**Submission Model Documentation**

Name: Pablo Vargas Ibarra

Location: Madrid

Email: [pablovargasibarra12@gmail.com](mailto:pablovargasibarra12@gmail.com)

Competition: Predict Future Sales

1. **Summary**

I have spent my time in 70% feature engineering / 30% machine learning. The most important insight was to understand the data leakage. It was related to the items in the test set. I had to generate the feature matrix and the cross validation set accordingly. The most important features were related to the number of sales in recent months and different mean encodings for categorical variables. I have used Python, the ensemble method Random Forest from scikit-learn and the library light gradient boosting.

1. Feature Selection / Extraction

The technique used was the usual in these cases, “error trial”. The most important features were related to the number of sales in recent months and different mean encodings for categorical variables. The most important aspect was understand the difference between test and training distribution in order to create good features.

1. Training Methods.

The models we have trained are two ensembles, Random Forest and Gradient Boosting. We have used the predictions of these models as meta-features for stacking. The meta-model was a simple linear regression.

1. Background

I am working as consultancy/data scientist in Madrid. My background is in mathematical engineering with focus on computational statistic.

**Appendix**

**A1. Dependencies**

numpy 1.14.2

pandas 0.22.0

sklearn 0.19.1

scipy 0.19.1

lightgbm 2.1.0

seaborn 0.8.0

**A2. References**

How to win Kaggle Competitions - Coursera