WNS Analytics Wizard 2019 Report

Roman Piankov (TheRealRoman)

3 September 2019

1 Brief description

My solution consists mainly of feature engineering. I generated features based on user_id and app_code characteristics. This is described in more detail below. Light GBM trained on 10 different subsamples is used as a model. As a final prediction, averaging of these models by rank is used.

2 Validation scheme

For validation I used StratifiedKFold(from sklearn.model_selection) with following parameters:

- n_splits = 10;
- random_state = 228;
- shuffle = True.

3 Feature engineering

1. I encoded the string values of the 'os_version' columns to integer values(category number).

2. From the column 'impression_time' i made two new features:

- hour;
- minute.

3. For ‘user_id’ I calculated number of unique values from column ‘app_code’(‘cnt_unique_app’). Calculated the difference and ratio between number of unique values and number of all values for user_id(’gg_1_diff’, ’gg_1_ratio’).

4. For ‘app_code’ I calculated number of unique values from column ‘user_id’(‘app_code_cnt_unique_u Calculated the difference and ratio between number of unique values and number of all values for app_code_id(’gg_2_diff’, ’gg_2_ratio’).

5. For each ‘user_id’ I calculated mean, min, max difference between two consecutive ‘impression_time’(’diff_time_mean’, ’diff_time_max’, ’diff_time_min’).

6. For each user_id and impression_time I calculated the time from the current impression_time to the previous and next user action. Calculated average number of action, which user clicked for previous time(’diff_time_user_id_last’, ’diff_time_user_id_next’, ’value_mean_user_id’).

7. For each app_code and impression_time I calculated the time from the current impression_time to the previous and next app_code action. Calculated average number of action, which users with this app_code clicked for previous time(’diff_time_app_code_last’, ’diff_time_app_code_next’, ’value_mean_app_code’).

8. For each user_id and impression_time I calculated the number of action in view log for previous time(’value_cnt_view_user_id’).

9. For each user_id and impression_time I calculated mean, min, max difference between two consecutive ‘server_time’ in view log for previous time(’value_diff_time_view_user_id ’value_diff_time_view_user_id_max’, ’value_diff_time_view_user_id_min’).

10. For each user_id and impression_time I calculated number of unique values of the columns from view_log with item_data: [‘session_id’, ‘item_id’, ‘category_1’, ‘category_2’, ‘category_3’, ‘product_type’](’user_id_unique_session_id’, ’user_id_unique_item_id’, ’user_id_unique_category_1’, ’user_id_unique_category_2’, ’user_id_unique_category_3’, ’user_id_unique_product_type’).

11. For each user_id and impression_time I calculated mode of the values of the columns from view_log with item_data: [‘item_id’, ‘category_1’, ‘category_2’, ‘category_3’, ‘product_type’](’user_id_mode_item_id’, ’user_id_mode_category_1’, ’user_id_mode_category_2’, ’user_id_mode_category_3’, ’user_id_mode_product_type’).

12. for each col in [’user_id_mode_item_id’, ’user_id_mode_category_1’, ’user_id_mode_category_2’, ’user_id_mode_category_3’, ’user_id_mode_product_type’, ’app_code’, ’user_id’]

I calculated frequency response for each unique value('vc_app_code', 'vc_user_id', 'vc_user_id_mode_item_id', 'vc_user_id_mode_category_1', 'vc_user_id_mode_category_2', 'vc_user_id_mode_category_3', 'vc_user_id_mode_product_type')

4 Model

As a model I used LightGBM with following parameters:

- 'bagging_fraction': 0.8,
- 'bagging_freq': 1,
- 'boost': 'gbdt',
- 'feature_fraction': 0.8,
- 'learning_rate': 0.01,
- 'metric':'auc',
- 'num_leaves': 31,
- 'num_threads': 8,
- 'objective': 'binary'.

I used a StratifiedKFold with 10 folds, so I got 10 models that were used to predict the test data. Rank averaging among these 10 predictions was used as the final prediction. On local validation i got the following average ROC_AUC value: 0.7383677000.

5 things a participant must focus

- feature engineering - it's most important in this competion;
- building competent validation;
- competent work with categorical features;

- setting model parameters;
- overfit control.