Tim Vogel

Applause Assessment

Data Science Thought Exercise

In order to predict the number of bugs a new tester would find, the first step would be organizing the data. This would entail aggregating the number of bugs found by user, so that we could establish our target variable, “total\_bugs\_found.” The next step would be creating the feature columns that I’d train the model on. Features would include country, time since last login (as a proxy for a tester’s activity), and device availability data (dummified values to indicate whether a tester possesses said device or not). I’d then change the values for country from string data to numerical values e.g. “US” = 1, “GB” = 2, “JP” = 3. At this point, the dataframe would be ready for modeling.

Before training the model, I’d split the data into three groups (train/validation/test). I would randomly select 70%, or 70,000 rows on which the model would be fit. I’d use a linear regression model as the target variable we hope to predict, total\_bugs\_found, is a continuous variable. The metric I’d employ to gauge the success of the model, would be R^2, to see how much variance in predicted results the model can account for. I’d select this metric as the regressors/features in question are independent of each other, i.e. a tester’s possession of one device does not necessarily indicate their possession of another. Further, this would allow the model to learn profiles of testers, and the devices at their disposal to understand/predict the number of bugs such a profile would find. Finally, I would test the model on the remaining 15% of the data that the model has not yet seen. I could then view the residuals and see how successful the model was in minimizing error in predictions on the test data. Ideally, the model would be able to account for more than 50% of variance and therefore be performing better than chance when making predictions on the numbers of bugs.