NYU Shanghai - NYU Stern

Final Project

SHBI-GB 7311 B1: Machine Learning for Business (Summer 2021)

Due at 1:00PM, on Saturday, August 21, 2021

The final project will count 20% (+2% extra credits) towards your final grade of this course. You will finish 2 tasks, one on prediction and the other on causal inference. Each task will have several sub-questions. You can code in ANY programming language that you feel comfortable with. You are allowed to discuss with anyone about this project (including myself), 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.

1 Prediction Task

For this task, please submit to Brightspace (i) a brief PDF report articulating your approach and results for each sub-question; and (ii) the code containing your analysis. Submitting *R* markdown or Jupyter Notebook is also acceptable. You should also submit a CSV file of your prediction results to Kaggle.

Please first register an account at https://www.kaggle.com/ and join the competition through https://www.kaggle.com/c/machinelearningforbusiness2021.

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.

You job in this task is to build machine learning models to predict the expected time of the delivery-man's next action (pick-up or delivery). The features you may use are the historical decisions and current status of the delivery-man.

Data

The data set can be directly downloaded from the Kaggle website. There are 4 files thereof.

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, the `expected_use_time` variable is

masked, and you are asked to predict it 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 traversable 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.

`expected_use_time`:

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

Questions

Please address the following questions.

(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) (3 points) **Baseline Model**. Create a simple model by doing an initial selection of features, doing appropriate pre-processing and cross-validating a linear model. Report the MAE of your baseline model.

(c) (3 points) **Any Model**. Try more complex models (k-NN, CART, RF, XGBT, 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) (3 points) **Further Feature Engineering**. Introduce new features through, e.g., re-scaling, polynomial features, clustering etc. Think about the business logic behind your 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 your

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

(e) (3 points) **Result Submission and Model Interpretation**. Based on the best model you build, make predictions on the test set and submit your result to Kaggle. 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 time of the next action for the delivery man? In order to receive a full credit in this question, the prediction error (MAE) on the private leader board (which will be accessible only after the competition ends) must be no greater than 190.

2 Causal Inference Task

For this task, please submit to Brightspace (i) a brief PDF report articulating your approach and results for each sub-question; and (ii) the code containing your analysis. Submitting R markdown or Jupyter Notebook is also acceptable.

The second task of the Final Project aims to provide you an opportunity to apply causal inference techniques to real experimental data.

Background

In 2008, a group of uninsured low-income adults in Oregon was selected by lottery to be given the chance to apply for Medicaid. This lottery provides a unique opportunity to gauge the effects of expanding access to public health insurance on the health care use, financial strain, and health of low-income adults using a randomized controlled design. The Oregon Health Insurance Experiment followed and compared those selected in the lottery (treatments) with those not selected (controls). You may visit this website https://www.nber.org/programs-projects/projects-and-centers/oregon-health-insurance-experiment?page=1&perPage=50 for more information about this experiment and the subsequent research and public policy based on this experiment.

You job in this task is to estimate the causal effect of being selected by the lottery and enrolling into the medicaid program on emergency department utilization.

Data

Please download the datasets and the relevant documentations from GitHub and Brightspace. The datasets are stored in `.dta` format (the data format of Stata), which can be read into R using function `read.dta()` in the `foreign` package. A sample R code to load `.dta` data is also provided. If you choose to work with Python, you can load `.dta` data as a Pandas data frame using the function `pd.read_stata("oregonhie_descriptive_vars.dta")`.

Randomization and Treatment Assignment. Oregon selected roughly 30,000 individuals by lottery from a waiting list of about 90,000 for an otherwise closed Medicaid program. The state conducted eight lottery drawings from March through September 2008. Selected individuals won

the opportunity – for themselves and any household member – to APPLY for health insurance benefits through a Medicaid program called Oregon Health Plan Standard (OHP Standard). OHP Standard provides benefits to low-income adults who are not categorically eligible for Oregon's traditional Medicaid program (OHP Plus); to be eligible individuals must be adults ages 19 – 64, not otherwise eligible for Medicaid or other public insurance, Oregon residents, U.S. citizens or legal immigrants, have been without health insurance for six months, have income below the federal poverty level, and have assets below $2,000. The randomly selected individuals chosen by the lottery who completed the application process and met the eligibility criteria were enrolled in OHP Standard. Following some selection rules, the data set contains only these 74,922 individuals. Of these individuals, 29,834 were selected as treatments (i.e. won the lottery and were given the chance to apply for health insurance); **treatment** status is indicated by the variable `treatment` in `oregonhie_descriptive_vars.dta`.

