Evidence 2. Data Science Project

Mathematics and Data Science for Decision Making

Data-Driven Nutrition: Harnessing Data Science for Healthy Living

Febrero-junio 2023

Gabriel Eduardo Meléndez Zavala A01638293

Miercoles 14 de junio de 2023

Phase 1. Business Understanding

The client could be anyone who decides to have their diet tracked for 14 weeks to predict a change in their person. This analysis would be trying to solve uncertainty people have regarding their diets, some might be doubtful that if their diet works while others worry about how bad or inconsistent their current diet is. Data Science will provide a robust solution based on the data it analyses, the solution it gives might give people relief that their diet is working or encourage people to change their current diet for a more balanced one. We need to learn what the desired change in person the client wants (the objective) and their diet for the past 14 weeks for a proper analysis. Fundamental understanding of statistics and data science, base understanding of nutrition and a balanced diet, and proper software to collect and analyze data to give well-made predictions.

Phase 2. Data understanding

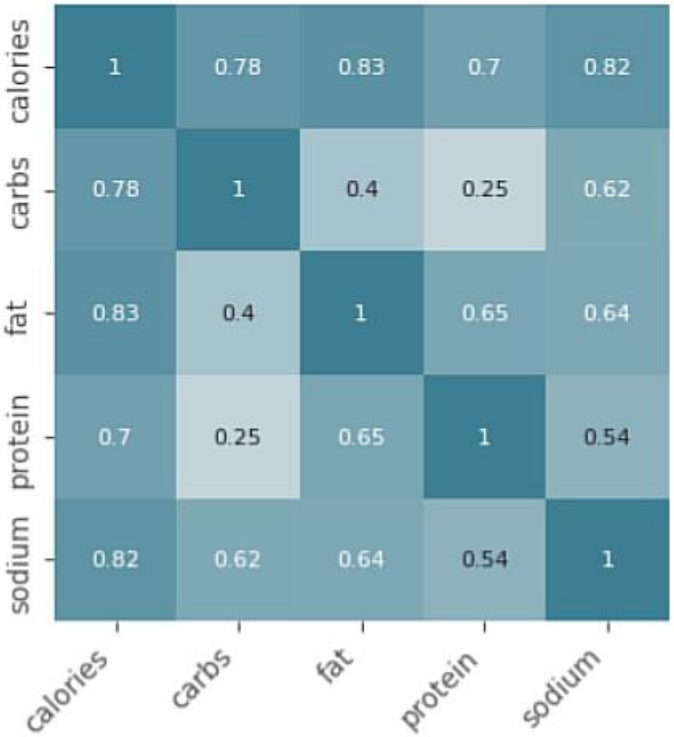
The type of data needed to push forward the idea is Numerical continuous data. The user must record every meal they eat for meaningful results to be given. The data is then obtained from what the customer provides. All the data gathered will have meaning towards the final analysis. The most promising attributes in the database are Calories, Carbs, Fat, and Protein. I believe this will provide valuable significance to the data analysis. The attributes that seem irrelevant are sodium which can greatly vary between data. The goal is to have at least 100 data entries and on my personal analysis there are more than 30 data entries which would be the minimum to achieve relevance for multiple analysis.

Phase 3. Data preparation

In my case there is no need to merge the data because there is no other database to merge columns or data entries with the original data. Since the only client in this project is me, there are no other data entries or parameters required in this project. In the case of subsets, I didn’t use them to discriminate a certain data group, however they can be applied for training, testing, and validation. They are used to facilitate efficient analysis, focus on relevant information, reduce computational complexity, and address specific research or analysis objectives. Moreover, regarding adding more data, we can add more data and it can affect the analysis. However, it is not always necessary or feasible. We must consider the specific context, goals, and available resources to determine whether adding more data is appropriate for the analysis. The data attributes for the analysis are correct and none need to be modified for a specific purpose. Each attribute is used as it was intended, and no changes were required. In addition, the data does not need to be sorted to perform the analysis, no matter the order the analysis will continue to be relevant and output meaningful results. The data was complete, therefore no blank values needed to be removed, replaced, or dealt with in an appropriate manner.

