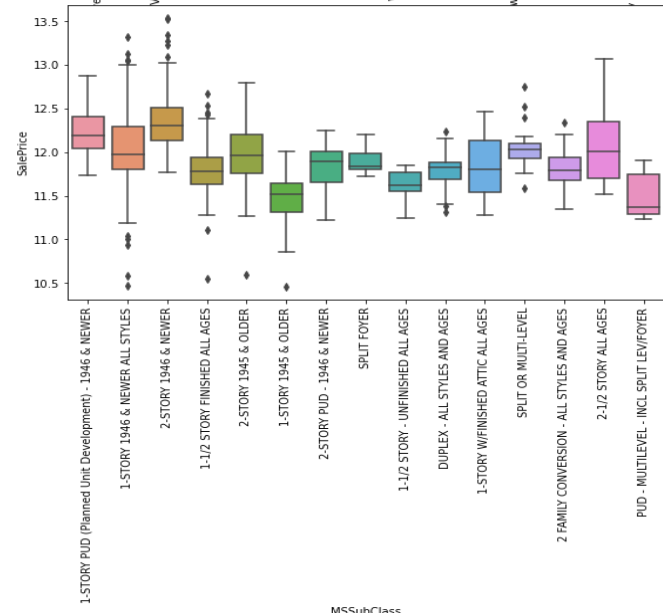
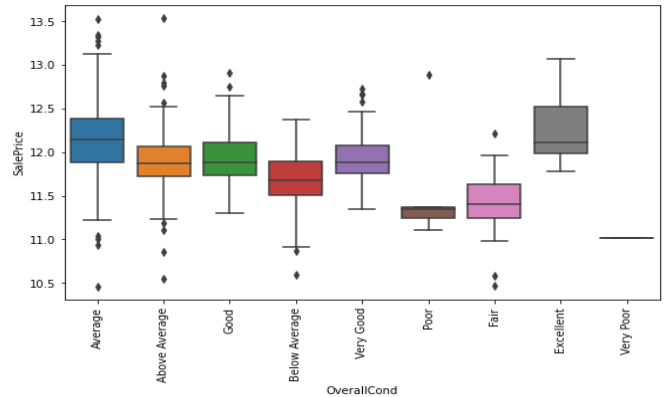
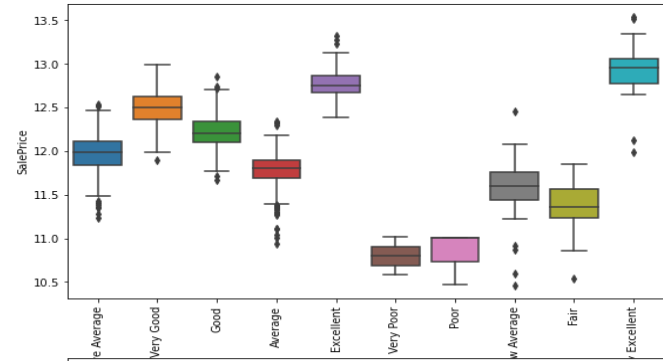
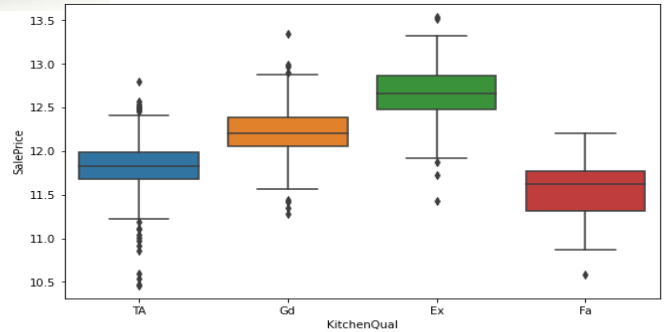
#Visualising the variables with missing values





