CUHK Business

Predictive Analytics for a Large-Scale Food Delivery Platform

DSME 6756BA/B: Business Intelligence Technologies and Applications

Due at 23:59PM, on Thursday, February 10, 2022

Please complete the following task and submit (a) CSV files of your prediction results on Kaggle, and (b) a brief PDF report articulating your approach and results and Jupyter Notebook containing your analysis on Blackboard. We also design several sub-questions which may guide your analysis. This project counts towards **12%** of the final grade for this course. You are allowed to discuss with anyone about this project, but you should perform the analysis and write the report on your own. Please make the PDF report, without compromising on quality and clarity, as concise as possible.

Please first register an account at https://www.kaggle.com/ and join the competitions through https://www.kaggle.com/c/ba2021classification/ (the classification problem) and https://www.kaggle.com/c/ba2021regression (the regression problem).

Background

Within the context of Eleme delivery service, at every moment the command platform of "Smart Logistics" sends orders from customers to delivery-men for instant delivery. The decisions being made by the delivery-men are mainly two-fold: (i) pick-up from a store, and (ii) delivery to a customer. At any specific moment, a delivery-man may receive new orders assigned to him, while he has some unfinished orders which were previously assigned. In this situation, he needs to decide what to do next (to pick up a new order or to deliver an old one), based on his order status and geographical location.

There are two main tasks for this project, each counting 6% towards your final course grade. The first task is a classification problem, in which you are asked to build a model to predict whether the delivery-man's next move is to pick up an order or to deliver an order. The second task is a regression problem, in which you are asked to build a model to predict the expected time of the delivery-man's next action. For both tasks, you will use his historical decisions and current status as features.

Data

The data sets can be directly downloaded from Kaggle. There are 4 files for each competition.

The main datasets you need to work with are `dataframe_train.csv` and `dataframe_test.csv`. The difference between them is that for the test data set, `action_type` (for the classification problem) and the `expected_use_time` for the regression problem variable are masked, and you are asked to predict them for the orders in the test data.

The other 2 files are a sample submission file, the format of which you should follow when prepare your own submissions, and a Q&A from some other students not in this course who have used the original data set. To make your life easier, we have already pre-processed the data. Below is some brief overview of the variables, the rest can be explored on your own. You may also refer to the Q&A file on Kaggle for more information.

`courier_id`, `wave_index`, `tracking_id`, `date`, `group`, `id`:

These are the demographic information of the order and its courier. The column names are somewhat self-explanatory. In particular, a wave on Eleme means a batch of orders the platform processes together to assign the delivery men.

`courier_wave_start_lng`, `courier_wave_start_lat`:

These are the starting longitude and latitude of that wave of a certain courier.

`level`, `speed`, `max_load`:

These are the courier information: The level of the courier, the speed of the courier and the max load of the courier.

`weather_grade`:

This is the weather condition at the time of the order.

`aoi_id`:

The id of the Area of Interest (i.e. the delivery destination).

`shop_id`:

The id of the shop.

`source_type`, `source_tracking_id`, `source_lng`, `source_lat`:

The information of the courier's previous action.

`target_lng`, `target_lat`:

The geographical information (longitude and latitude) of the target.

`grid_distance`:

The shortest travel distance to the target provided by the GPS.

`hour`:

The hour in the day.

`urgency`

Identifies how urgent the order is

`action_type`

The type of the action, delivery or pick-up. This variable is masked in the test set for the classification problem.

`expected_use_time`:

This is the label of the prediction task (measured in seconds) and is, therefore, masked in the test set for the regression problem.

Questions

Please address the following questions for both the classification and the regression problems.

(a) (2 points) **Initial Data Pre-processing**. Prepare a new data set by selecting a subset of features that you feel relevant for your prediction task. Convert certain factor variables into numeric ones. Remove unreasonable data observations such as outliers. Down-sample a certain proportion of the data observations so that your computer could handle the subsequent training and testing procedures efficiently. It is suggested that you downsample to no more than 20,000 data observations first and use a larger set after you finalize your model. You may also do some exlporative data analysis such as computing summary statistics and conducting data visualization to guide you building the initial data sample.

(b) (2 points) **Baseline Model**. Create a simple model by doing an initial selection of features, doing appropriate pre-processing and cross-validating a **linear model**[1]. Report the F_1 score (classification) and the MAE (regression) of your baseline models.

(c) (2 points) **Any Model**. Try more complex models (e.g., k-NN, decision trees, random forests, gradient boosting trees, etc.) to strengthen your prediction. You may need to (and should) change your pre-processing and feature engineering to be suitable for the model. You are NOT required to try all of these models. Tune the parameters as appropriate.

(d) (2 points) **Further Feature Engineering**. Introduce new features through, e.g., re-scaling, polynomial features, clustering etc. Think about the business logic behind your feature engineering procedures. For example, it may not be necessary to cluster the data observations using all features, but the geographical information will suffice. Identify useful features from the original features and those created through feature engineering that are important for

[1] By a linear model, we mean a linear regression model or a logistic regression model.

your best model. Re-build and re-tune your model with the selected features to improve your prediction.

(e) (4 points) **Result Submission and Model Interpretation**. Based on the best model you build, make predictions on the test set and submit your results to Kaggle. The submission to the classification problem should be in 0-1 format, with 1 representing the "DELIVERY" action. Based on your model and result, discuss, if any, actionable business insights you can recommend to the Eleme platform. For example, which feature(s) do you think is/are most relevant for predicting the type and time of the next action for the delivery man? In order to receive a full credit in this question, the prediction accuracy for the classification problem (the F_1 score) should at least 0.88 on the private leader board, whereas the prediction error for the regression problem (MAE) on the private leader board must be no greater than 190. Please note that the private leader board will be accessible only after the competition ends.

Hints

Below we provide some hints for this project:

1. See https://en.wikipedia.org/wiki/F-score for the definition of the F_1 score; and https://en.wikipedia.org/wiki/Mean_absolute_error for the definition of the MAE metric.

2. You should submit your results to both the classification competition (https://www.kaggle.com/c/ba2021classification/) and the regression competition (https://www.kaggle.com/c/ba2021regression), respectively. Please check with the sample submissions provided on Kaggle to make sure your submissions are in the correct format.

3. The MAXIMUM number of submissions for each competition in a single day is **3**.