Crucially, the lottery selected individuals, but the opportunity to apply for health insurance was extended to **all household members** of lottery winners: **treatment selection is random only conditional on the number of household members on the waiting list** (this is given by the variable `numhh_list` in `oregonhie_descriptive_vars.dta`. For example, an individual could sign up his or herself as well as a spouse for the lottery, and both have equal probability of being chosen. Thus, this person and his or her spouse are twice as likely to win the opportunity to apply for health insurance as someone who only added their own name to the list, without adding other household members. In short, those in a larger household are more likely to be selected into the **treatment** condition.

Merging Datasets. All datasets contain observations at the individual level. Observations can be linked across different `.dta` files by the unique identifier `person_id`, which appears in all datasets. No other variable appears across multiple datasets.

Dataset Descriptions. Below we describe the 3 datasets concerned in this task:

`oregonhie_descriptive_vars.dta`

This dataset contains demographic characteristics that were recorded when individuals signed up for the lottery and lottery selection. You may refer to the code book `oregonhie_descriptivevars_codebook.pdf` for descriptions of the variables in this dataset.

`oregonhie_stateprograms_vars.dta`

This dataset contains information from the state of Oregon on individuals' participation in the following state programs: Medicaid, the Supplemental Nutrition Assistance Program (SNAP), and Temporary Assistance to Needy Families (TANF). You may refer to the code book `oregonhie_stateprograms_codebook.pdf` for descriptions of the variables in this dataset.

`oregonhie_ed_vars.dta`

This dataset contains variables derived from administrative data of all visits to twelve hospital

emergency departments in the area of **Portland, Oregon**. You may refer to the code book `oregonhie_ed_codebook.pdf` for descriptions of the variables in this dataset.

For more detailed descriptions of the entire data set, please read the documents `ohie_startguide.pdf` and `ohie_userguide.pdf`. All the data and their descriptions can be found here: https://www.nber.org/research/data/oregon-health-insurance-experiment-data.

Questions

You job in this task is to estimate **the causal effect of being selected by the lottery and enrolling into the medicaid program on emergency department utilization**. Please address the following questions. You may need to merge different data sets together.

(a) (3 points) **Initial Data Pre-processing and Balance Check**. Because the individuals selected by the lottery have the opportunity to apply for the OHP Standard program, you need to create dummy variables for the number of people in household on lottery list. Why should we use dummy instead of numeric variables for this setting? Because the ED visit data is only available for the Portland area, we will mainly work with the data observations in this area. Please use the OLS approach to check the balance of the treatment and control groups for the individuals in the data sample. Specifically, you need to regress the variable which you want to conduct balance check on (i) the treatment variable, and (ii) the dummy variables for the number of people in household on lottery list. Please conduct balance check for the following variable with the full OHIE data sample (N=74,922):

- Included in the emergency department (ED) sample, i.e., the Portland area, (N=24,646).

Please conduct balance check for the following variable with the ED data sample (N=24,646):

- Year of birth
- Female
- Signed up self for lottery
- Any ED visit, pre-randomization (censored)
- Number of ED visits, pre-randomization (censored)

(b) (3 points) **Causal Effect of Being Selected by Lottery**. Next, for the data sample in the Portland area (N=24,646), please estimate the causal effect of being selected by the lottery on the following outcome:

- Whether an individual was enrolled in any Medicaid program (including the OHP Standard) between the earliest notification date in the sample (10 March 2008) and 30 September 2009.

Please include as appropriate necessary features into your regression model. In particular, you need to include the dummy variables for the number of people in household on lottery list (why?). Please discuss/justify your choice of features included in the regression model. What is the average treatment effect of being selected by the lottery on being enrolled in any Medicaid program?

(c) (2 points, extra credit) **Causal Effect of Enrolling into a Medicaid Program on ED Visits.** Read the reading on Instrumental Variables (11a) and use 2-stage-least-squares to estimate the average treatment effect of enrolling into a medicaid program on (i) the probability of any ED visits during the study period and (ii) the (censored) number of ED visits in the study period. Again, you need to include certain features into the regressions to remove the bias and/or reduce the variance of your estimation. Please report your estimation results, including the 95% confidence intervals.

Note: This is actually the core part of the analysis, which leads to a published paper in the Science Journal, "Medicaid Increases Emergency-Department Use: Evidence from Oregon's Health Insurance Experiment". Unfortunately, we did not discuss the IV analysis in detail in our class, so we leave it as extra credits.