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 second one is related wth the prices of cars, and involves several features and the main objective on these datasets is to 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)

Article	Year	Techniques	Data	Results
Consumers' preferences on the Swiss car market: A revealed preference approach	2019	hedonic pricing approach and linear regression	Were obtained on the auto-Schweiz website	How lastly the costumers have been prefered lighter cars than heaviear despite the increasing of the weight cars and how it is an important factor for the costumer the efficence on the fuel consumtion
İkinci El Otomobil Fiyat Artışına Etki Eden Faktörlerin Yapısal Eşitlik Modeli ile Tespit Edilmesi: Van İli Örneği(Determination of the Factors Affecting the Used Car Price Increase with the Structural Equation Model: The Case of Van Province)	2023	AFA, and structured equation model	Were obtaind due to surveyrs on the province of Van	It has been found how the several features as economics, marketing, strategies and supplying are correlated with the amount of prices, however the economics are is stronger against the other factors

4 Methods and materials

The material used for this analysis were the usage of python and its 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 chosen.

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.

Also, the main models that has been studied were Linear regression and Knn regresor

4.1 Linear regresion

When we try to represent something complex it is usual that it has to be create a structure that simplifies the process however, it must contain the enough data and values that can approachthe behaviour of phenomen. In science areas, it is commonnly called model. A model can explain a complex phenomen in simple terms that can be easy handled and understandable. One of the most simplest models that exist on data science is the linear regression. Knowing:

$$f : \mathbb{R} \to \mathbb{R}$$

we have:

$$f(x) = mx + b$$

This is a linear function(Figure 1), one of the simpliest functions that can be viewed on calculus, sometimes a simple phenomen can be viewed without getting in a huge amount of complexity and could be described with simple models, and when we have a problem or a situation in the reality when we need to create a machine hat can predict future data, what can we do? In those

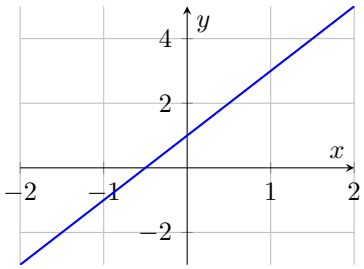

Figure 1: An example of a lineal function

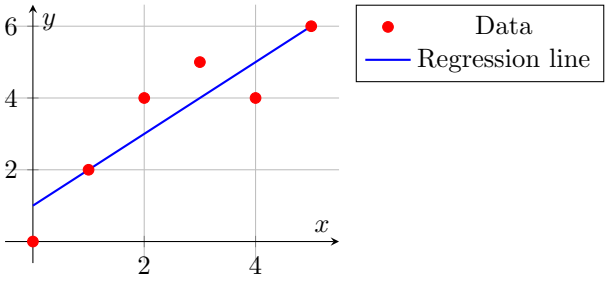

Figure 2: A simple example of a lineal regression

cases, as it has been portrayed, a linear function can be used as a mathematical feature that attempts to fit on the data. Despite that nowadays exist several types of models that can predict future data, lineal regression as it is one of the simpliest due to it just use a mathematical function that is easy to use, a lot of new models can be tested with a contraposition of lineal regression. So, the Lineal regression is described as

$$y = \beta_0 x + \beta_1 + e$$

β_0 it is the slope and β_1 is the intercept. And e is the error tha exist on the model.

However it'll be required to know some metrics that will help us to find out how efficent is our model and to know the error that we have in our model too

- R^2: This metric can measure how much the model can describe the phenomen. The $R^2 \in [0, 1]$ and if the model tends to 1 the model can describe better the phenomen and can be more precise in the predictions.
- MSE and $RMSE =$ Are metrics that measure the distance between the value predict against the real value.

$$MSE = \frac{1}{n}\Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

The difference between MSE and RMSE are that MSE is susceptible to detect outliers values due to the nature of a square function, and also it helps to know the variance that we got in our model. Nevertheless, RMSE use the same scale of the values and want to measure the error as it is on the data.

