Final Project Check-In

Data Science II (STAT 301-2)

Derex Wangmang

February 25, 2021

Contents

Data Collection

I successfully obtained the data by downloading the dataset from the University of California at Irvine's Machine Learning Repository, found here. Since the data had one observation per row, with variables in each column, I considered the dataset to be tidy and not requiring additional organization.

Data Processing

When reading the variable names, I noticed that some characters were problematic. For instance, one variable had the degree symbol within its title (`Temperature(°C)`). To read in the data properly, I used a different encoding: `bike <- read_csv("data/unprocessed/SeoulBikeData.csv", locale = locale(encoding = "ISO-8859-1"))`. Then, I tidied the names by using `janitor::clean_names()`., ensuring all names were standardized.

No missing values were present. However, I observed that one variable had one level for a majority of the data (~97%) and a second level for the rest of the data. I processed the data by only including the 97% of the data where the variable is a given value and afterwards dropping the variable from the dataset.

Additionally, some variables, such as the `date` and `seasons`, seemed unnecessary. The date, for instance, allows the user to recognize the season. At the same time, the exact date may constrict the data to small groupings. The data may be better grouped by seasonality or month. After consulting with the instructional team, I decided to remove `seasons` and only keep the month of the date, stored as a numeric variable.

No issues existed with the target variable.

I have time-dependent data. I have a `hour` variable that tracks the number of rental bikes during each hour.

Changing Dataset

Since my data did not present major issues, I will not be changing my dataset.

Data Splitting

I will be using an initial split with 70% training, 30% testing. I plan to use stratified sampling, splitting by the target variable, `rented_bike_count`. Since I dropped the `date` variable, I simplified the dataset and removed that data which that could be used for time series. Instead, I plan to use time-based variables, such as `month` and `hour`, as numeric variables in my models.

Current Progress

After processing my data according to the above, I developed folds using repeated *k* cross validation: `bike_folds <- vfold_cv(bike_train, v = 10, repeats = 3)`. I also created a recipe by `bike_recipe <- recipe(rented_bike_count ~ ., bike_train) %>% step_dummy(all_nominal()) %>% step_normalize(all_predictors())`. I used this recipe for my two models so far, a linear regression model and a random forest model. Next, I plan to use the recipe within other models, such as a ridge or lasso model.

Citations

Seoul Bike Sharing Demand Data Set. (2020). UCI Machine Learning Repository [https://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand]. Irvine, CA: University of California, School of Information and Computer Science.