UDACITY

< Back to Machine Learning Engineer Nanodegree

Finding Donors for CharityML

REVIEW

CODE REVIEW

HISTORY

Meets Specifications

Dear student,

Great re-submission 👍

You have corrected and re-ran most of the sections and I really appreciate your work. I would recommend you to work on some kaggle competitions on supervised models. Initially you just observe the winning solutions and their reason of selection of model for given data set. Later you can test with your own choice and compare the results. And this is a good way to strengthen your understanding and easy to work on real world applications as well.

keep up the good work!

Exploring the Data

Student's implementation correctly calculates the following:

- **Number of records**
- **Number of individuals with income >$50,000**
- **Number of individuals with income <=$50,000**
- **Percentage of individuals with income > $50,000**

Preparing the Data

Student correctly implements one-hot encoding for the feature and income data.

Evaluating Model Performance

Student correctly calculates the benchmark score of the naive predictor for both accuracy and F1 scores.

You correctly calculated both accuracy and f-score. Good work!

You could check this link for further understanding precision and recall.

The pros and cons or application for each model is provided with reasonable justification why each model was chosen to be explored.

Please list all the references you use while listing out your pros and cons.

Please check below videos on svm, decision trees and ensemble methods to futher understand the concepts.

svm
Decision trees
Ensemble methods

Student successfully implements a pipeline in code that will train and predict on the supervised learning algorithm given.

Student correctly implements three supervised learning models and produces a performance visualization.

You will notice that there is no f-score for 1% sample size in SVC. This is happening because the sample size is so small that there is no positive class correctly predicted, therefore TP = 0. If you look at the formula, you will see that it depends on precision and recall on both nominator and denominator. Both of them depends directly on TP, that is Precision = TP/(FP+TP) and Recall = TP/(FN+TP). Since TP=0 for 1% sample size, both Precision and Recall are equal to zero. This makes F-score = 0/0, which is undefined.

Improving Results

Justification is provided for which model appears to be the best to use given computational cost, model performance, and the characteristics of the data.

Student is able to clearly and concisely describe how the optimal model works in layman's terms to someone who is not familiar with machine learning nor has a technical background.

The final model chosen is correctly tuned using grid search with at least one parameter using at least three settings. If the model does not need any parameter tuning it is explicitly stated with reasonable justification.

You could actually go well beyond grid search and implement 'pipelines' where the whole machine learning process becomes 'grid-searchable' and you can parameterize and search the whole process though cross validation.
http://scikit-learn.org/stable/modules/generated/sklearn.pipeline.Pipeline.html
And yes you can try out several algorithms automatically as well too! Watch out though this is pretty advanced stuff, here is a great, informative, top notch tutorial from Zac Sewart!
http://zacstewart.com/2014/08/05/pipelines-of-featureunions-of-pipelines.html

Student reports the accuracy and F1 score of the optimized, unoptimized, models correctly in the table provided. Student compares the final model results to previous results obtained.

Feature Importance

Student ranks five features which they believe to be the most relevant for predicting an individual's' income. Discussion is provided for why these features were chosen.

Student correctly implements a supervised learning model that makes use of the `feature_importances_` attribute. Additionally, student discusses the differences or similarities between the features they considered relevant and the reported relevant features.

An alternative feature selection approach consists in leveraging the power of Recursive Feature Selection to automate the selection process and find a good indication of the number of relevant features (it is not suitable for this problem because that is not what is required by the project rubric, though it is generally a very good approach).
http://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html

Student analyzes the final model's performance when only the top 5 features are used and compares this performance to the optimized model from Question 5.

DOWNLOAD PROJECT

RETURN TO PATH