Model Documentation

Background

- **Competition Name:** Sales Prediction
- **Private Leaderboard Score:** 0.93
- **Name:** Vincenzo Palermo
- **Location:** Italy
- Master in Computer Science
- I have worked on little projects of data science
- I spent about 7 days in this competition

Summary

- I had used GBDT with LightGbm since it's more fast than Xgboost and Sklearn. I also noticed that GBDT are really fast to implement, with good results and a little hyper parameter tune.
- Libraries: Pandas (to manage the data frames), Numpy (to manipulate vectors), Matplotlib, Sklearn (for label encoding and mean square metrics), Pickle (to dump the models and processed sets).
- Tools: Python, Jupyter notebook and Virtual Env.

Features Selection and Engineering

- Due to a low computational resources, i had use a manual search for the hyperparameters
- The best features are lagged month intervals with 1 month and items
- I had analyzed features impact with the tool "feature importance" of LightGBM

Training

- I had used LightGBM
- I initially tried with a linear combination of Random Forest, GBDT, Linear Regression and KNN, however i noticed how was more better use only GBDT.

Interesting Findings

- The best trick i had used is lagged mean, since we had temporal datas and also an upper trend of selles.

Model Exécution Time

- Training time: 10 minutes
- Prediction time: 10 seconds

Dependencies

- Programming language: Python 3.7
- Libraries:
  - Pandas 0.25.3
  - Numpy 1.18.1
  - LightGbm 2.3.1
  - Sklearn 0.22.1
  - Matplotlib 3.1.2
  - Pickle 0.7.5
- Operating system: Windows 10 1904