Data Agnosticism: Feature Engineering Without Domain Expertise In Python

Nicholas Kridler
Accretive Health

April 2, 2013

Abstract

Bits are bits. Whether you are searching for whales in audio clips or trying to predicit hospitalization rates based on insurance claims, the process is the same: clean the data, generate features, build a model, and iterate. Better features lead to a better model, but without domain expertise it is often difficult to extract those features. Numpy/Scipy, Matplotlib, Pandas, and Sci-kit Learn provide an excellent framework for data analysis and feature discovery. This is evidenced by high performing models in the Heritage Health Prize and the Marinexplore Right Whale Detection challenge. In both competitions, the largest performance gains came from identifying better features. This required being able to repeatedly visualize and characterize model successes and failures. Python provides this capability as well as the ability to rapidly implement and test new features. This talk will discuss how Python was used to develop competitive predictive models based on derived features discovered through data analysis.