Solution Documentation

Background

Competition: Predict Future Sales
Team: none, just me
Name: Manuel Alejandro Martínez Flores
Location: Guatemala
E-mail: manuelalejandromartinezf@gmail.com
Public leaderboard: 0.997806
Private leaderboard: 0.999039

I am an Applied Mathematics student and enthusiast about ML and DL. I am enrolled at 'Advanced Machine Learning' specialization that includes the curse 'How to win a data science competition: learn from top Kagglers'.

I spent around 18 hours working on this solution.

Model Summary

The features generated were sales lags from each shop-item combination, features extracted from categories names and mean encodings.

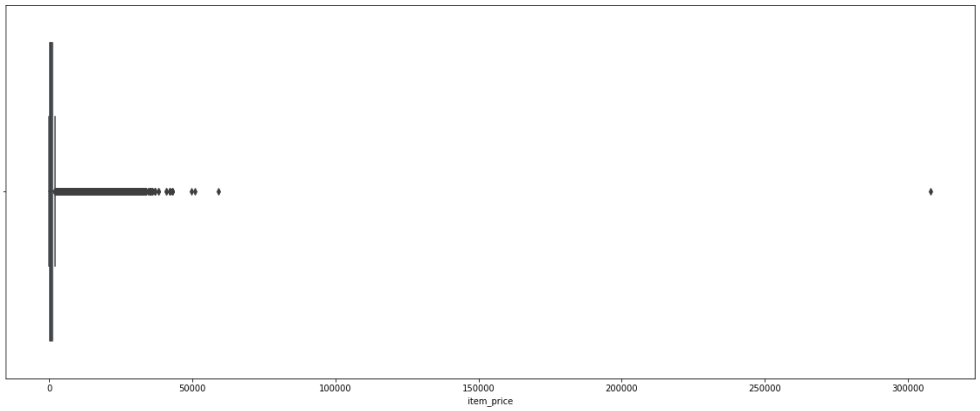
The model consisted of a stacking of two levels but a CatBoost regressor gives a similar result by itself.

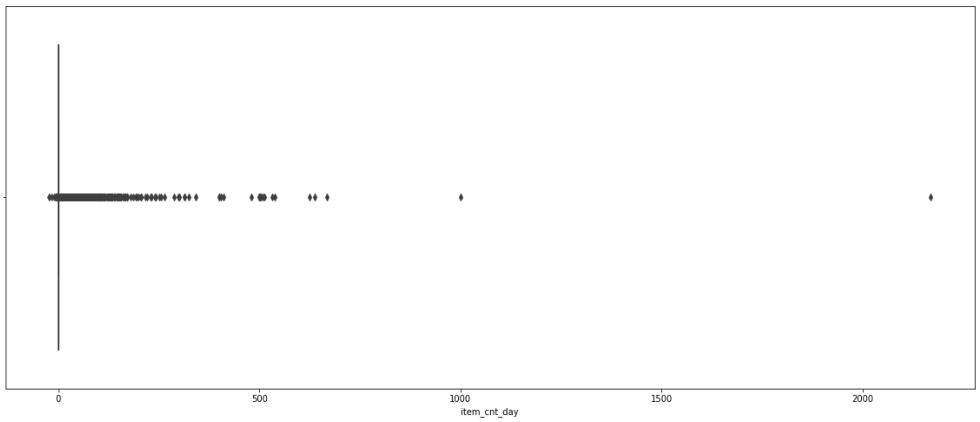
The main used tools were:

- numpy 1.1.5
- pandas 1.9.5
- sklearn 0.22.2
- catboost 0.24.4

Feature Selection/Engineering

An initial exploration (distributions, correlation, etc.) showed that there were some intense outliers in the 'item_price' and 'item_cnt_day'. These were removed from the dataset





In feature engineering, a pivot table from pandas was used to create the lags from the previous sales from each item-shop combination. Additionally, the 'item_category_id' categorical field was mapped from one of the datasets provided.

