Machine Learning Approaches to "Predict Future Sales" Kaggle Competition

Jocelyn McConnon

April 8th, 2020

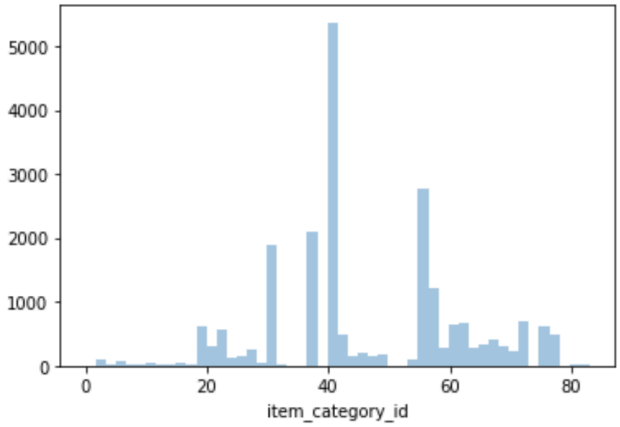
Intro to Machine Learning Applications

Lally School of Management
Rensselaer Polytechnic Institute
Troy, Ny 12180

The following report will discuss the Kaggle prediction competition “Predict Future Sales”. This competition uses sales data from a Russian software firm named 1C Company. Key features of the data include when, where, and how many of their products are purchased. The objective is to predict the number of each product which will be sold by each store for the upcoming month. Submissions to this contest are scored using the root-mean-squared error. RMSE is computed by taking the square root of the average of the squared errors. Three solutions will be discussed in this report. The modeling approach as well as method of feature selection will be explained, and leaderboard scores are taken into consideration.

The dataset consists of 6 CSV files. All the files that compose the data contain no missing or repeating information. Therefore, there is no need to clean the data at this point. One file contains the 84 categories of items. The category name is given along with a unique ID number. It is worth noting here that the categories names are in Russian. One of the solutions discussed in this report translated the category names to English, and created a supercategory feature. Figure 1 contains a histogram which describes the distribution of items among the categories. The category named “Кино - DVD” which translates to English as “Movies - DVD” (ID number 40) contains close to 5,500 items. 1C Company has 22,170 items for sale and thus roughly 25% of their inventory is in DVD Movies.

Figure 1: Histogram of Number of Items vs. Item Category ID



Another file for the data used in this competition contains the item name, item ID number, and item category ID. There are 22,170 items, and like the category names, the item names are given in Russian. An example of an item name translate to English would be “100 Best romantic melodies (mp3-CD) (Digipack)”.

There is a file which contains the 60 shop names (In Russian), and corresponding shop ID number. The shop name inherently contains the city in which the shop is located. For example, “Krasnoyarsk TC "Vzletka Plaza”” is one of the shop names given (Krasnoyarsk is a city in Siberia, and Vzletka Plaza is where the shop is located). Figure 2 shows the distribution of sales amongst 1C Company’s 60 shops. Their Vzletka Plaza location, for example, is shop ID number 17 and experiences a moderate volume of sales.

Figure 2 is made using the sales data contained in the test file. The testing data file contains 6 columns and 10,34 rows. The features are date, consecutive month number of the date, shop ID, item ID, item price, and number of products sold. The number of products sold is the predicted value. In other words, given the selected features, the objective is to predict the number of each item sold at each sore location in the upcoming month. Thus, the testing file contains three columns. One column

Figure 2: Histogram of Number of Sales vs. Shop ID

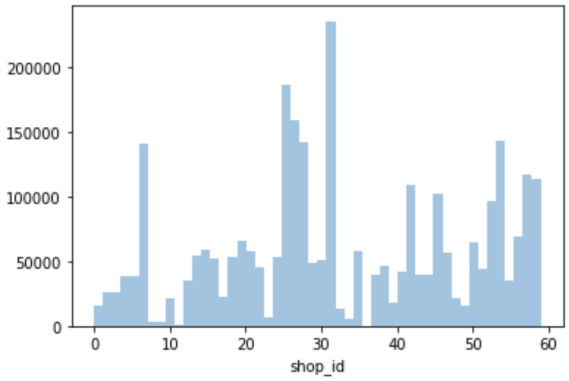
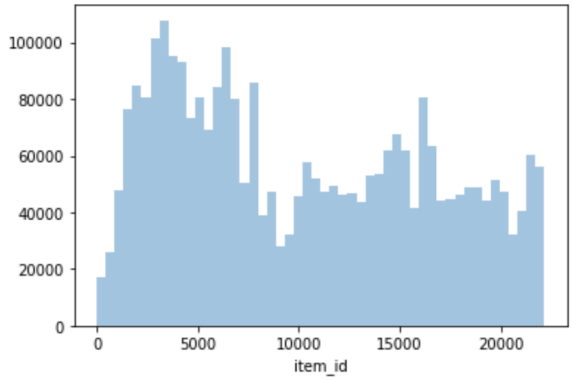


Figure 3: Histogram of Number of Sales vs. Item ID



contains shop IDs, another contains item IDs, and the last column contains a ID for this tuple.

