Ryan Harper
Sep 20th, 2018

Thinkful Capstone Proposal:
Assessment of Bike Accidents in Portland, Oregon

Data Set:
Oregon Department of Transportation Crash Data

Capstone Summary:

For the final capstone, I will focus on analyzing bike accidents in Portland, Oregon. I will use the *Oregon Department of Transportation crash data* as my data set for this analysis. Traffic in Portland (like most big cities) is getting worse so it will be of growing necessity for ODOT to assess the variables behind any increase in accidents by year (location, speed, vehicle type, weather, etc.). I emphasize bicycle accidents, because Portland is famous for its bike friendly roads and because fatal bike accidents are frequently reported in local news.

Data Wrangling:

Because I have already worked this data set, I am familiar with a lot of the preprocessing necessities. The data set has a very large number of features and also includes a code manual (https://www.oregon.gov/ODOT/Data/documents/CDS_Code_Manual.pdf). The first problem with the data is that each row in the csv file includes every vehicle participant of an accident (a lot of repeat data and missing values for vehicles that just witnessed the accident). Another big issue is that the geo tags are not always accurate and can sometimes be off by a few blocks (important for assessing accidents on highways). Fortunately, I have already been able to address most of these issues but would like to use new strategies I learned from Thinkful to better clean the data (ex. Imputation vs. deletion).

Preprocessing:

The goal of this capstone is to estimate future bike accidents from previous years. Because of this, I will be working with a time series and will likely be wanting to try and estimate a polynomial fit with some form of regression. This will be an opportunity for me to practice using boosting methods (like Gradient Bossting from Thinkful) to better fit the data. I also intend to use PCA (or SVC) to analyze the large number of features in the data.

Models and Metrics:

The MVP of the capstone is to predict a polynomial fit to bike accidents for 2016/2017 based on existing data (training=2009-2015, test=2016/2017). I'll establish a baseline RMSE value from a simple linear fit and then compare that to the RMSE values from a number of different regression models.

Problems:

There is also a lot of information in the data to explore and I know it will be easy for me to get sidetracked with understanding all of the features in the data. The portion of reported bike accidents is not big so it might be hard to analyze the sample. Also, I haven't done a lot of work with time series in models, so I anticipate other unforeseen problems with polynomial fits in my predictions from linear regression models.