UDACITY

Return to Classroom

Finding Donors for CharityML

REVIEW

CODE REVIEW

HISTORY

Meets Specifications

Excellent job you got it!

🎉👏 Congratulations! You met all the requirements smoothly! It was a pleasure to go through your whole code. Please, take into account some comments I added to the remaining part of the rubric below. I'll also reiterate the suggestion of creating some documentation to showcase this work since it is a great project that you achieved here. 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.

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.

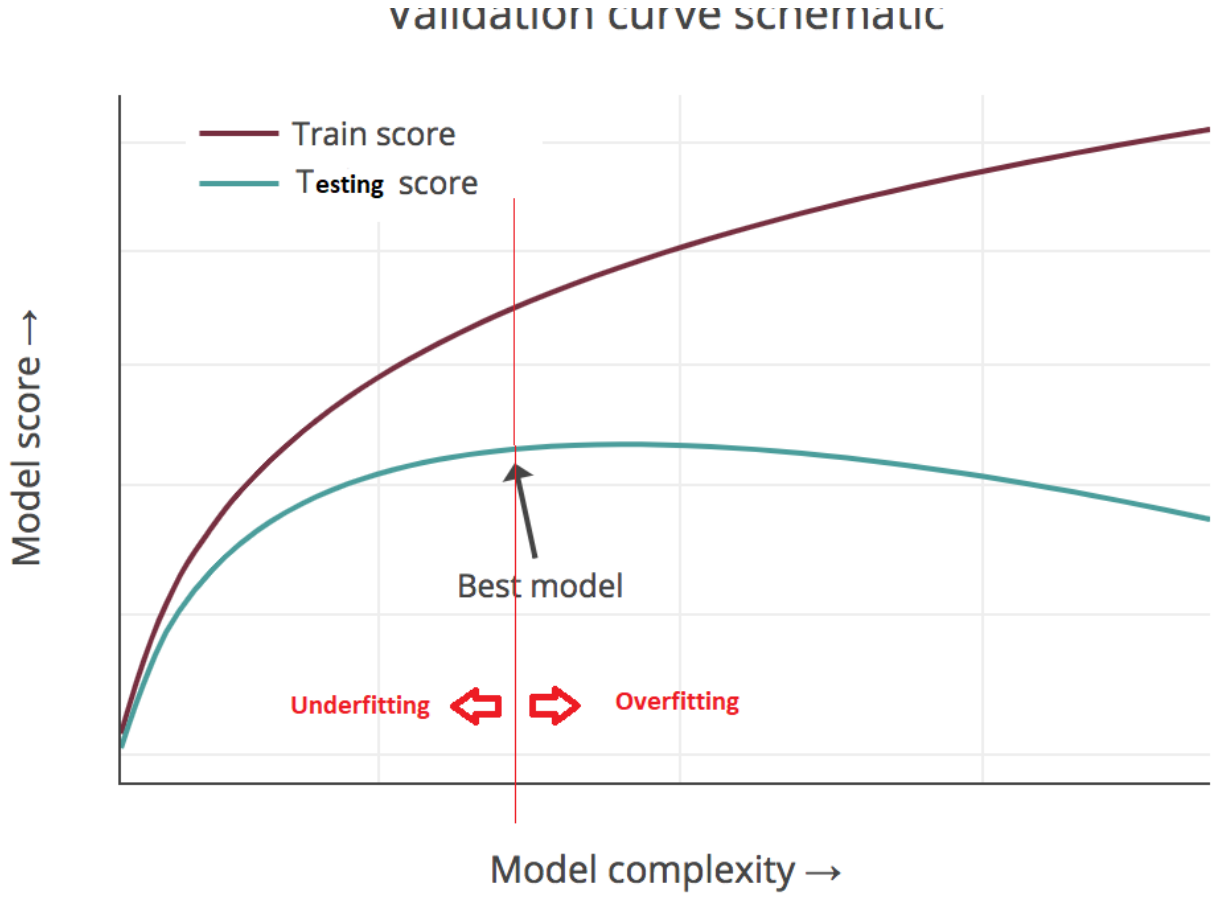
Student successfully implements a pipeline in code that will train and predict on the supervised learning algorithm given.

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

- classifiers trained with:
  - 361 samples (1%) ✔
  - 3617 samples (10%) ✔
  - 36177 samples (100%) ✔
- Plots performed correctly ✔
- random_state parameter set on classifiers ✔
- All models are set with their default arguments as required in the notebook ✔

Perfect! Thank you for fixing this, we are now able to replicate your results! 🙂

Comment: We wanted you to perform the comparison with default parameters because there are many models that overfit to the data. Hence, we want you to experience that. For instance, It would be great if you also mention that DT is overfitting to the data. You can notice this by looking at the training score vs the testing score where the difference in performance is considerable. In machine learning, we should opt for the model that better generalizes the data. That's why we prefer to avoid any form of overfitting in our model as is the case in this situation. Below you can see an illustration of the idea:



Note: You could've still selected an overfitted model (in case you wanted to) but that would've meant that you need to fine-tune its parameters to guarantee that the model is no longer overfitting. In this project, this could be done in the grid search section. For this project, we are not checking for overfitting, but I think it is very important that you are aware of this problem in case you want to enhance your work even more.

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.

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

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.

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

Rate this review

START