4.2 Knn model

It consist in to look out on the data and predict a value according to the vecinity of the values nearest from the data that is presented. Knn presents two basci hyperparameters, the first one is the number of vecinity and the second is the type of distante that it'll be chosen. Knn model has some different parameters if we measure it as classifier. However at this case it'll be used the Knn model as a regressor, so it'll be usable the metrics that it has been described before. (figure 3) After knowing the main features that were used at this study, let's

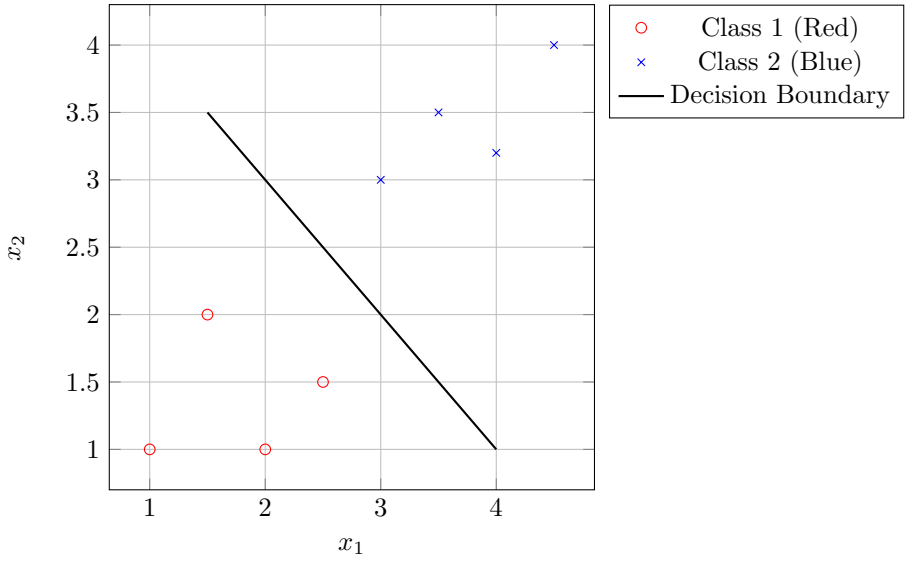

Figure 3: A simple example of knn regressor

look out for the process that has been followed. As it'll have seen, the process aims to use the scientifical method, however, with the feauters of a data analyst

