METCS 521 Final Project

Predicting Housing Prices with Linear Regression

Authors: KARTIK BALAJI KUNDETI and XIAN LI

INTRODUCTION

Real estate companies attach great importance to the sales price. In the actual sales process, there may be many factors that can cause the change of the sales price. It is meaningful to explore which factors and variables will have an impact on the sales price and how much impact it will have, because through these results, real estate companies can adjust their company's strategy and invest more time, energy, and capital in more important factors, so that the company's sales price can be increased, so that the company can obtain more profits.

In this project we have created jupyter notebook scripts that takes the user through various steps including:

1. Downloading the dataset
2. An exploratory analysis to identify ideal variables for future prediction
3. Linear regression of variables and prediction of test dataset
4. Linear Regression class

By first exploring the data with a correlation analysis we identified variables that would be better at predicting the housing prices. We then used these variables and performed an ordinary least squares analysis and compared how well they predicted the housing prices in the same train and test dataset. Finally, we implemented a regression class that takes X variable, y variable, type of regression and mode of regression and trains and tests data and gives us the results.

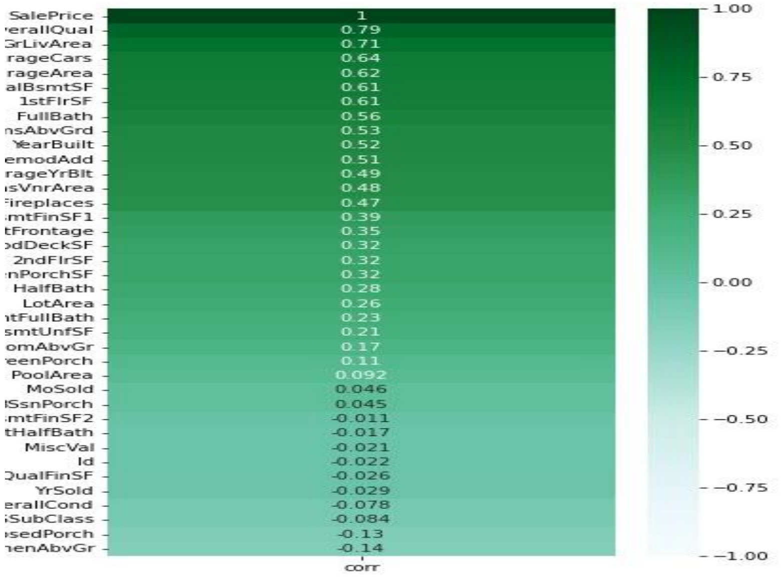
DOWNLOADING THE DATASET

To download the dataset the user needs to run the **Data Download.ipynb** script. The script will write a csv file to your working directory that will contain all the housing data. The file will also read this csv file into a dataframe to be used for exploration and prediction of housing prices. Each cell in the file should be run sequentially to reproduce the results in this paper.

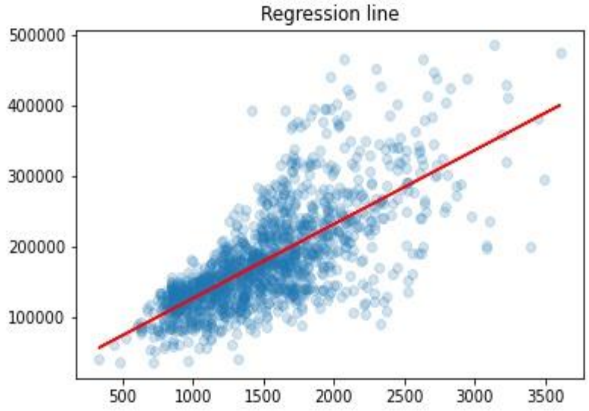
DATA EXPLORATION

In the process of exploration, we should first have an overall understanding of the data. After understanding the information contained in the data, such as the number of rows and columns, we observe the characteristics of the data and find that some variables in the data are obvious numerical variables, some are classified variables, but some variables whose output results only contain several levels can also be regarded as classified variables. So we use the dummy function to deal with these variables that are not obvious classification variables, turn them into classification variables, and then remove the Nas. Later, we selected some variables that have more influence on the ‘SalePrice’ variable through correlation and drew some pictures about these variables.

