



[Return to "Machine Learning Engineer Nanodegree" in the classroom](#)

DISCUSS ON STUDENT HUB

Finding Donors for CharityML

REVIEW

CODE REVIEW

HISTORY

Requires Changes

3 SPECIFICATIONS REQUIRE CHANGES

Very impressive submission here!

It seems that you have acquired great knowledge of supervised learning algorithms and are now ready to apply it on your own problems.

However, a few very minor changes are required to meet our rubric.

Keep doing this great job! Keep experimenting and be curious.

All the best for your future.

Happy Learning!!

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

Nice job finding out the correct numbers!

I would further encourage you to do more exploratory data analysis (EDA) on this dataset using pandas and other libraries like seaborn. You may refer to the following resources to know more about how EDA is done using these libraries. [link1](#) [link2](#)

Preparing the Data

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

Nice job encoding the features using `get_dummies`! Nice job converting the target labels to correct numerical values!! However, it would have been better if you had converted `income` such that records with `<=50K` is set to `0` and records with `>50K` set to `1` instead of creating two columns. It would have been achieved by a code line like this -

```
income = pd.get_dummies(income_raw)
income = income['>50K']
```

Please note that there are different encoding strategies suitable for different tasks. I would encourage you to check out the following resources to know more about encoding strategies. [link1](#) [link2](#)

Evaluating Model Performance

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

Impressive calculations! Kudos!

Also, please note that, this dataset is not balanced, i.e., the number of samples of positive class (income > 50k) is very less compared to the number of samples of negative class (income <= 50k). In such cases, accuracy is not the correct metric to evaluate the performance of any model. F1 or F-beta scores are more useful. You may check out the following resource to know more about metrics. [link1](#)

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.

Good discussion on the real-world applications, strengths, and weaknesses of the models you chose! It seems that you have a very good understanding of the models.

However, you need to elaborate more on the candidacy of the models given what you know about the data. Please check the properties of the dataset and the strengths of the models you chose. It would be good if you can comment on how the models you chose would leverage the properties of this dataset. One of the goals of this section is to encourage you to develop a solid understanding on model selection, i.e., which model to choose for a particular problem at hand. This is a 'must-have' skill for any machine learning

practitioner. Please take a look at the following resources which will give you a rough guideline. [link1](#) [link2](#) [link3](#)

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

Great job implementing the training and prediction pipeline!

Please note that, it is always good to write a function like `train_predict` which can be called from other modules with different samples of train and test data. It makes the code cleaner, organised and easy to debug.

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

Great job training the models with different sample sizes and getting the predictions on test set! Nice use of random seeds! Nice visualizations to compare the performances of the models!

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.

Good justification for choosing Gradient Boosting.

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.

Very good job indeed. It seems that you have a very good understanding of the model. However, please describe what are weak learners and what are decision trees. We can't expect a non-technical person to know these. Also, please elaborate on what you mean by "functional gradient descent". The goal is to explain how the model of your choice (in this case Gradient Boosting) gets trained and makes predictions in simple intuitive terms without using any mathematical terms.

I would further encourage you to refer the following resource which is one of the best intuitive explanations on Gradient Boosting. [link](#)

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.

Great job running so many grid searches. Very nice implementation of GridSearch and the choice of hyperparameters and their values were great. This is one of the best scores I have seen on this dataset. Awesome job! Kudos!

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.

Good work comparing your tuned model, the untuned model and the Naive one!

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.

Nice intuition, these are some great features to check out!

However, you need to give an intuitive reason on why you think these features would be most important, e.g., people with higher education generally tends to have better jobs.

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.

Great job implementing Gradient Boosting to get feature importance!

Please note that different algorithms use different strategies to get feature importance. So, the result could be different if you choose a different model. I would encourage you to try AdaBoost or Random Forest to get feature importance and check if your intuition matches with their results.

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.

Nice analysis of the final model's performance with only the top 5 features.

 RESUBMIT

 [DOWNLOAD PROJECT](#)



Best practices for your project resubmission

Ben shares 5 helpful tips to get you through revising and resubmitting your project.

[Watch Video](#) (3:01)

RETURN TO PATH

Rate this review