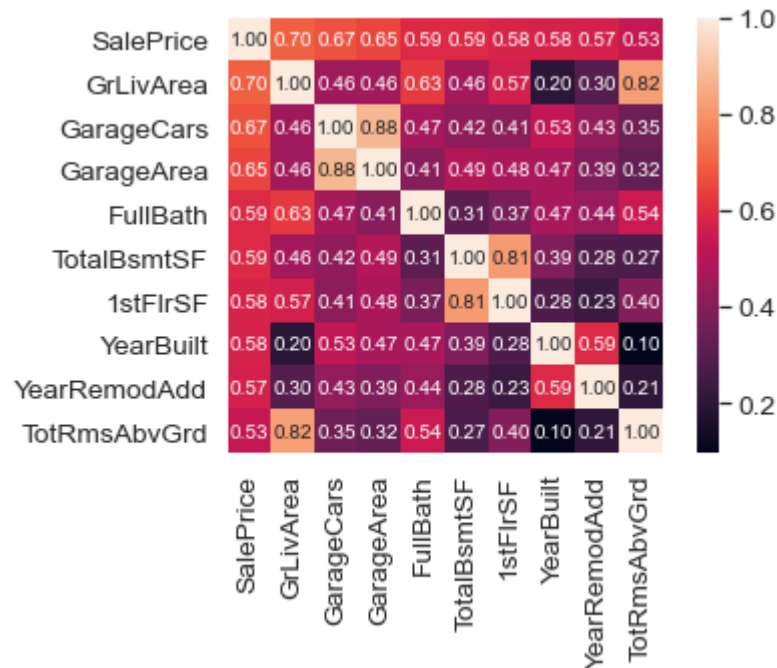
The boxplot is a convenient way of graphically depicting groups of numerical data through their quartiles. The quartiles are in the box where Q1 is the 25% of the data, Q2 50% of the data, also known as the median (the red line), and Q3 is the 75% of the data. Box plots may also have lines extending vertically from the boxes (whiskers) indicating variability outside the upper and lower quartiles. The end of the whiskers represent respectively the highest data within 1.5 IQR of the upper quartile and the lowest data within 1.5 IQR of the lower quartile. The points plotted outside the whiskers are marked as outliers. Furthermore, the position of the box also indicates how the data are distributed given a variable. If the box is on left, the distribution is skewed on the right, if the box is on the right, the distribution is skewed on the left. So just using the boxplot, I had a better vision of how are structured my data according the variable specified, there are outliers and only number of bedrooms seems to be normally distributed.





Correlation Analysis And Redundancies

After seeing the distribution of my variables, I computed the correlation between the different attributes and how each one is correlated to the price. Indeed, two attributes highly correlated (not using price) could be useless because they will not have a great impact on the regression result and should be reduced to one (cf: Principal Component Analysis). On the other hand, an attribute low correlated with the dependent variable (Price) could not be really influent on the result.



#saleprice correlation matrix

Outliers Detection



Outliers Detection According to different measures and charts of my data, there was no doubt about the presence of outliers. The outliers, in my case, are tuple taking abnormal value such as very large or very small, even 0, in one or many of variables. These outliers can affect greatly the results of my learning algorithm. They are several types of outliers :

- 1. Univariate, outliers having an extreme value on one variable.**
- 2. Multivariate, outliers having a combination of unusual scores on at least two variables.**

Negatively Skewed, apply power greater than 1 for every value of a given attribute (x , x , etc..)

- Positively Skewed, apply power less than 1 for every value of a given attribute (x , x , x , tc...), or try .





Data Exploration



Data exploration is the first step in data analysis and typically involves summarizing the main characteristics of a data set, including its size, accuracy, initial patterns in the data and other attributes. It is commonly conducted by data analysts using visual analytics tools, but it can also be done in more advanced statistical software, Python. Before it can conduct analysis on data collected by multiple data sources and stored in data warehouses, an organization must know how many cases are in a data set, what variables are included, how many missing values there are and what general hypotheses the data is likely to support. An initial exploration of the data set can help answer these questions by familiarizing analysts with the data with which they are working.





Data Selection



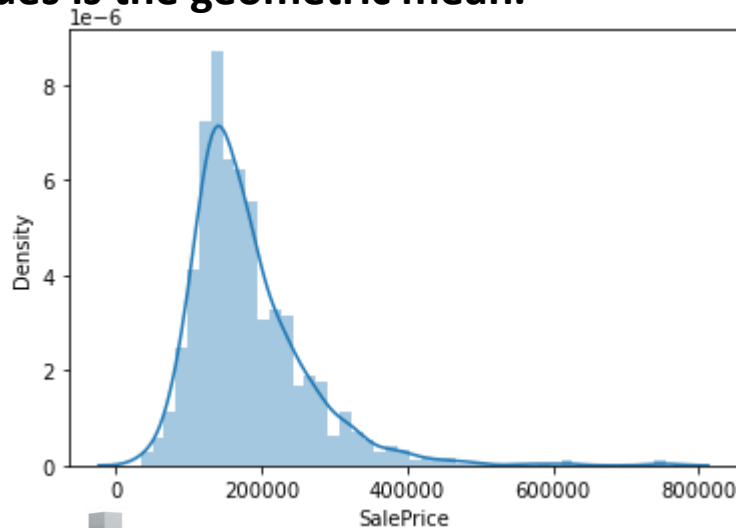
Data selection is defined as the process of determining the appropriate data type and source, as well as suitable instruments to collect data. Data selection precedes the actual practice of data collection. This definition distinguishes data selection from selective data reporting (selectively excluding data that is not supportive of a research hypothesis) and interactive/active data selection (using collected data for monitoring activities/events, or conducting secondary data analyses). The process of selecting suitable data for a research project can impact data integrity.