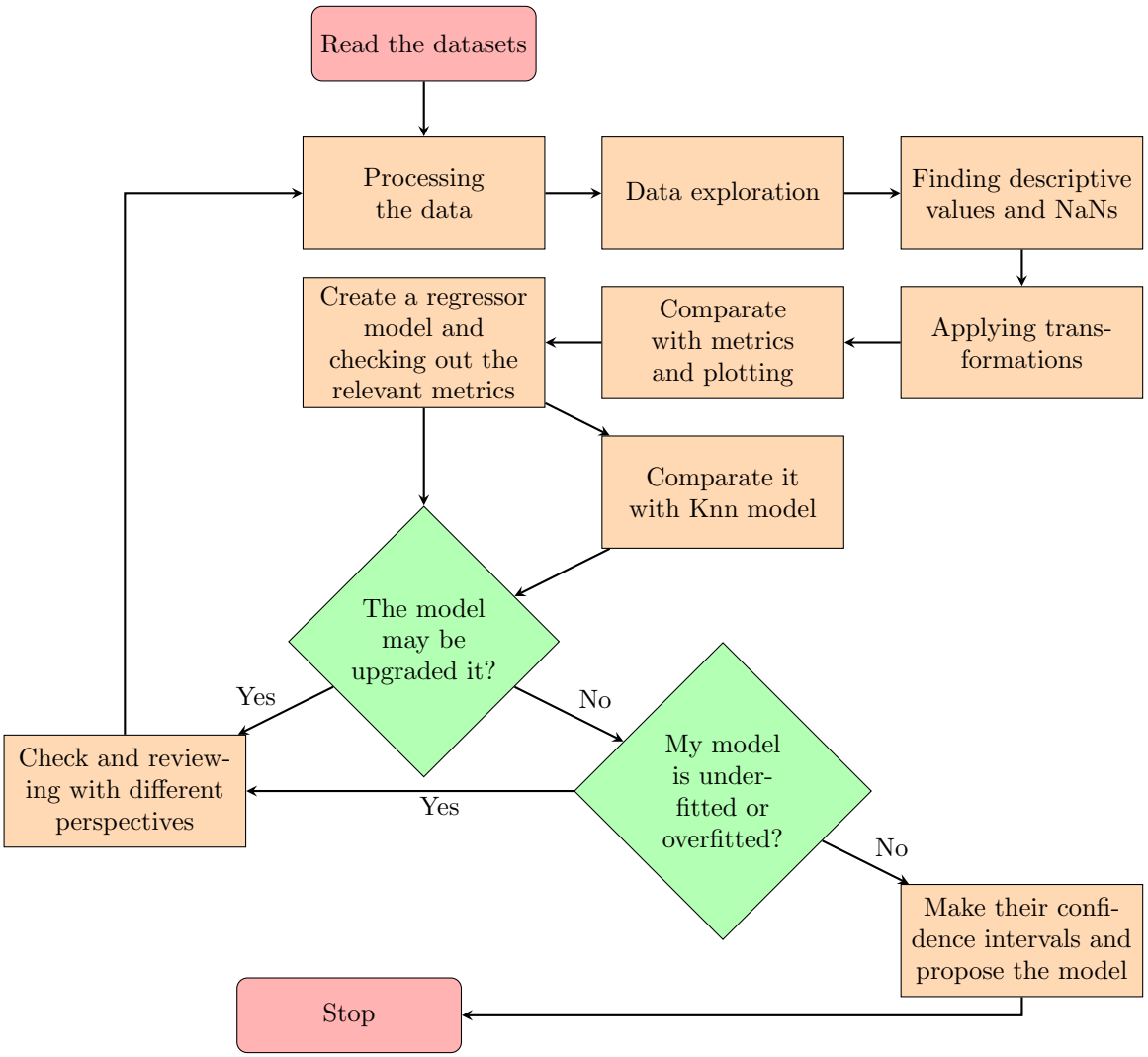

Figure 4: The procedural followed at the investigation

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, it'll be useful to look for more sources to create other hypothesis.

5 Data analysis

5.1 first dataset

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

parameter	y	x
number of values	506.000000	506.000000
mean	22.528854	3.613524
std	9.182176	8.601545
minimum	5.000000	0.006320
percentile 25	17.025000	0.082045
percentile 50	21.200000	0.256510
percentile 75	25.000000	3.677083
maximum	50.000000	88.976200

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

	y	x
y	1.000000	-0.389582
x	-0.389582	1.000000

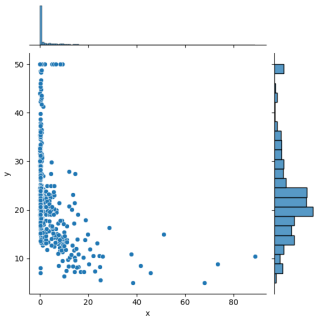

Figure 5: E

The correlation is quite low between x,y. That means that it'll be hard to find a relation with the x and y with linear regression. 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. Other way to demostrate it could making a hypothesis test and comprobating it with a Shapiro-walk; it consist to propose a null hypothesis, that if the value of statistic is closer to 1 or great, it will reject the hypothesis, therefore the distribution doesn't follow

a normal distribution. By other hand if we encounter a value smaller it will indicate is that probably the distribution follows a normal dsitrbution. And also we have de p value that if is greater than 0.05 will indicate that is a normal distribution, in the oppostive way it is a non-normal distribution. The test concluded with a distribution of $4.85 * 10^{-}28$ so it is not a normal distribution. At this point we can choose several ways to follow

- Trying to normalizing
- Trying with models more associated with exponential distributions or models that will no depend on the distribution.
- Infere with other features

Due to the objectives of the work, the first attempt it'll be used. At this dataset has been followed 4 approaches to look the efficence of methods at the cleaning of the data. However, it has been tested several types of transformations to check out if it'll be more useful the logarithm transformation according with the parameter of skew

type-skew	x	y
original	5.223148798243851	1.1109118502479587
log	1.2692005882725572	-0.24563979611568673
sqrt	2.024382103123676	0.4381663127860419
1/x	-0.5772191040682719	1.9393215717506038
box-cox	0.093649	...
outliers	0.40335	1.058543

It could seems some techniques are better however logarithm tranformation has a better resolution in the skew of both variables.

5.2 first method

The fisrt approach to make a model was transforming with logarithm and making a cross validation And therefore it was contrasted with the Knn model,

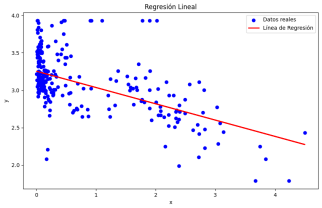

Figure 6: Lineal regression

getting the next values

metric	Lineal regression	Knn
R^2l	35.488840060699256%	43.57
MSE	0.09611374428381783	0.11904
RMSE	0.30971201639466434	0.33166

The model has been contrasted with both models and with the same cleaned data, as also trained at the same way respectively at the model However, the Lineal regressor has been predicting the data within the confidence. For example, it added a value x on 2.3 and the result in y is 2.7534 The confidence value on 95% interval is from $(2.172214750599946, 3.334689169873629)$ and is has been seen the value was fitted correctly. However it has been made a gen-