Here is a graph about correlation, which intuitively shows the correlation between different variables and the ‘SalePrice’ variable. You can see that the first one is ‘SalePrice’ itself, and the correlation is 1. And we combine the information judgment of logic and data, and select some variables as the input variables of the regression equation.
For example, we chose the ‘GrLivArea’ variable as one of the input variables. Firstly, because the correlation is very high, it is 0.71, and this is in line with common sense. Generally, when the above grade (ground) living area square feet is larger, the house price will be higher, and the data type of this variable is an obvious data variable, which can be used as an independent variable of the regression equation.



Continuing to take this variable as an example, we draw the scatter diagram of this variable and the ‘SalePrice’ variable, and draw the fitting curve. After removing some extreme values, the image performs very well and the fitting effect is very good. It can be seen in the figure that the main points are distributed around the fitting curve.



LINEAR REGRESSION

For simple linear regression, we first consider **“Sale Price”** as the **dependent variable** and **“GrLivArea”** (above ground living area) as the **independent variable**. Regression gives us an R-squared of 0.5. For this, I first create a new data frame with only the two required variables. As we have some outliers, we clean the data to remove values above 4000 for GrLivArea and above 500000 for SalePrice. We choose these specific values because, for GrLivArea, plotting a boxplot shows that there are only 5 points above the value of 4000; all the other values which are 1.5 standard deviations above the mean are closer, and there are a considerable number of such values. Removing all such values will leave us with less data. For the same reason, all the values above 500000 are removed for the variable SalePrice. Our data cleaned, we proceed to perform the OLS regression. Fitting the model gives us the value of intercept to be 20483.69 and that of the coefficient of GrLivArea to be 105.18. While the intercept does not have a meaning by itself, the interpretation of the coefficient is this: an increase in the average above ground living area by one unit increases the sale price of the house by $105.18. The R-squared value of this model is 0.510. That means, about 50 percent of the variation in sale price of the house is explained by the above ground living area.

TEST AND TRAIN

In an attempt to improve the model, we try the “test and train” method. These two together form an important part of machine learning. As a part of this process, we divide the data at hand into two (sometimes it can be divided into 3 parts). One part of the data is used by the algorithm to learn the data and the other part of the model is used by the algorithm to test the data. This is done to avoid overfitting of the model which, if present, will pose a problem to the predictive power of our model. Doing this revised our coefficient estimate to $106.75. So an increase in the average above ground living area by one unit increases the sale price of the house by $106.75.

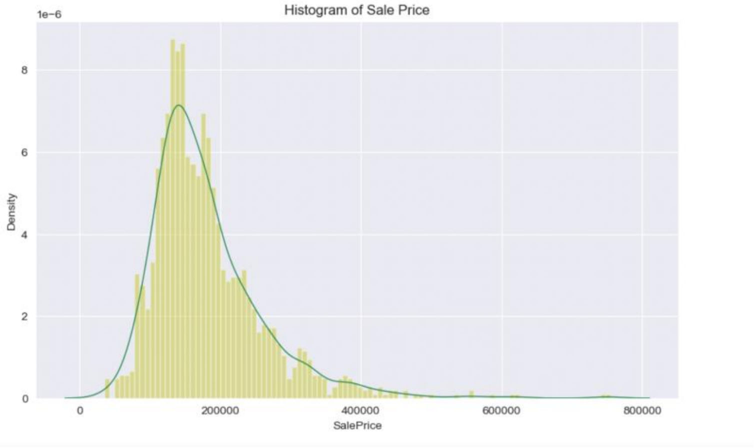
Next, we perform multiple regression. For this, in addition to using GrLivArea we have to decide what variables to include in our model. In doing so, we create a correlation heatmap for all the (numeric) variables present in our data set to pick out the most relevant (in that they have high correlation with SalePrice) variables. First criteria we set is that our correlation coefficient should be higher than 0.5. And the second criterion is that they are continuous variables. This is because using categorical variables makes our regression equation complicated. Which leads us to have **GrLivArea, GarageArea, TotalBsmtSF** in our regression. Additionally, we keep one binary variable (**CentralAir**, which is 1 if the house has a central air conditioner and 0 if not) because it is reasonable to expect that central air conditioner goes into choices for people's decision to buy a house. With our variables selected, we first create dummies for CentralAir, as they are in the format of Y and N. We use 1 for Y and 0 for N. Next, we clean the data to remove outliers. After that we train the data and then perform the OLS regression. The p-values of all the independent variables is 0.00, which means the coefficients are statistically significant at 5 percent level of significance. Also, F-stat is also significant at 5 percent level (even at 1 percent level) which means that the coefficients are also jointly statistically significant from 0. The R-squared value of this model is 0.7. So, this model explains about 70% variation in SalePrice. To compare this one with our linear regression model, however, we need to consider the adjusted R-squared value, as the r-squared value generally increases with the number of independent variables. So the adjusted R-squared value for this model is 0.711 as opposed to that of 0.506 in the previous model. Clearly this is a better model.