The primary objective of data selection is the determination of appropriate data type, source, and instrument(s) that allow investigators to adequately answer research questions. This determination is often discipline-specific and is primarily driven by the nature of the investigation, existing literature, and accessibility to necessary data sources.

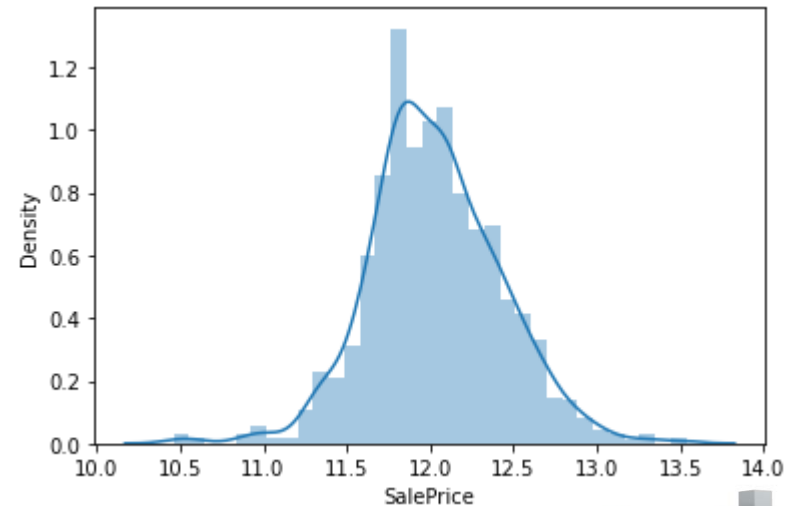


Data Transformation The log transformation

can be used to make highly skewed distributions less skewed. This can be valuable both for making patterns in the data more interpretable and for helping to meet the assumptions of inferential statistics. It is hard to discern a pattern in the upper panel whereas the strong relationship is shown clearly in the lower panel. The comparison of the means of log-transformed data is actually a comparison of geometric means. This occurs because, as shown below, the anti-log of the arithmetic mean of log-transformed values is the geometric mean.



Skewed Price



Normal Price



Model Building and Evaluation

Ridge and Lasso Regression



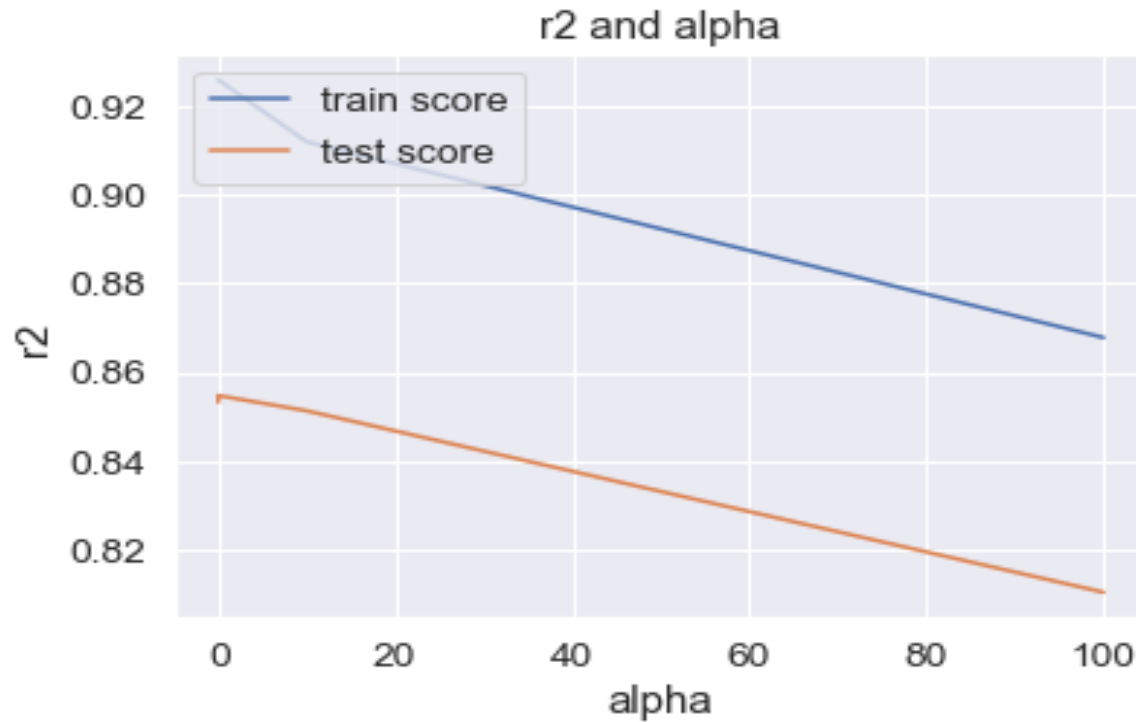
There are three popular regularization techniques, each of them aiming at decreasing the size of the coefficients:

Ridge Regression, which penalizes sum of squared coefficients (L2 penalty).

