**Submission Model Documentation**

Name: Siddhant Prashant Kulkarni

Location: Pune, India

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

Competition: Predict Future Sales

1. **Summary**

I have spent most of my time in feature engineering and small amount of time in 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. I have used Python, the ensemble method I used is simple averaging of the two models and it worked out best for me.

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 a linear regressor and Gradient Boosting. We have used the predictions of these models for simple averaging to get the submission.

1. Background

I am an electronics engineering student with an avid interest in data science.

**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