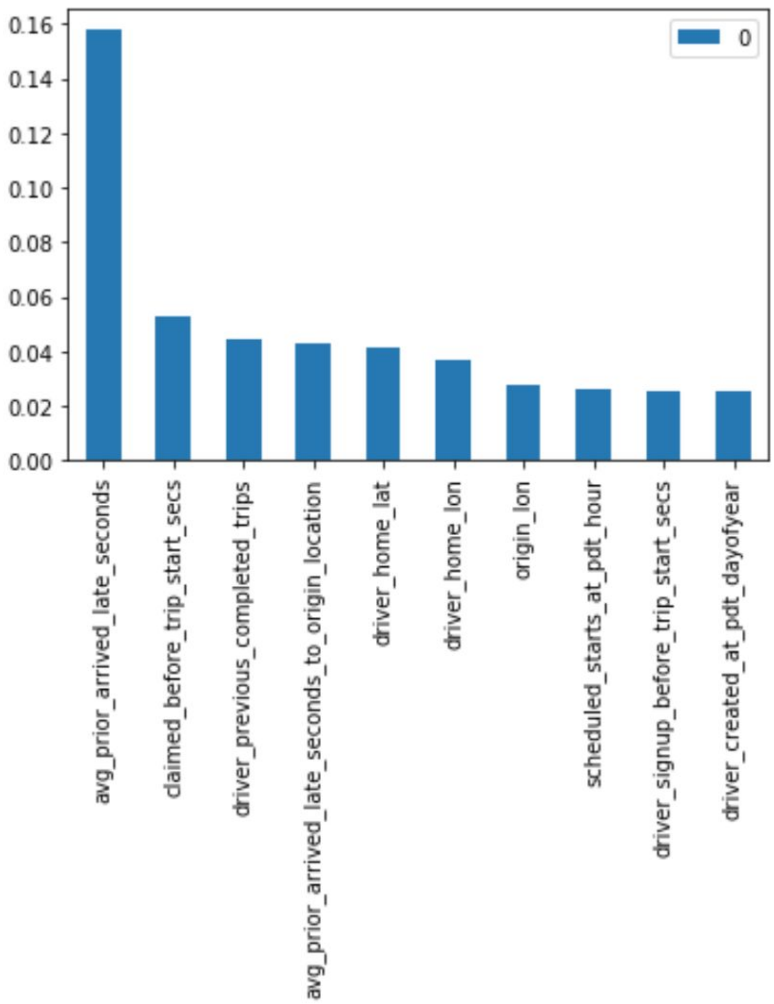
Inferential statistics steps performed for capstone project.

Bob Newstadt
Springboard.com Intermediate Data Science

My capstone project was titled *“Will a scheduled ride arrive on-time?”* I built data science models to predict when the vehicle would arrive at the pickup location relative to the scheduled start time. This was based on a dataset from a ride hailing company. For those of you who want to follow along the capstone project, presentation, and jupyter notebooks are available on github.

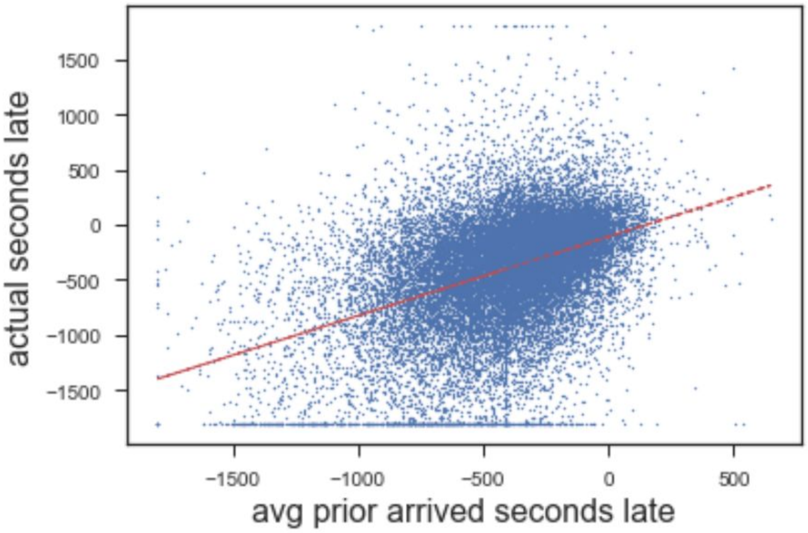
Inferential statistics were ***not*** used to validate hypotheses about the dataset or produce p-values. The focus of the project was instead to get the best machine learning model performance. The models build determined which features were important to optimizing the regression. Measuring and reducing correlations between pairs of independent variables was done indirectly through trials with different feature sets. I’m sure there are many correlated features in the data set which affected interpretability of the simple linear models. To deal with this I switched to models which could tolerate these correlations well (gradient boosted regression trees).

