Amit Levizky - 203911813
Alex Abramov - 314129438

Predict Future Sales - Report

The task we investigated is predicting the number of units that will be sold in stores in the next month. Like every ML/DL task, the first thing we would like to do is to preprocess our data if he is not already ready to use. These steps were critical to our model in terms of both the model's complexity and its correctness. The processes we performed are:

1. discard all rows that have value (-1) in the '*item_cnt_day*' column.
2. clear all rows in the training dataset that contains shops and items that not in test-set
3. Fixing the date representation, so later on we could use him as features.
4. Fixing the test-set by adding the month we want to predict.
5. Grouping the records so that the units counter is by months and not by days.
6. Discard values that are far less common and if so can significantly impact the results of the model.
7. Dividing the training set to 80% training and 20% validation.

The next thing we were asked to do is do the training and prediction on the data set and test using a classic ML algorithm. The results obtained were not bad, but there seemed to be room for improvement.

At this point, we started to build an RNN model when we want to extract features from the data. We did this with Embedding. First we only used 3 features and we built a neuron network for the data in the training set. Already at this stage, when we only run the model with 3 features we got a better result than the previous step (ML).

Now we have decided to add another feature that we think will affect the result and this is the year feature. Once again it was noticeable that there was a positive change in the prediction results.

The final step in the task was to perform a feature-extractor process from the models we trained in the previous step. To complete the task:

- We pulled the last layer from both models.
- We created a new ML model. (We chose to use logistic regression).
- We performed prediction with the original models (without the last layer) for the test-set.
- The output from the previous step was entered as an input to the ML model.

Result: Surprisingly, the results we received were less accurate than the results we described in the previous steps.

The reasons that caused the surprising results in our opinion are:

1. Wrong choice of ML algorithm for the task that you are trying to perform.
2. The Feature Extractor process was performed incorrectly.

To summarize, The key points in our opinion are:

Amit Levizky - 203911813
Alex Abramov - 314129438

1. A thorough preprocess that not ignores important details and yet knows how to deal with less common or unimportant details.
2. Choosing features that will improve prediction based on the research done on data in the early stages. For example, we could see from the tables we attached above that in some months sales dropped and others went up. Therefore, choosing a month as a feature made the model prediction more accurate.