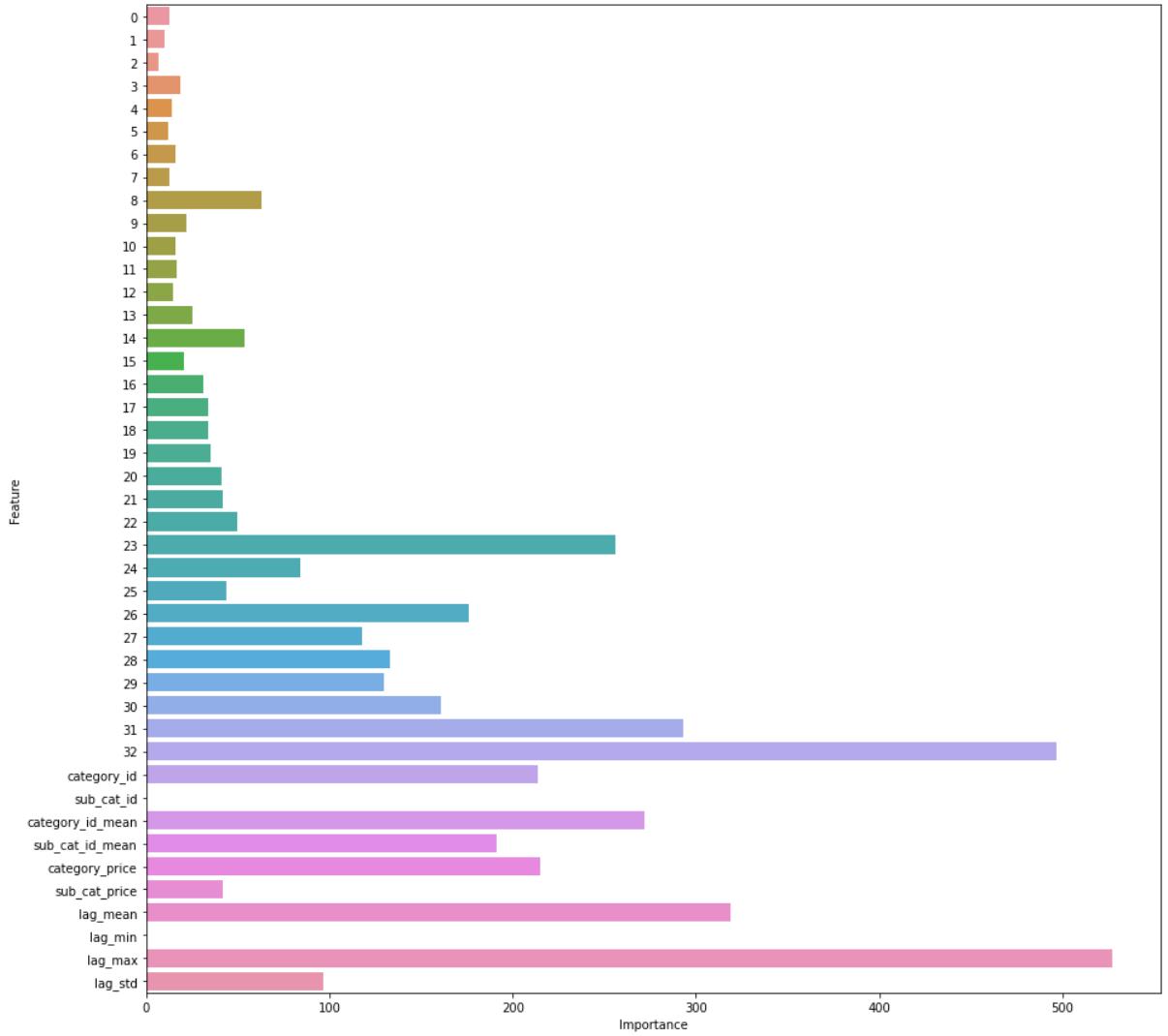
After exploring the category dataset, the names showed that each 'item_category_id' class belonged to a bigger category cluster. After some text mining, these new categories were also mapped to the main dataset. These new categorical features seemed to get the best out of the CatBoost model.

The same idea was implemented with the shop dataset, however, these categories seemed to add no new information after feature selection. Then resulting in dropping the 'shop_id' and 'item_id' because they seemed to affect the model negatively.

Therefore, mean encoding was implemented using the categorical features, the expanding mean method was used to avoid overfitting. A price mean by both categories was computed under the motivation of price-demand relationship, and it turned out to add new information. Similarly, statistical features (mean, min, max, std) from the lag distribution were computed and added significant information.

Feature importance

Using a preliminary LGBMRegressor, the feature importance was computed. The result shows that the most significant features are the max of all the lags and the lag from the past month. Mean encodings and other statistical features resulted to perform well too.



Modeling and training

Stacking was implemented in two levels using a cross-validation scheme. The first layer included: SVM, CatBoost Regressor, LGBM Classifier. The second layer consisted of a Ridge Regression after some meta-feature manipulation.

However, stacking did not have a significantly better performance compared to the CatBoost Regressor alone. The CatBoost Regressor was trained

using early stopping and using cross-validation. This allowed the model to use most of the training data without overfitting.

Interesting Findings

The CatBoost alone was performing as well as the ensemble. The categorical relations this model finds are really interesting

Model execution time

From data reading to model training to target prediction, it takes around 20 minutes to complete. This time was recorded using a Google Colab environment with 25GB RAM.

References

Coursera www.coursera.org
Kaggle www.kaggle.com
CatBoost https://catboost.ai/docs

Future work

Try a more diverse ensemble.
Try a different cross-validation strategy
Implement advanced text-mining features.