Task 2 - Data Science for Engineers

This document will consist of two brief parts. In the first part a model was developed only using continuous variables, since it is what I am most used to; in the second part a more comprehensive model was created, which includes both continuous and categorical variables to achieve a model with a better R-squared, which means a model that fits the data better (Ogee, A. et al, 2013).

Both parts are composed not only by the model development, but also by an exploratory data analysis section that helped me choose the best predictors for the data to the best of my knowledge. Note that only the most relevant annotations will be made, for more detail please check the attached Jupyter Notebook files.

First Part: Model based on Continuous Indicators

Exploratory Data Analysis (EDA)

I decided to use the Pearson correlation coefficient to check for both multicollinearity and correlation. Some of my insights were the following:

1. Regarding correlation, I decided it would be great to use Rooms, Bedroom, Bathroom, Car and Building Area, since they hold a significant relation with Price.
2. When checking for multicollinearity we might stumble across an error if we decide to use both Rooms and Bedrooms variables, since they are highly correlated (which makes sense, since one contains the other). I will stick with Rooms since it has a higher correlation with Price. It happens similar with Bathrooms, so I will try both combinations.

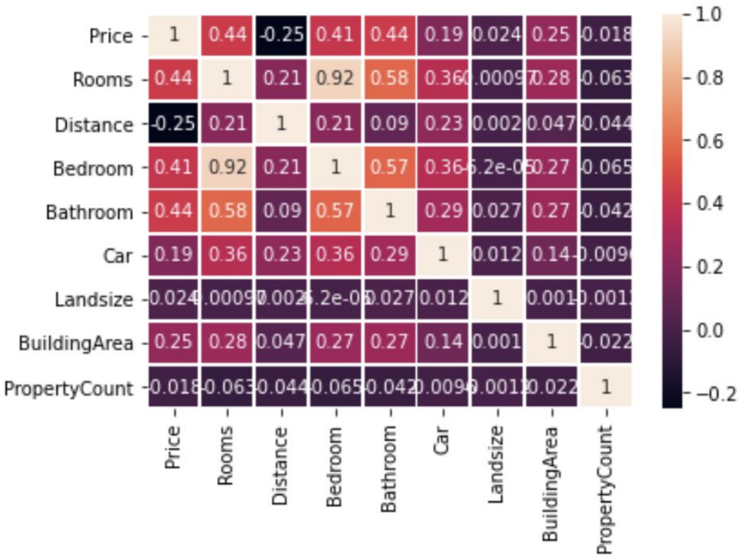


Image 1. Heatmap of Pearson correlation

Model Development

As a first task, I had to clean the data. The first thing to do was to fill empty cells on numerical fields, such as Car or Building Area variables. To do this I decided to use the median of that column to replace NaNs, which is a common and safe technique to handle empty cells and avoids errors on Pandas (Ohri, S. 2020).

After that, I developed two models with the variables I decided to test on the EDA section. On the first model I decided to use Rooms, Distance, Bathroom, Car, Landsize and BuildingArea, I followed a simple process which consisted of three additional steps.

1. By using Scikit-learn model selection I splitted the data as follows: 40% of it was destined for testing and 60% of it for training the model.
2. With my X and Y trained, which correspond to the independent and dependent variables respectively, I launched a Linear Regression model.
3. Once I had the model, I obtained the following equation:

$$Price \;=\; 360,600 + 236,505\ Rooms - 43,529\ Distance + 225,562\ Bathroom + 52,646\ Car + 2.37\ Landsize \;+\; 531\ BuildingArea$$

As well as the following information:

$$R^2 = 37\%$$

$$R^2(adj) = 36.90\%$$

The dropout between R^2 and $R^2(adj)$ is small and the plotted model is good. Scikit is a great tool because we don't have to test ourselves for evidence on the predictions, rather we can plot and analyze the results of them with Seaborn, as shown.

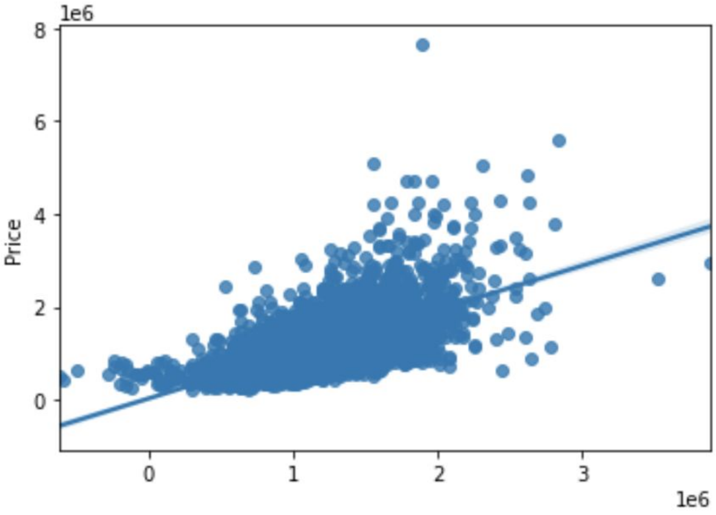


Image 2: Predictions vs. our y_test

As I said earlier, I also developed the model without Bathroom variable, to avoid multicollinearity, but it seems to be a worst model if we do this since it provides an R^2 of 33%. So I will stick with the first model presented. Even though we don't have to test for evidence regarding our predictions, since Scikit did that for us, I will do one hand written prediction below:

