Finding Donors for CharityML

REVISÃO

REVISÃO DE CÓDIGO

HISTORY

Meets Specifications

Olá,

Ótimo projeto! Suas explicações estão muito completas. Continue com o bom trabalho

Atenciosamente,

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.

Ótimo! Apenas com algumas linhas você pode aplicar as transformações.

Para sua referência, você pode conferir este artigo que explica quando e por que usamos One Hot Encoding.

Algo a se notar aqui é que também podemos usar o Label Encoder como uma alternativa Multi Class. Label Encoder pode ser implementado como a seguir:

```
encoder = LabelEncoder() income = encoder.fit_transform(income_raw)
```

Sugestão: Este link fornece 7 estratégias de codificação diferentes. Codificação binária, é uma ótima opção para casos em que o número de categorias é alto.

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.

Muito bom trabalho ao mencionar algumas aplicações do mundo real, pontos fortes / fracos e raciocínio para sua escolha! Gostei da imagem também.

Aqui podem estar algumas ideias para pensar nos modelos a escolher:

- O poder preditivo do modelo
- O tempo de execução do modelo e como ele será dimensionado para muito mais dados
- A interpretabilidade do modelo
- Com que frequência precisaremos executar o modelo e / ou se ele suporta o aprendizado online.
- Distribuição da variável target
- Dados não lineares?
- Outliers?
- Dados faltando?

✓

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.

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.

Boa descrição do seu modelo. Isso seria útil para alguém que não esteja familiarizado com o aprendizado de máquina.

✓

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.

Ótimo uso do GridSearch aqui com todas as combinações e estados aleatórios.

Dica Pro: Com um conjunto de dados desequilibrado como este, uma ideia para garantir que os labels sejam divididos igualmente entre os conjuntos de validação seria usar o StratifiedShuffleSplit do sklearn: http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedShuffleSplit.html

```
from sklearn.model_selection import StratifiedShuffleSplit
cv = StratifiedShuffleSplit(...)
grid_obj = GridSearchCV(elf, parameters, scoring=scorer, cv=cv)
```

✓

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.

✓

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.

BAIXAR PROJETO

RETORNAR

Avalie esta revisão