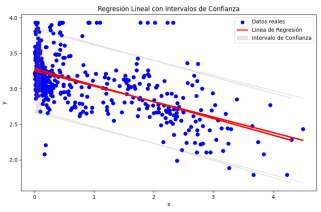

Figure 7: Lineal regresion with the intervals set

eralized model to linear regression in a set of dots (fig7) The equation for this model is

$$y = -0.21773690693332431x + 3.2542468461834333$$

5.3 second method

The second approach it consisted to just transforming the value with logarithm without CV 5x2

metric	Lineal regression	Knn
R^2	15.0589474050694%	22.562569548739275%
MSE	0.1127456496305317	0.11904364904719832
RMSE	0.3357761897909554	0.33166

5.4 third method

The outliers were fixed with capping and flooring methos using the median

5.5 fourth method

The last methods were used at this model

The methods were contrasted by the simplest model and also the intervals were obtained, however that section it comprehends the results discussion

And one of the most important things that iat has been obtained is this table that compiles all the metrics and parameters that has been obtained

Métrica	**model1**	**model2**
R^2	36.9%	39.9%
MSE	0.0961	0.1127
RMSE	0.3097	0.3358
EQUATION	$y = -0.2177x + 3.2542$	$y = -0.2479x + 3.2824$
Adjusted R^2	36.7%	39.8%
F-statistic	147.1	267.1
$P(F - statistic)$	5.97e-27	2.11e-46
Log-Likelihood	-50.079	-91.384
AIC	104.2	186.8
BIC	111.2	194.8
std error	const=0.024, x=0.018	const=0.019, x=0.015
σ	const=135.161, x=-12.128	const=170.036, x=-16.343
$P > \|t\|$	const: 0.000, x: 0.000	const: 0.000, x: 0.000
Confidence Intervals	const=[3.2068, 3.3017], x=[-0.2531, -0.1824]	const=[3.2444, 3.3203], x=[-0.2777, -0.2181]
Omnibus	41.738	37.308
Prob(Omnibus)	0.000	0.000
Jarque-Bera (JB)	64.331	64.192
Prob(JB)	1.07e-14	1.15e-14
Skew	0.960	0.581
Kurtosis	4.553	4.570
Durbin-Watson	0.983	1.902
Condition Number	2.26	2.16

Table 1: Comparación de los modelos 1 y 2

The next dataset that it has been studied was about cars. It is a car's magazine from 1985 that compiles prices and features from several cars. For this dataset it has been read some papers(include the cite) to understand some of the features and its relationship.

As a one of the first steps was the cleaning the data, and a data exploration was made to look for missing values and

6 Discussion

6.1 First data-set

6.2 Second data-set

7 Conclussions and future work

(Weber, 2019) (Polat and Bulut, 2023)

Métrica	**model3**	**model4**
R^2	33.8%	37.1%
MSE	0.014957	0.0807
RMSE	0.12224	0.2839
EQUATION	$y = -0.30537x + 1.85673$	$y = -0.2052x + 3.2467$
Adjusted R^2	33.5%	36.8%
F-statistic	128.2	147.8
$P(F - statistic)$	2.76e-24	4.72e-27
Log-Likelihood	181.27	-32.091
AIC	-385.5	68.18
BIC	-351.5	75.25
std error	const=1.8567, x=-0.3054	const=0.022, x=0.017
σ	const=0.015, x=0.027	const=144.464, x=-12.159
$P > \|t\|$	const: 0.000, x: 0.000	const: 0.000, x: 0.000
Confidence Intervals	const=[1.8279, 1.886], x=[-0.3584, -0.2522]	const=[3.2024, 3.2909], x=[-0.2384, -0.1719]
Omnibus	40.637	35.555
Prob(Omnibus)	0.000	0.000
Jarque-Bera (JB)	56.377	46.953
Prob(JB)	5.73e-13	6.37e-11
Skew	1.021	0.938
Kurtosis	4.087	3.966
Durbin-Watson	0.966	0.958
Condition Number	4.45	2.26

Table 2: Comparación de los modelos 3 y 4

References

Polat, M. and Bulut, C. (2023). İkinci el otomobil fiyat artışına etki eden faktörlerin yapısal eşitlik modeli ile tespit edilmesi: Van İli Örneği. *Sosyoekonomi*, 31(55):347–369.

Weber, S. (2019). Consumers' preferences on the swiss car market: A revealed preference approach. *Transport Policy*, 75:109–118.