Finding Donors for CharityML

REVIEW | HISTORY

Meets Specifications

Hi there, it's Cláudio! Thanks for sending all the required files for the review process and for all code executing without any issue.
Congratulations for this project submission and for the quality presented in this project. You really did a great job.

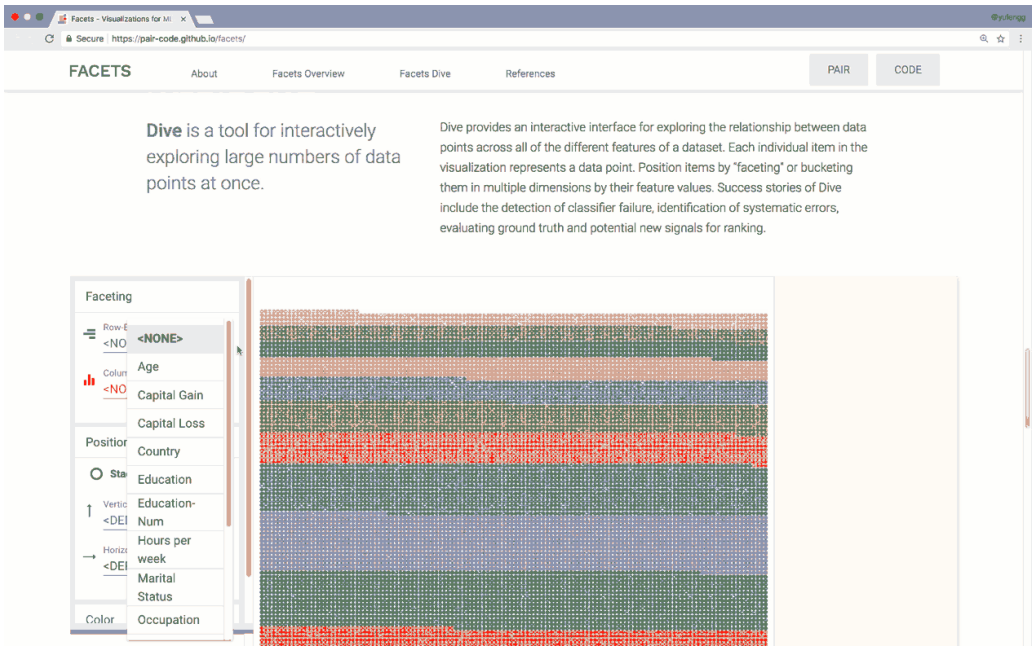
I hope you had enjoyed doing this project and put in practice good and important concepts from machine learning. I will leave my contact below in case you have any doubt about this review as well to get connected.

That's all. Enjoy learning and keep it up the great work!

Finally, I wanted to share a interesting tool from google that helps machine learning engineers to understand the data really fast and then make a decision on what type of algorithm it will fit better that data, called: Facets - Visualizations (https://pair-code.github.io/facets/). Definitely check it out.

The power of machine learning comes from its ability to learn patterns from large amounts of data. Understanding your data is critical to building a powerful machine learning system.

Facets contains two robust visualizations to aid in understanding and analyzing machine learning datasets. Get a sense of the shape of each feature of your dataset using Facets Overview, or explore individual observations using Facets Dive.



Thank you.
Cláudio

Email: cgimenest@uol.com.br
Linkedin: https://www.linkedin.com/in/claudiogimenestoledo/

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**

Great job. You have correctly reported the numbers:

Total number of records: 45222 ✅
Individuals making more than $50,000: 11208 ✅
Individuals making at most $50,000: 34014 ✅
Percentage of individuals making more than $50,000: 24.784% ✅

Suggestion:

- You may also want to use the value_counts () (https://pandas.pydata.org/pandas-docs/stable/generated/pandas.Series.value_counts.html)

```
n_at_most_50k, n_greater_50k = data.income.value_counts()
```

  This will return the count for each value you do have in the columns specified

Bonus:

- Usually it's a great idea to start the exploratory analysis getting the statistics from the dataset. I usually get that from Pandas library using the method describe:

```
pd.dataframe.describe()
```

Link for reference: https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.describe.html

Preparing the Data

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

Awesome. You have implemented correctly the one-hot encoding and have the correct number of features:

103 total features after one-hot encoding.