Lasso Regression, which penalizes the sum of absolute values of the coefficients (L1 penalty).



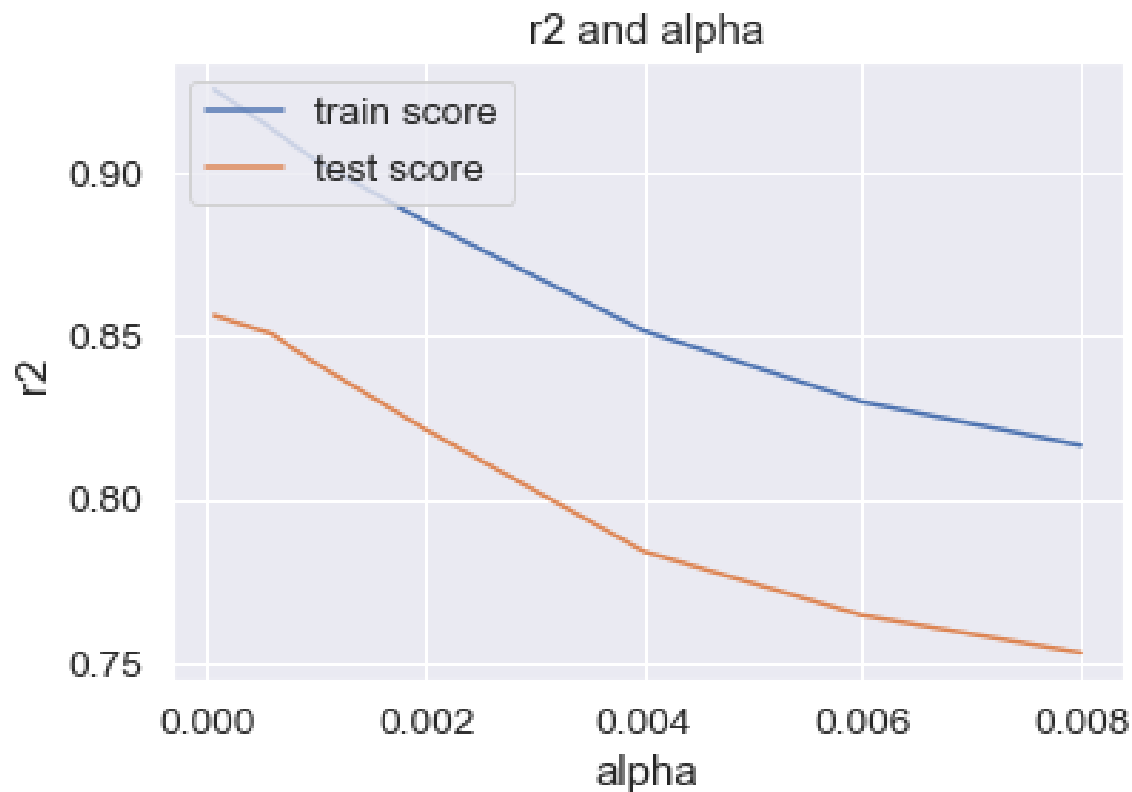
plotting mean test and train scores with alpha





Lasso

The goal of lasso regression is to obtain the subset of predictors that minimizes prediction error for a quantitative response variable. The lasso does this by imposing a constraint on the model parameters that causes regression coefficients for some variables to shrink toward zero.



Conclusion :



we got a decent score for both Ridge
and Lasso regression.

Ridge : Train :91.7 Test :75.8

Lasso : Train :90.1 Test :74.4





Top 5 most significant variables in Ridge are:

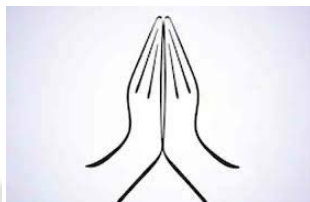
('SaleCondition_Partial', 0.143)

('SaleCondition_Others', 0.105)

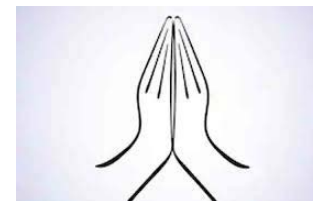
('SaleCondition_Normal', 0.099)

('GarageFinish_Unf', 0.094)

('GarageFinish_RFn', 0.092)



Thank You





These Variables are directly proportional to each other.



Optimal Value of lamda for ridge : 10

Optimal Value of lamda for Lasso : 0.001

Because of Feature selection as well we can choose Lasso regression in this case.

