Flip oo project final report

Rongxin Xu

Hunan University, China

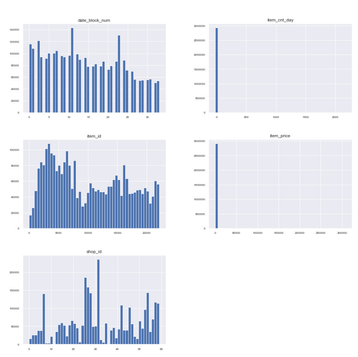
Introduction

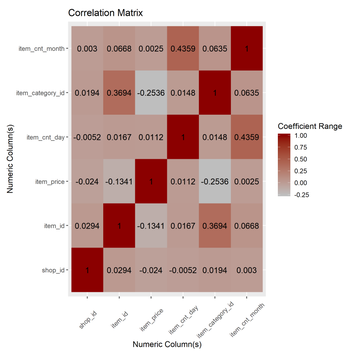
This is a problem with time-series prediction. After a month of making scientific observations and taking careful measurements, can predict total sales for every product and store in the next month. The raw dataset contains train set with 2935849 samples and 214200 unlabeled samples as test set. Through the train data, predict total sales for every product and store in the next month.

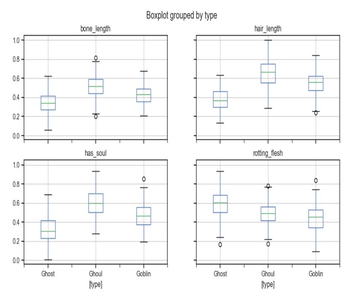
id an Id that represents a (Shop, Item) tuple within the test set.
shop_id unique identifier of a shop.
item_id unique identifier of a product.
item_category_id unique identifier of item category.
item_cnt_day percentage of soul in the creature.
item_price current price of an item.
date date in format dd/mm/yyyy.
date_block_num unique identifier of item category.
item_name name of item.
shop_name name of shop.
item_category_name name of item category.

Data Visualization

The Histogram shows the distribution of the various attributes. It seems that item_id and shop_id has a huge impact on sales and sales tend to decline with the date. Through the Boxplot, we know that the outliers are very small, so can be ignored.







Feature Engineering

Using the Feature Importance this function of Random Forest to select the most important features to form a new train data. I take these features to form a new train datad.

Algorithm

To combine the base models as 1st level model predictions, I'll use a simple linear regression. As I'm only feeding the model with predictions I don't need a complex model.

- Base Models
  - RandomForest
  - XGBoost
  - LSTM
  - Linear regression
  - KNN
- Ensemble Model
  - Linear regression

Experiment Result

The tables below shows that the rmse of each model, RandomForest and XGBoost perform better, and there is not much difference between the other models.

- Forecast Result of Base Models

	RandomForest	XGBoost	LSTM	Linear regression	KNN
Train rmse	0.8358	0.8327	0.9276	0.8572	0.6976
Validation rmse	0.8810	0.8959	0.6611	0.8806	0.8946

The train rmse of Ensemble model is 0.764973649571408.

Conclusion

Exploratory data analysis is very important for the competition, Discover the imperfections of the data and have a certain understanding of the overall appearance of the data, which will help later modeling and analysis.

The data that we have, needed processed in many cases. Data preprocessing includes deal with missing data and outliers, We must think carefully about the outliers, such as ignoring them.

The most important thing is feature engineering. We have to think carefully and deal with outliers, such as ignoring or deleting them.

There is no best model, only the best model. We should try as many models as possible to get the best prediction results.

Feature engineering is very important and even plays a decisive role in this competition.

The Ensemble model may perform better than a single model in some complex problems.

Acknowledgement

- Thanks!