$$rooms = 3\ ,\ distance = 2.5,\ bathroom = 3,\ car = 1,\ landSize = 166,\ buildingArea = 174$$

$$price(y) \;=\; 360,564 + 236,505 * 3 - 43,529 \;* 2.5 + 225,562 \;* 3 + 52,646 \;* 1 \;+\; 2.37 * 166 + 531 * 174$$

$$price(y) = \gamma \;=\; 1,783,285$$

The expected result was 1,447,500, which is very close to our prediction and we are not even considering the standard error. So overall it was a **good prediction.** With the above prediction, as well as Image 2, where we can observe out scatter plot follows the regression line, we can say we have a **good model,** but it definitely can be improved since the R^2 was not that high. Let us continue with a model that contains both continuous and categorical predictors.

Second Part: Model based on Continuous and Categorical Indicators

Exploratory Data Analysis

In this part, choosing which predictors to use was not an easy task, but after a good analysis I got a nice set of variables. First I decided to ditch variables that had repetitive, useless or uncategorizable data, such as Latitude, Longitude, Address and Date. As I already knew, Pearson's correlation won't work for this situation since it is not suitable for categorical variables so I decided to develop scatterplots to check for correlation. I did this for many variables.

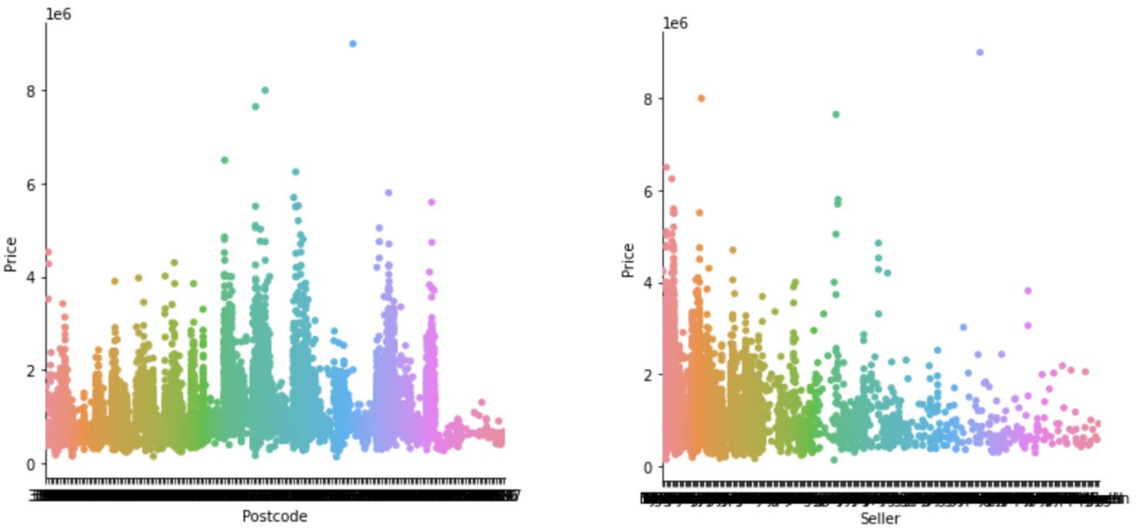
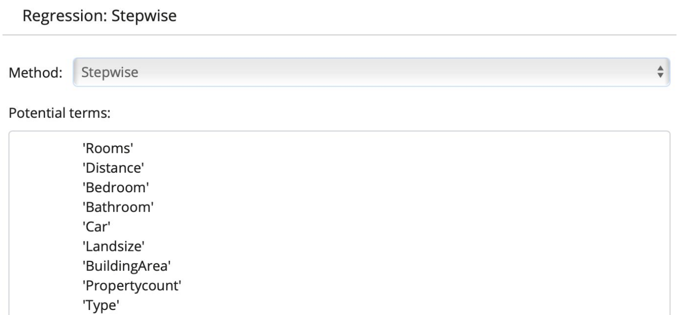


Image 3. Seller and Postcode scatter plot

For example, with the information above, I decided to keep variables such as Seller, because we can see a linear correlation, and to ditch sparse variables as Postcode.