EXTENSION: REGRESSION CLASS

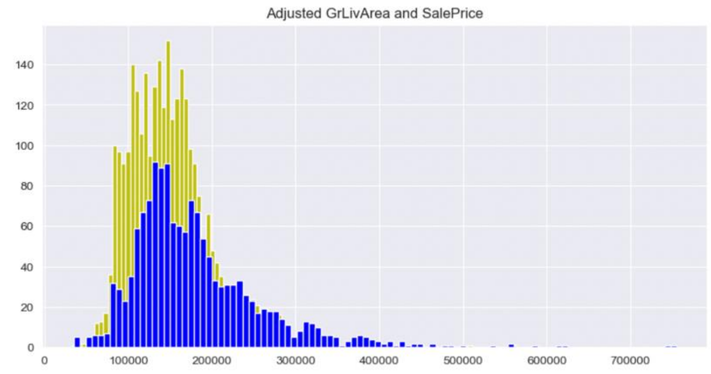
For one of the two extensions, we implement a linear regression class. For this we first define the init method with input parameters of self, X which takes all the dependent variables, y which points to the independent variable, model pertains to the model we are using (which here is defined by default as LinearRegression), and reg_type refers to the type of regression (that is, whether it is simple linear regression or multiple regression). In the init method we also define test, train and split, specifying that 80 percent of the data should be used for training and the remaining 20 percent for testing. Next, we write a method for training and predicting. In this, we fit the model with the tested X as the independent variable(s) and the tested y as the dependent variable. We also define the y_predicted which simply is that we are now testing our trained model on the 20 percent of the data we earlier mentioned. After this, we write another method for visualizing the results. Here, we want to visualize the intercept and coefficient(s) of our model and for the linear regression case, how predicted and actual data plots against the X-values. In the first condition, we print our intercepts and coefficients. Since for both linear regression and multiple regression we have only one intercept, printing it is easy. However, we have more than one coefficient for multiple regression. For that we specify that if there is only one intercept (linear regression case) print it and if there are more than one, we convert them into string and print them by separating each one of them with a comma. After this, for the linear regression case, we plot a scatter plot showing how actual and predicted values of the dependent variable vary with the independent variable. In the penultimate method, we write a score method, which will show us two specific scores in understanding our model. The first one is the R-squared value and the second one is the explained variance. While both values try to explain how much of the variance in the dependent variable is captured by the independent variables, the difference is that the explained variance uses the biased variance to determine what fraction of the variance is explained. R-Squared uses the raw sums of squares. If the error of the predictor is unbiased, the two scores should be the same. Finally, we define an execute method to execute our regression. In this way, by simply running our RegClass class, we can test and train the data and obtain predictions for our model. An important caveat here is that, before using this linear regression class, it is important to make sure that there are no missing values in the data. It is also advised that the user clean/filter the data before running the regression.

EXTENSION: VISUALIZE AND EVALUATE THE DATA IN NEW WAYS

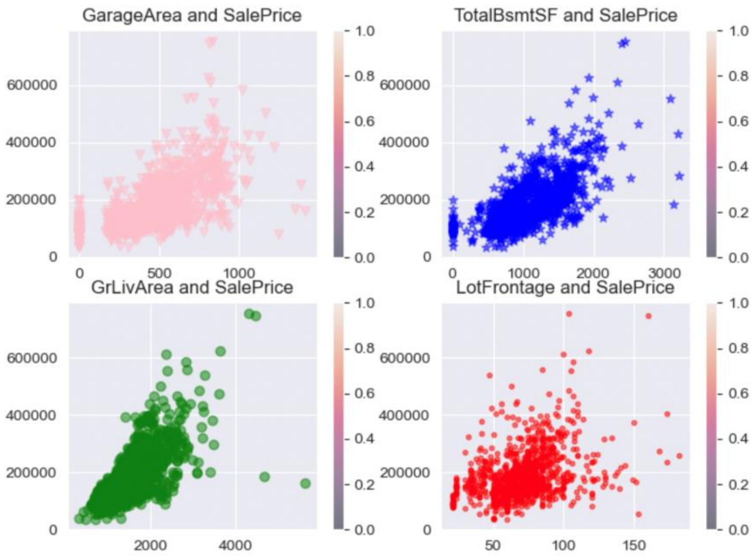
From this graph, we can see the distribution of house sales prices in different ranges. Most of the prices are about $200000. The difference between this figure and the previous figure is that the fitting curve is added to make it easier to see the change trend.