Features with the highest correlations to the dependent variable were identified by this ensemble model. This figure shows the top correlated features of Model 22. Study of these so-called “importances” lead me to infer that the two top causes of lateness were the driver themselves and the driver replacement events. The driver’s past performance, the driver’s experience, and the driver’s location are all useful attributes to the model. The driver replacement event is correlated to lateness based on the claim time being shortly before the start time of the ride. The original driver who claimed the ride previously was removed from their ride for some reason. Driver replacement happened frequently enough to be a problem which appears in the data.

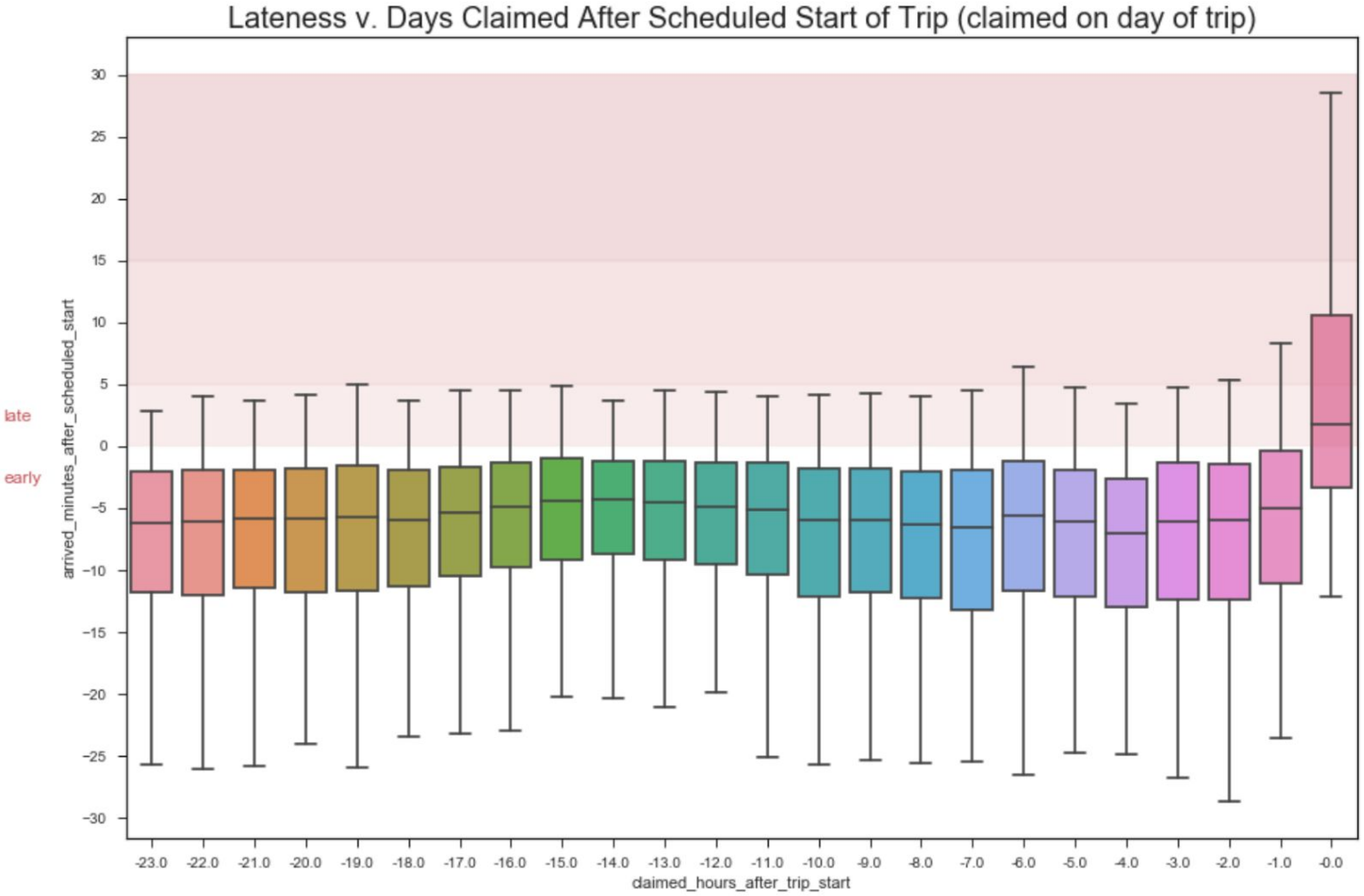


Earlier, Model 21 used 3 “ID” features in the top 10 which I suspected would not generalize well. “ID” features are unique numbers which identify a driver or location. When I removed all the “ID” features the performance on the test set improved and was closer to the performance on the training set.

This scatter plot shows the correlation to the target (dependent) variable with the top feature used in the model. This feature measures how late the driver has been on the most recent trips before this one. Drivers who have tended to be late in the past tend to be late.



Other correlations were confirmed by box plots. This one shows the lateness correlates with the number of hours before the start of the trip that the ride was claimed by the driver. Note that very near the start of the trip the driver replacement events tend to happen. This feature is the #2 most important feature used to build the predictive model.


Lateness v. Days Claimed After Scheduled Start of Trip (claimed on day of trip)

These visualizations are from an additional notebook which was used to assess relationships between variables discovered by the models.