Machine Learning Engineer Nanodegree

Capstone Proposal

Al Piran

May 25, 2018

Proposal

Domain Background

Being a wine enthusiast I've always been curious as to what makes one wine better than the other. Is it just a matter of opinion or are there characteristics that are common between wines that are considered to be high quality. Making wine is both an art and science. The process of going from grapes to an actual bottle of wine is complex and involves many variables that affect the quality of the wine. Those variables can be measured and tweaked, but the ultimate judge of quality is the human palette. This projects intends to use machine learning and the Wine Quality Dataset from UCI in order to show what objective factors in the wine making process have the biggest effect on the subjective quality of wine.

Problem Statement

The subjective quality assessment that wine experts perform has a lot riding on it. A rating above seven on a scale of 10 in a magazine such as wine spectator can have a big effect on the price and marketability of the product. But can this assessment be understood in terms of several measurable objective variables? If so to what extent can this be done? and of the 11 variables that we have in this dataset which are the 2 or 3 most important ones for being able to predict the subjective quality assessment.

Datasets and Inputs

The UCI red wine dataset will be used which is hosted below at UMass.

http://mlr.cs.umass.edu/ml/machine-learning-databases/wine-quality/winequality-red.csv

Solution Statement

The solution will tell us several things, namely:

1. How good is the benchmark model in predicting the output based on an R2 score and error.
2. How much better would Polynomial Regression or Random Forrest be, again using R2 score and error, to establish the accuracy of each model.
3. What are the 2 or 3 key variables in predicting the output. Recursive Feature Elimination (or RFE) works by recursively removing attributes and building a model on those attributes that remain. It uses the model accuracy to identify which attributes (and combination of attributes) contribute the most to predicting the target attribute.

Benchmark Model

Linear Regression will be used as a benchmark model.

Evaluation Metrics

The R2 score is the main metric that will be used to evaluate both the benchmark and the solution model. The R2 (or R Squared) metric provides an indication of the goodness of fit of a set of predictions to the actual values. In

statistical literature, this measure is called the coefficient of determination. This is a value between 0 and 1 for no-fit and perfect fit respectively.

Project Design

The project will have the following steps:

1. Explore the data including. This includes the min/max, mean for all the variables as well as graphing them to visualize the distributions.
2. Establish the benchmark model, in this case it will be linear regression, and evaluate using the above mentioned metrics.
3. The data will be evaluated using the solution model namely polynomial regression and the ensemble algorithm random forest.
4. The best model will be used and fine tuned using hyperparameters.
5. The most relevant features will be established using Recursive Feature Elimination (or RFE)

References

https://archive.ics.uci.edu/ml/datasets/wine
https://machinelearningmastery.com/metrics-evaluate-machine-learning-algorithms-python/
https://machinelearningmastery.com/feature-selection-machine-learning-python/