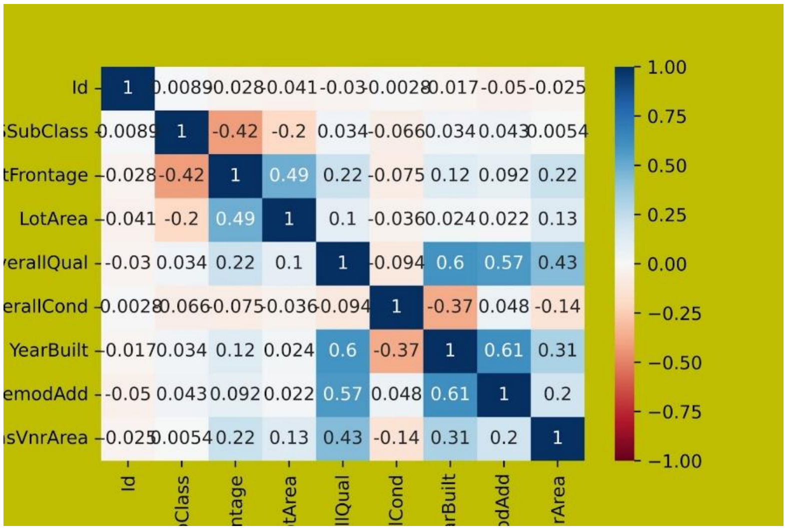
The graph below shows GrLivArea and SalePrice variables together in the form of histogram. Because the order of magnitude is different, I adjusted GrLivArea variable because the purpose is to observe the change trend of the two variables, and the change of absolute value does not affect the observation. It can be seen that the distribution of the two variables is similar.



In the graph below, I selected four variables with high correlation coefficient with sales price. We can see the relationship between them and the sales price. We adjusted the data and deleted some extreme values to make the image look better. And put the relationship images between different variables and sales price together, which can be compared more intuitively. It can be seen that the relationship between LotFrontage and sales price is not as close as other variables, because its points are more scattered in the graph.



In addition, we also made a Heatmap image below(Part of the image). Show the correlation between all variables in a more intuitive way. It can be seen from the image that different colors represent different correlations, and the darker the color, the higher the correlation. Because the result of the diagonal is the correlation of the same variables, the color of the diagonal is the deepest, representing the correlation of 1.



	Id	Class	ntage	tArea	lQual	lCond	arBuilt	odAdd	rArea
Id	1	0.0089	0.028	-0.041	-0.03	-0.0028	0.017	-0.05	-0.025
SubClass	0.0089	1	-0.42	-0.2	0.034	-0.066	0.034	0.043	0.0054
tFrontage	0.028	-0.42	1	0.49	0.22	-0.075	0.12	0.092	0.22
LotArea	0.041	-0.2	0.49	1	0.1	-0.036	0.024	0.022	0.13
verallQual	-0.03	0.034	0.22	0.1	1	-0.094	0.6	0.57	0.43
erallCond	0.0028	0.066	-0.075	-0.036	-0.094	1	-0.37	0.048	-0.14
YearBuilt	0.017	0.034	0.12	0.024	0.6	-0.37	1	0.61	0.31
emodAdd	-0.05	0.043	0.092	0.022	0.57	0.048	0.61	1	0.2
sVnrArea	0.025	0.0054	0.22	0.13	0.43	-0.14	0.31	0.2	1

CONCLUSION

This paper mainly studies the impact of different variables on real estate prices through regression method, simple linear regression studies the impact of a single variable on sales prices, and multiple linear regression studies the impact of multiple variables on sales prices. In this process, we convert some classification variables in order to use them in the regression equation. Finally, we get some results and verify the significance and R-square value to ensure that our results are meaningful.