Capstone Project - Data Wrangling
Intermediate Data Science
Bob Newstadt
December 14, 2018

Describe the data wrangling steps used to clean your capstone project data set. What kind of cleaning steps did you perform? How did you deal with missing values, if any? Were there outliers, and how did you decide to handle them?

The data set contained 217,348 observations, each one a trip, and 74 features.

Trips were removed from the dataset when the target value could not be derived for logical reasons (if the trip was canceled before the driver arrived to the pick-up location).

Ten interesting pairs of timestamps were selected and used to calculate duration features.

Timestamp attributes were processed into features representing the quarter, year, day, hour, weekday, week of year, and day of year features. Any NULL values were replaced with the lowest possible value.

Categorical features were turned into 1 hot features, n-1 boolean features where there were n possible values..

When calculating complex features such as avg_prior_arrived_late_seconds any NULL values were replaced by the overall mean of the dataset.

Boolean columns containing True/False were wrangled into features containing 1/0, respectively.

Finally any remaining NULLs were filled with 0s.

The outlier values of target variable (arrived_seconds_after_scheduled_start) were truncated to +/- 30 minutes. The reasoning was that values beyond those limits were irrelevant for prediction and distracted from the visualization of the data.

The final wrangled test set contained 125,675 observations, each one a trip, and 216 features. The biggest reason for the 60% reduction in observations was for trips canceled far in advance of the scheduled start.