Model summary

1. Background

Competition: Predict Future Sales

Name: Christian Urrego

Location: Bogotá, Colombia.

Email: christianurrego3@gmail.com

I am an electronic engineer but I work as a programmer in a software development company.

I am starting in the field of data science; I use weekends and holidays to study and improve my knowledge of all related topics.

This is one of the first projects that I carry out in this field.

2. Summary

I spend 80% of the time interpreting and understanding the data as it is the most important part, the number and quality of the characteristics that are identified or generated become very relevant when training the model. The most important features were related to different mean encodings for categorical variables. For the modelling I try first with a simple method like Linear Regression but the results but the results were not good, then I try with a XGB Regressor to get better results.

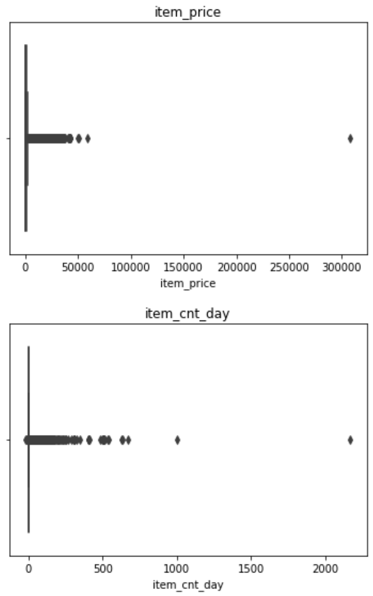

Figure 1 Outliers

Tools used to the project:

- numpy 1.19.0
- pandas 0.23.0
- seaborn 0.8.1
- sklearn 0.23.1
- xgboost 1.1.1

3. Feature selection / extraction

It was necessary to remove some outliers and fix some values that did not match with data specifically in prices and in the quantity per days. Data visualization was crucial to identify these outliers.

A matrix was created with all the possible combinations item-shop for each month and the dataframe began to be built with information from the other datasets .

The main features were generated with mean encoding of categorical variables like the category, sub-category, item. Also were created some others like the month and the city.

4. Models and training

In a first attempt a linear regression is used but the results are not good enough and a 5/10 score is obtained. Looking for a better method, XGB Regressor was tried. the predictions obtained a 10/10 grade.

All the process was made in a Laptop Acer A315-53G with an intel core I5-8250U 1.6 Ghz and DDR4 8 Gb RAM. I tried with some free instances in the IBM cloud, but I had several operating problems with the kernel and I could not use it, I configured the whole environment on my computer and with all the resources to the maximum it was executed successfully.

The model training took 5h and 34m, perform 394 iterations to find the better relation between the training-rmse and validation-rmse (0.92511,0.87850).

5. Conclusion

Is very important work with the data first, it was difficult to my because I´m a beginner but is evident when you try to training a model even when is simple like the linear regression that the result is not good enough. You can find important features in the data like the city in the shop names. When graphing the features importance, it is seen that one of those generated with the mean encoding has a great weight when making predictions, unlike the others.

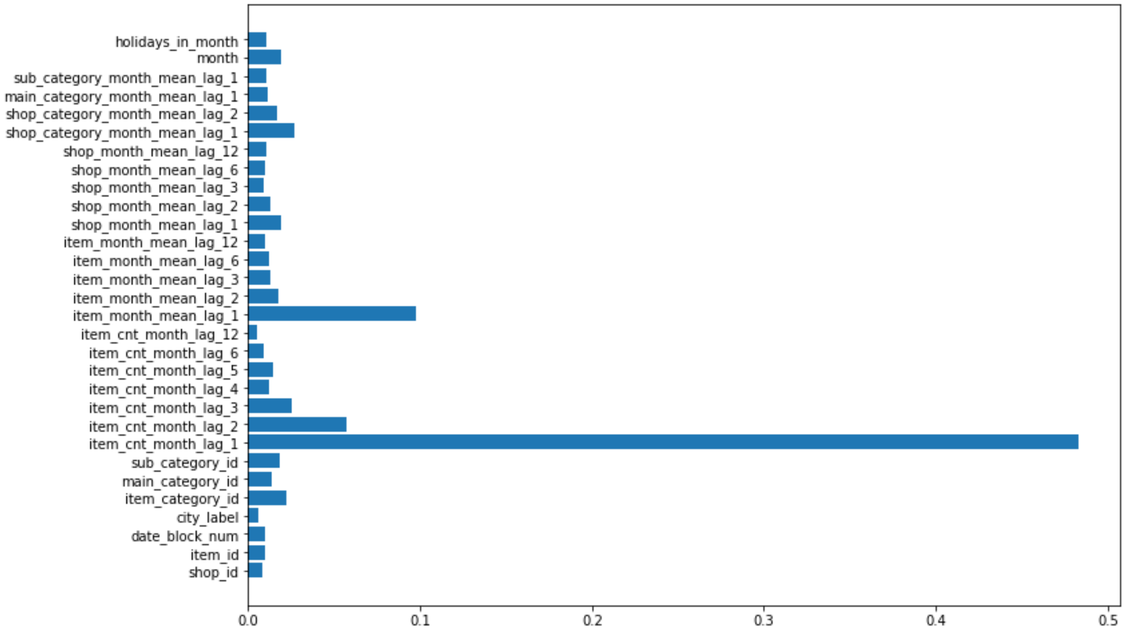

Figure 2 Features importance

Have a good hardware resources could help you to improve your work, in my case took a long time to training the model , it consumed and this consumed a lot of my available time.

I would like try more approaches to study the data and other models to compare the results. I think these exercises are very useful to learn and practice many skills in data science.

References

How to win Kaggle Competitions – Coursera

scikit-learn Documentation

Kaggle , Predict future sales competition notebooks .