Model Documentation

Name: Mervyn Lee

Location: Malaysia

Email: leewd1994@live.com

Competition: Final Project - Sales Prediction

1. Background

- Bachelor in Computer Science
- I attended KKBox Churn Prediction and get top 9% ranking. I use xgboost in that competition, allows me to reuse the xgboost easily.
- I spent about 5 days in this competition.

2. Summary

4-6 sentences summarizing the most important aspects of your model and analysis, such as:

- The training method I used is XGBoost
- The most important features in my opinion is the lagged month intervals
- Pandas, numpy, sklearn, xgboost library, matplotlib and jupyter notebook
- 30 minutes

3. Features Selection / Engineering

- Most important features are the lagged mean encoded values from the categorical data.
- I select features with XGBoost plot on feature importance.
- I did not use any external data.

List of the 5 most important features: item_cnt_day_lag_1, item_category_id, shop_id, item_id_sum_item_cnt_day_lag_1, item_id

4. Training Method(s)

- I use XGboost
- I tried on stacking with Random Forest, Extra Tree and Linear Regression as first level model and use XGBoost as second level model without fine tuning and it give terrible result. Due to time constraint, we decide not to use ensemble here.

5. Interesting findings

- What was the most important trick you used?
- What do you think set you apart from others in the competition?
- Did you find any interesting relationships in the data that don't fit in the sections above?

Appendix

This section is for a technical audience who are trying to run your solution. Please make sure your code is well commented.

A1. Model Execution Time

Many customers care about how long the winning models take to train and generate predictions:

- What software did you use for training and prediction?
    - Linux terminal
- What hardware (CPUS spec, number of CPU cores, memory)?
    - i7-7700K, 4.20 Ghz x 8 cores, 64gb memory (Limited time usage)
    - I7-7700, 3.60 Ghz x 8 cores, 16gb memory (Unlimited time usage)
- How long does it take to train your model?
    - 30 minutes
- How long does it take to generate predictions using your model?
    - Less than 1 minutes

A2. Dependencies

List of all dependencies including:

- programming language/statistical tool
    - Python
- libraries or packages
    - Pandas: 0.20.3
    - Sklearn: 0.19.1

    - Xgboost: 0.6a2
    - Matplotlib: 2.1.0
    - Numpy: 1.13.3
- operating system
    - Ubuntu 16.04

A3. How To Generate the Solution (aka README file)

Provide step-by-step instructions for how to create the solutions file from the code provided. Include that description here and in a separate README file to accompany the code.

1. Run data_processing.ipynb in Jupyter Notebook
2. Change lag features month to [1, 2, 3, 4, 5, 6, 9, 12] to compute full features if enough memory
3. Run training_xgboost.ipynb for prediction. Comment away training code to skip training and use pre trained model from xgb.pickle.dat