Running a Chi Square was a great option to test for relationships between the categorical variables (StatisticsSolutions, 2021), but I decided to use an external tool to help me keep choosing the adequate variables. I used Minitab's Sepwise Procedure to help me choose the variables that had a p value lesser than .15, which is acceptable (please see Minitab files in repository). This tool performs all the combinations not only to check for correlation but to avoid multicollinearity and returns the best set of variables to use. With this information I then proceeded to develop the model with Scikit.



Regression: Stepwise

Method: Stepwise

Potential terms:

'Rooms'
'Distance'
'Bedroom'
'Bathroom'
'Car'
'Landsize'
'BuildingArea'
'Propertycount'
'Type'

Image 4. Stepwise Procedure in Minitab

Model Development

The followed procedure was very similar, but with slight changes on the categorical variables.

1. I first converted categorical variables into dummies, by using the pandas.get_dummies function, which returns a new data frame. Remember variables used were obtained from Minitab's analysis.
2. I then used Scikit-learn model selection and splitted the data in the same proportion.
3. With my X and Y trained I launched a Linear Regression model.
4. Once I had the model, I obtained a pretty big equation, whose coefficients can be found in this link and with an intercept of $231,409$. As well as the following information:

$$R^2 = 65\%$$

$$R^2(adj) = 63\%$$

The dropout between R^2 and $R^2(adj)$ is small and both R^2 and $R^2(adj)$ are significantly higher than the ones obtained in the previous model. So overall, this model is better.

Now let us run manually one prediction, using the whole set of variables from the model to check for the precision of our model. Remember obtained coefficients are in this link.

$$rooms = 2,\ distance = 5.5,\ bathroom = 1,\ car = 3,\ landSize = 453,\ buildingArea = 93,\ propertyCount = 11$$

$$type = h,\ method = S,\ seller = Barry,\ councilArea = Darebin, regionName = Northern\ Metropolitan$$

$$price(y) \ = \ 595,512 + 166,022 * 2 - 40,784 \ * 5.5 + 161,189 \ * 1 \ + 53,638 * 3 + 1.66 * 453 + 370.54 * 93$$
$$- 1.78 * 11,364 + 101,643 * 1 - 86,810 * 1 + 124,\ 200 * 1 - 148,531 * 1$$

$$price(y) = \gamma = \ 1,030,833$$

The expected result was 1,117,000, which is very close to our prediction and we are not even considering the standard error. So overall it was a **good prediction.**

$$rooms = 3,\ distance = 14.5,\ bathroom = 2,\ car = 2,\ landSize = 650,\ buildingArea = 150,\ propertyCount =$$

$$type = h,\ method = S,\ seller = Nelson,\ councilArea = Brimbank, regionName = Western\ Metropolitan$$

$$price(y) \ = \ 595,512 + 166,022 * 3 - 40,784 \ * 14.5 + 161,189 \ * 2 \ + 53,638 * 3 + 1.66 * 650 + 370.54 * 150$$

$$- 1.78 * 1,119 + 101,643 * 1 - 38,109 * 1 - 90,499 * 1 - 168,382 * 1$$

$$price(y) = \gamma = 844,823$$

In this second prediction, the expected result was 773,000, which was also considerably close to our prediction. So overall it was a **good prediction.** With the above predictions, as well as our high R^2 and $R^2(adj)$ we conclude we have a **good model.**

While dealing with categorical variables can be hard, there are a lot of resources out there that we can leverage, though knowledge on statistics and its basic concepts is required. As future work I think we could skip the use of Minitab by using other Scikit feature selection packages such as SelectKBest and chi2.

References

1. A. Ogee, et al. “Regression Analysis: How Do I Interpret R-squared and Assess the Goodness-of-Fit?,” Minitab Blog, 30-May-2013. [Online]. Available: https://blog.minitab.com/blog/adventures-in-statistics-2/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit#:~:text=R%2Dsquared%20is%20a%20statistical,multiple%20determination%20for%20multiple%20regression. [Accessed: 31-Jan-2021].
2. S. Ohri, “Pandas: Replace NaN with mean or average in Dataframe using fillna(),” *thispointer.com*, 24-Aug-2020. [Online]. Available: https://thispointer.com/pandas-replace-nan-with-mean-or-average-in-dataframe-using-fillna/. [Accessed: 31-Jan-2021].
3. S. Solutions, “Using Chi-Square Statistics in Research,” *Statistics Solutions*, 09-Apr-2020. [Online]. Available: https://www.statisticssolutions.com/using-chi-square-statistic-in-research/#:~:text=The%20Chi%20Square%20statistic%20is,the%20population%3B%20they%20are%20independent. [Accessed: 31-Jan-2021].