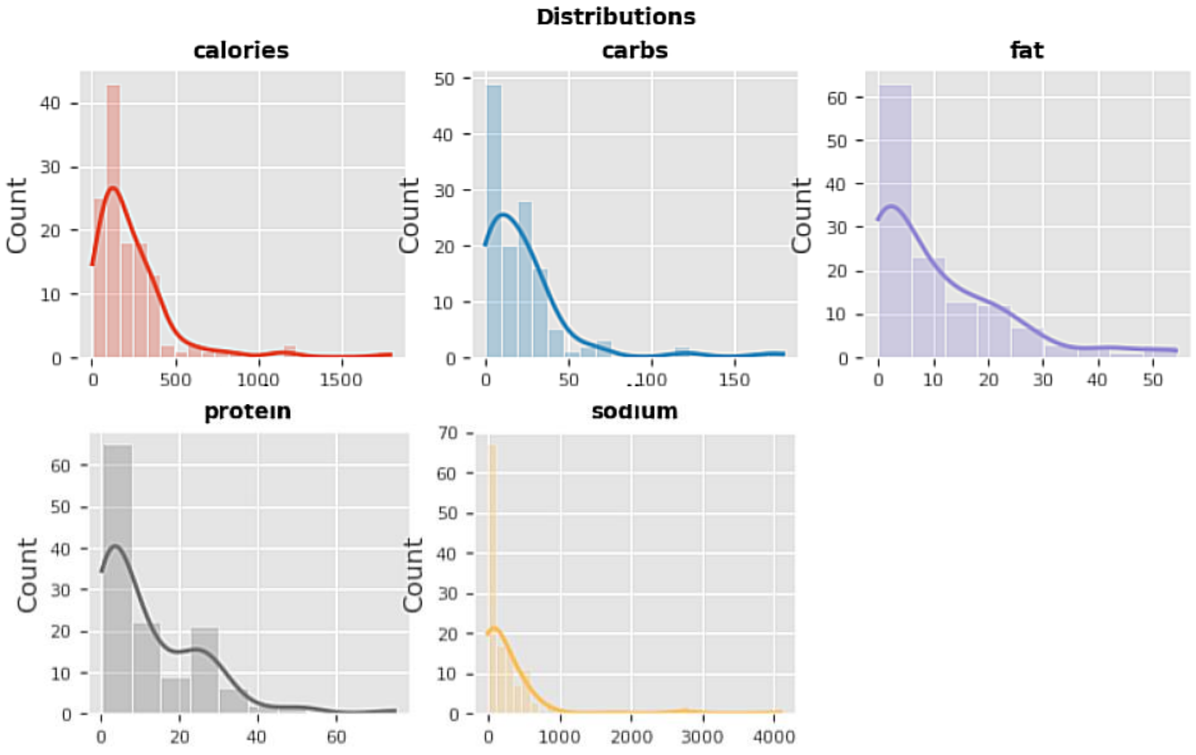
Phase 4. Data modeling

As part of my analysis, we obtained various results ranging from heat maps, distributions, and regression analysis using jupyter notebook as well as libraries like numpy, matplotlib, pandas, and sklearn.



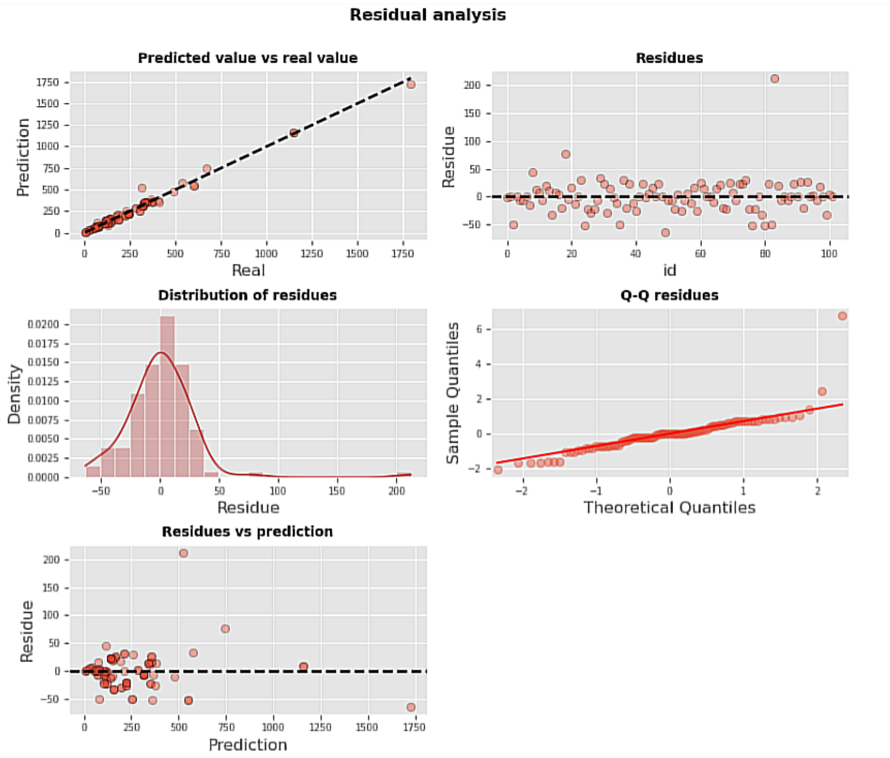
	calories	carbs	fat	protein	sodium
calories	1	0.78	0.83	0.7	0.82
carbs	0.78	1	0.4	0.25	0.62
fat	0.83	0.4	1	0.65	0.64
protein	0.7	0.25	0.65	1	0.54
sodium	0.82	0.62	0.64	0.54	1

(This is a heatmap for the correlation matrix between each variable.)



(Distribution plots for each numerical variable)

And finally, a residual analysis:



Lastly the multiple linear regression that we obtained,

carbs=3.6373+7.1983xfat+4.7404xprotein+0.0807xsodium

can explain 99.1% of the observed sales variance (R-squared: 0.992, Adj. R-squared: 0.991). The F test is significant (p -value: 0.001). Therefore, there is clear evidence that the variance explained by the model is greater than that of random chance.

However, the test mean square error is higher that the 10% of the mean response value which means that linear regression is not the model that best fits the data for our response value. And more analysis need to be conducted.

Phase 5. Final reflection

To conclude, data science is important because it allows for storytelling and “magic”. What I mean by this is that the analysis of data uncovers the history and its origins, reasons and motives, and it can even tell us what’s coming in the future and predict what something will be like ahead of time. And personally, I believe that’s magic and one of the reasons many people enjoy data science. Also, the ethics concerning data science are of great importance precisely because it can be used can cause bias harming fairness and equity in results, privacy and security concerns could be collected or sensitive information might be handled, and transparency and explicability because sometimes algorithms can be hard to interpret. Moreover, data science can enhance performance, solve complex problems, innovate and revolutionize current and future products, and detect fraud, malfunctions, or failure of a system. Overall, data science is essential because it empowers organizations from business performance to healthcare outcomes.