In Figure 3, we can observe the distribution of sales among 22170 items. The range is relatively tight, meaning that 1C Company's products perform similarly. There are some products that sell better than others, but there are no products which do not sell at all.

The first solution considered is named Feature Engineering/LightGBM/Exploring Performance and was submitted by Luke M. Specifically, version 135 will be discussed in the following section. This submission has a score of 0.85389, which is the lowest of the considered solutions and thus the most accurate method discussed in this report. The approach to building this model was predicated upon translating the Russian category, item, and shop names into English. The English translation of the data was one of the three inputs for this submission. Using the translated data, a supercategory of category names was added as a feature. For example,"Accessories" would become the supercategory of "Accessories - PS2". As part of the data preprocessing, outliers were removed from the training sales data. Shop IDs that do not appear in the testing set were also removed from the training data. Features such as day, month, and year were added using the date given in the training data. The training data was then grouped by month, shop ID, and item ID. K-means clustering was then performed. The best silhouette score was found using k = 7, and so shops were clustered into 7 clusters using their category sales as features. A feature was added representing the amount of time elapsed from the start of each month until an item was sold. This will be used in the model to determine the past performance of items. Other features added include (but are not limited to) sales of items by shop, sales within categories by shop, mean monthly sales by shop and by item, average price of items and categories. Since so many features are added, the first two month blocks were discounted from the training data for the reason that the author felt some metrics may not be representative. This solution was unique in that it improved upon an initial model which was included in the inputs. A Light Gradient Boosting (LGB) model was used with a learning rate of 0.01. The training RMSE was 0.62278, and the testing RSME was 0.740195.

The next solution to be considered is named “Feature engineering, xgboost” and was submitted by Denis Larionov. Specifically, discussed in this report is the 2658 version of the notebook, which has obtained a leaderboard score of 0.90684. To prepare the data, items outing prices and low sales were removed from the training data. Additional features were added by exploiting the naming structure of the shops and categories. Specifically, the city contained in the name of the shop was added as a feature. Like in the first solution we examined, the supercategory of categories was also added as a feature. MeanEncoding was used on the features to convert the names into numerical values. Trend features were added to describe the sales of items for the past 6 months, as well as the performance of shops in terns of sales. All features, including the newly created ones, were used. A XGBoost model was used, and produced a training RMSE of 0.813137 and a testing RMSE of 0.905782.

Figure 4: Performance summary

	Solution 1 Feature Engineering/ LightGBM/Exploring Performance	**Solution 2** Feature engineering, xgboost	**Solution 3** A beginner guide for sale data prediction
Score	0.85389	0.90684	1.25011
Feature selection	K-means clustering	Existing features were supplemented.	Existing feature were supplemented.
Model	LGB	XGBRegressor	LSTM
Runtime	6447.6 seconds	3976 seconds	3684.5 seconds
Training accuracy	0.622781	0.813137	3.85
Testing accuracy	0.740195	0.905782	4.29

The third solution I will consider is named “A beginner guide for sale data prediction” and was submitted by Triet Chau. Version 1171 will be referenced in this report, which had a Kagle performance score of 1.25011. This score is the highest of the submissions we have discussed, meaning this method is the lease accurate. This method uses Long Short-Term Memory architecture. First, this method checks to see which shops are in the trainings and testing data. The model accounts for the spike in sales in the month of December, due to holiday shopping. The data is normalized, which is unique to this method. In fact, after normalizing the data there are several items for which the price is negative. The author uses LabelEncoder to transform the shop names into numerical values which are then vectorized. The model is built using Sequential, a single LSTM. After the model is evaluated, the training scores is 3.85 RMSE and the test score is 4.29 RMSE. These are ridiculous numbers and show that the model is very inaccurate.

Reference List

"Predict Future Sales" Kaggle Competition. https://www.kaggle.com/c/competitive-data-science-predict-future-sales/overview

"Feature Engineering/LightGBM/Exploring Performance". Luke M. Version 135. https://www.kaggle.com/deepdivelm/feature-engineering-lightgbm-exploring-performance

"Feature engineering, xgboost". Denis Larionov. Version 2658.https://www.kaggle.com/dlarionov/feature-engineering-xgboost

"A beginner guide for sale data prediction". Triet Chau. Version 1171. https://www.kaggle.com/minhtriet/a-beginner-guide-for-sale-data-prediction