You used correctly the method from pandas to do that:

```
# TODO: One-hot encode the 'features_log_minmax_transform' data using panda
s.get_dummies()
features_final = pd.get_dummies(features_log_minmax_transform)
```

Suggestion:

- You may also convert to numerical values the targets using an lambda function, example:

```
income = income_raw.apply(lambda x: 0 if x=='<=50K' else 1)
```

Bonus:

- Here I will leave some good articles about the importance of using one-hot encoding:
  https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/
  http://forums.fast.ai/t/to-label-encode-or-one-hot-encode/6057
  https://hackernoon.com/what-is-one-hot-encoding-why-and-when-do-you-have-to-use-

it-e3c6186d008f

Evaluating Model Performance

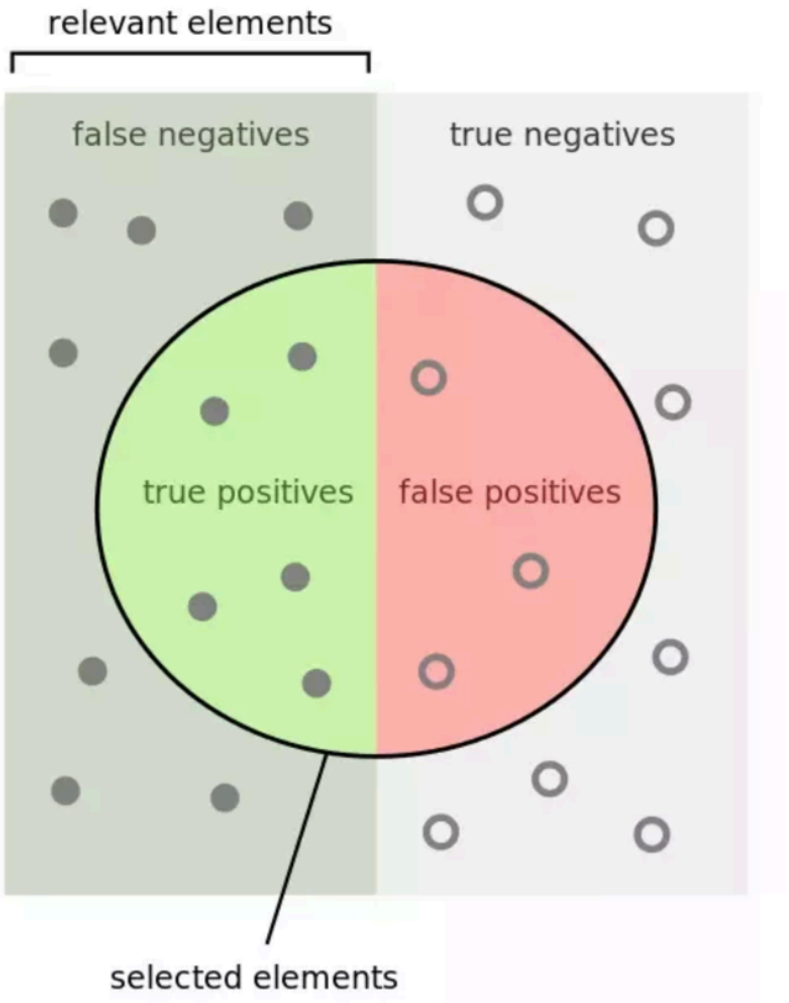
Student correctly calculates the benchmark score of the naive predictor for both accuracy and F1 scores.

Great job calculating correctly the accuracy score and fscore

> Naive Predictor: [Accuracy score: 0.2478, F-score: 0.2917]

Bonus:

- Here I will leave a great article about scores for precision and recall: https://www.quora.com/What-is-the-best-way-to-understand-the-terms-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.

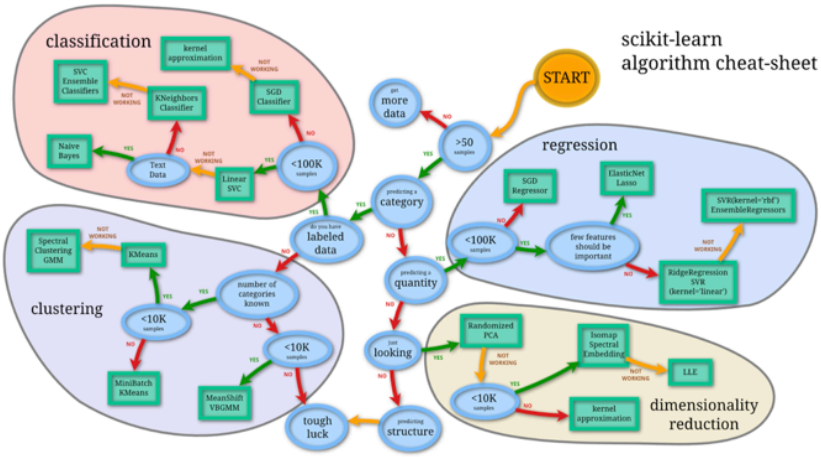
Please list all the references you use while listing out your pros and cons.

Awesome job describing pros and cons. Also noticed you have provided link for further reference. Good job.

Suggestion:

- It's always a good idea to link references, images and study cases about what we are trying to convey to get a more consistent response and reinforce our outcomes.

- A great summary about models and its application in a outlook view:



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

Awesome work done.

