Michael English

Springboard Data Sciences Certification Course

Machine Learning

You've collected your data, cleaned it up, wrangled it into shape and explored it. Now it's time to perform some in-depth data analysis using machine learning. This step depends on you and your mentor, but here are some suggestions to get you going.

1. **How do you frame your main question as a machine learning problem? Is it a supervised or unsupervised problem? If it is supervised, is it a regression or a classification?**

Original Research question: Can we approximate overcrowding in Hartsfield-Jackson in March 2019?

Research Question with Machine Learning in Mind: Are we able to develop a predictive model that can estimate population size at Hartsfield-Jackson in a given month and determine overcrowding?

It is a supervised problem, the basis of a function with independent and dependent variable (an output). Based on the flow of data, it has been determine that it is a regression.

1. **What are the main features (also called independent variables or predictors) that you'll use?**

My independent variable is the number of flights, aptly described as FpCity (Flights per City) in the dataset. This number will be used to estimate the dependent variable, Passengers, which is dependent on the size of aircraft and the number of flights that are available to the consumer.

1. **Which machine learning technique will you use?**

Because it is a regression, we needed to decide between a logistic and linear regression. The linear regression seemed to fit best as the number of passengers is not limited by any factors, and represent a variable with the potential increase over time.

1. **How will you evaluate the success of your machine learning technique? What metric will you use?**

For this project, a training dataset was created with 20% of the data from the original dataset and a cross-validation will be performed. The cross-validation will run the model with varying variable and determine which has the most optimal:

1. p-value, which displays the significance of the model compared to a null model, which is usually a model that displays averages of the dependent variable. It is a matter of laying down a baseline for accuracy of the model. The lower the p-value (ranging from 0 to 1), the more likely the model is more accurate at approximation than the null model, and the null model can be thrown out.
2. R2–value, which determines how “close” the points are to the regression line. The values range from 0 to 1, with values closer to 1 indicating that the data points are “closer” and more tightly correlated to the regression line. In layman’s terms, the closer the value is to 1, the better approximation of the data points the predictive model will give. If it is closer to 0, then the independent variable may need to be changed as it does not provide enough context or influence on the dependent variable and is thus not useful for the model.
3. Root Mean Square Error, which gives a “standard deviation” of the data points; the average of variance of each of the points from each other. This value can be infinite, but the smaller the RMSE, the better approximation – the values of the dataset are not as sparse. This is important as data points can line up well and be set up around another regression line that was not considered inititally.