Documentation

How to Win Data Science Competitions

Competition: Final Project Predict Future Sales

Author: Cesar Gustavo Seminario Calle (Systems Engineer)

Universidad Nacional Tecnológica de Lima Sur

Lima - Perú

Email: gustavosystemas@gmail.com

Summary

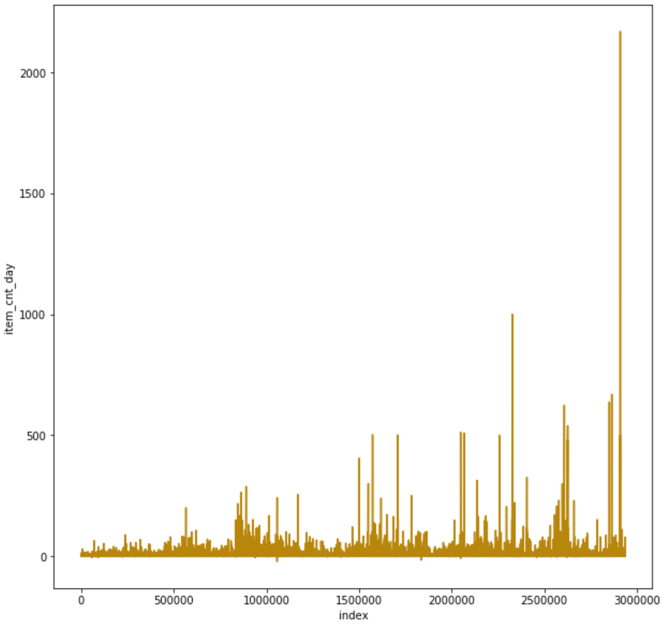
Exploratory Data Analysis

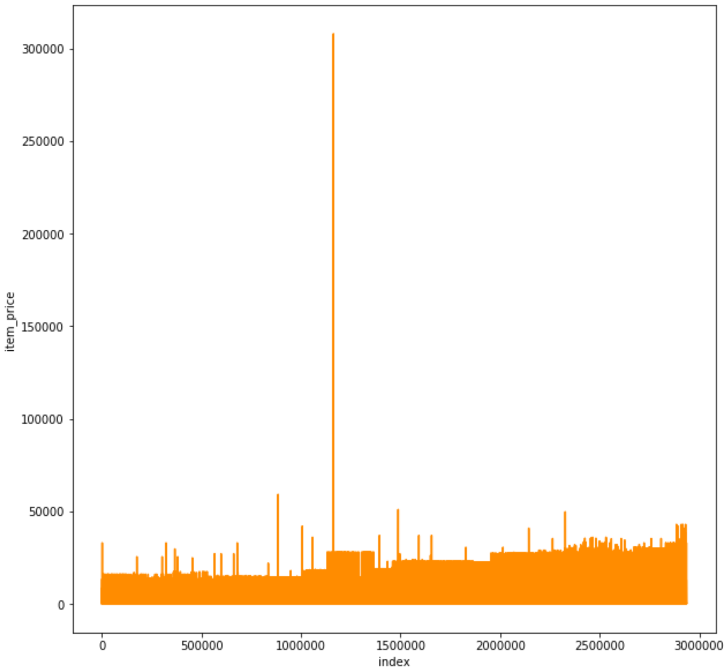
I started checking the distribution of the numeric features like item_cnt_day.

I check how the train and test were splitted so I found out that in test data they have make throrought multiplication between a set of shop_id against a set of item_id, It means that it migh be that those combinations don't even exist in training.

I also added new variables related to date time (month, year) just to plot and see how other features behave

Look for outliers in the numeric features, I found two possible outliers 1 in item_cnt_day and the other in item_price which were extremely high in comparison to the other values so I thought it would be worth trying exclude these values from the analysis.





Preprocessing

Then I aggregate the data monthly by using shop_id and item_id as my keys for grouping.

I translated category_name and item_name to English so I could extract another features.

Feature Engineering

I extract some labels from those features as is _audiobook, is_dvd, categories (I named it cats).

By this moment I was thinking about target encoding that would be udeful due to the fact that I could summary a set of months in a single period.

I try target encoding I use item_price and item_cnt_day to extract some measures like average, sum from the other categories in diferent months so I have to stack those new features.

Note:

This idea of using target encoding requires that your computer have at least **60 Gb ram**, what I did was use Google Cloud computer engine which give you free credits for a year

Another option will be using **Dask**.

Validation:

Hold Out validation was my choice because Kfold or LOO would take too much time due to memory consume.

Hold out with time window:

- Training Months From January 2013 to September 2015
- Validation Dataset period October 2015

I assumed that the test data were items from November 2015.

I try using Random Forest as my first option but it often run out of memory,

I try XGBoost and LightGBM

Final Solution

LightGBM give such good results on my solution.

The parameters that I foundnd useful to play with were:

- Early_stopping
- Feature_fraction
- Num_leaves

I try to use hyperopt to optimize my parameters but it takes too many hours that I decided to do it manually.