Bonus:

- The purpose of the pipeline is to assemble several steps that can be cross-validated together while setting different parameters. For this, it enables setting parameters of the various steps using their names and the parameter name separated by a '__', as in the example below. A step's estimator may be replaced entirely by setting the parameter with its name to another estimator, or a transformer removed by setting to None
Link for reference: http://scikit-learn.org/stable/modules/generated/sklearn.pipeline.Pipeline.html

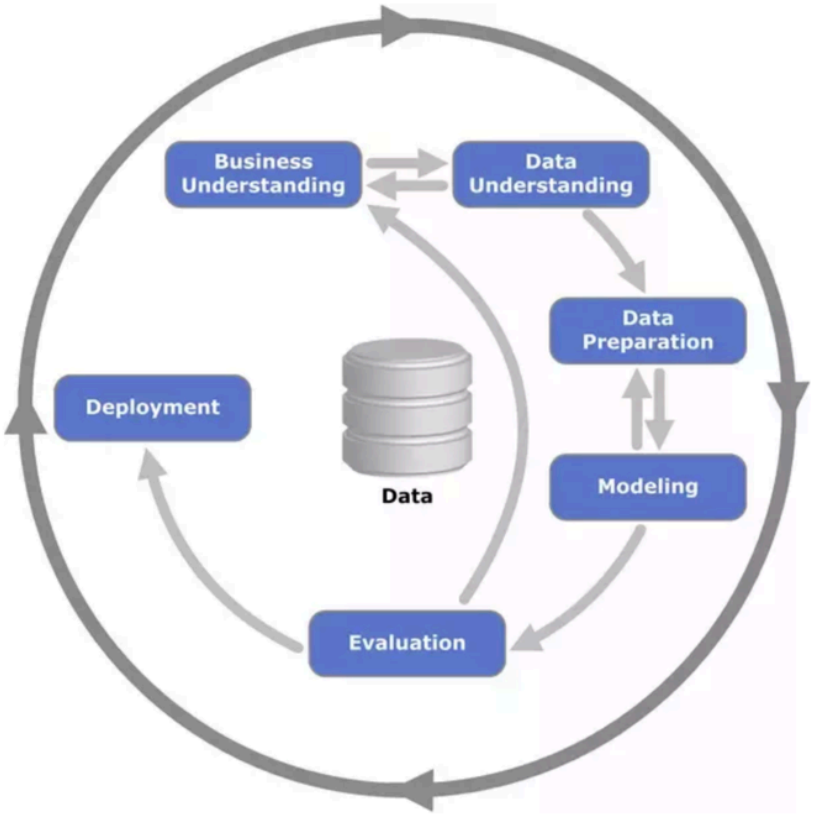
Student correctly implements three supervised learning models and produces a performance visualization.

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.

Nice explanation and I agree with your final model selection.

Below a great summary around the process for models and then check and evaluate:



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.

Good job. This is a very important step that we should carefully consider when explaining to all type of audience. Usually not all in the audience knows deeply the technical aspects and it's up to us to open the "black box" and convey a clear message on why we are choosing that model and that is the best solution we may have for the given problem.

Suggestion:

- Usually it's a good idea to provide link for further reference as well images, diagrams and business cases when describing in layman's terms.

Bonus:

- I will leave here some good resources about layman's terms in AI:
  https://www.quora.com/How-can-you-explain-artificial-intelligence-in-laymans-terms
  https://stopad.io/blog/artificial-intelligence-facts
  http://blog.aylien.com/10-machine-learning-terms-explained-in-simple/

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. You have correctly implemented the model tuning using the grid search. You have provided all the requirements from this part of the code for the model tuning.

Bonus:

- GridSearch is not the only technique available to us though! Another similar technique worth looking is RandomizedSearchCV
- Here I will leave some good articles about this subject:
  https://en.wikipedia.org/wiki/Hyperparameter_optimization
  https://machinelearningmastery.com/grid-search-hyperparameters-deep-learning-models-python-keras/
  https://www.quora.com/Machine-Learning-How-does-grid-search-work

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.

Great job.

You have done a great job indicating and ranking those features.

Suggestion:

- It's always a good idea to put some science behind of our opinions when it's possible. You can try to reinforce your arguments using statistics and show the correlations between then or show a sample where, for instance, you have higher occupations making more money than others.

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.

Awesome job. You have implemented correctly the feature importances from your chosen model.

Feature selection is also called variable selection or attribute selection.

It is the automatic selection of attributes in your data (such as columns in tabular data) that are most relevant to the predictive modeling problem you are working on.

feature selection... is the process of selecting a subset of relevant features for use in model construction

— Feature Selection, Wikipedia entry.

```
importances = model.featureimportances
```

Bonus:

- Here I will leave some good articles about the importance of feature selection:
https://www.kdnuggets.com/2017/06/practical-importance-feature-selection.html
http://www.simafore.com/blog/bid/61099/Reasons-why-feature-selection-is-i

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.

Great work.

Comparing those feature is really important and feature selection really matters when determining and using the supervised models of machine learning.

DOWNLOAD PROJECT

Rate this review