U1

Rodriguez González José Adrián

August 2024

1 Abstract

In data analysis we need to understand that every analysis problems requires its own time before to create a predictive model. This is because we need to verify if our model would it be useful. We have two problems to practice that it has been learned about data analysis. As first checkout it'll be a simple dataset that only have two features *xandy*. So at the first sight seems to be quick but before to create a predictive model, it'll be required to take a look into the data as first step. And also, we got another case about predicte the price of cars.

2 Introduction

With the objective to checkout how does it work the linear regression and look out for troubles on the real life. It has been presented two datasets. The first one consist in sintetic data. And the seconde one is related wth the prices of car, and involves several features and the main wit these datset is create a predictor that may predict future data.

3 Related works

Some of the related works(by now, I'll just let some of the books and the documentation that I could to see)

4 Methods and materials

The material used for this analysis were the usage of python and it libraries for data analysis and machine learning:

- Numpy: For mathematical calculations
- Pandas: For handle datasets
- scipy: to evaluate statistical parameters

- Scikit-learn: to train our model and check more parameters related with the model chose.

The methodology followed it's the mix of scientific methodology with the abstractacion for a data scientist.

- Obtain the data
- Make an exploration into the data.
- Check various parameters from the dataset.(These parts involves most of the section of exploratory anaylisis, as also, this step gives several hypothesis to check out at the dataset)
- Now that we have our Hypothesis planted, and also, with the help of th last step that it can be plotted the data. Now it'll be cleaned the data, and for this section it incquires in several steps
-
  - Hypothesis proposal
  - Transformations (logarithm, square, box-cox, capping and flooring)
  - Check the metrics of skewness, R^2, MSE,$RMSE$
  - Make the model of linear regression
- check its metrics
- Propouse another hypothesis.

These were the main steps followed at the work. The part of reading some material and check information from several sources, is mainly the first step before to begin the work. However if there's something missing when the hypothesis are covered and we are finding out in a rabbithole, it'll be useful to look for more sources to create other hypothesis

5 Data analysis

The first data analysis that has been studied was the dataset with sintetic data. As it had been mentioned, following the scientifical methodology we look for the data and it started with a data exploration. It obtained the following values ...

Now that we have assured that the data are completed we can see the value of the mean and the 50 percentile are quited different, mostly on the variable x, it requires to be cleaned the data. Nevertheless, it is useful to check the correlation with variable x to y too. The correlation matrix resulted on ...

The correlation is quite low between x,y. And also, something that we can notice with the plots that they are not normally distribuitted. The x-axis it seems more a exponential distribution than normal. So the pre-propressing procedure