Mohammed Saleh
MohammedSaleh@ieee.org

Finding Donors for *CharityML*

27th April 2020

OVERVIEW

For some of CharityML it’s important to understand the potential donors. Our goal with this implementation is to predict whether an individual makes more than $50,000. This sort of task can arise in a non-profit setting, where organizations survive on donations. Understanding an individual's income can help a non-profit better understand how large of a donation to request, or whether or not they should reach out to begin with.

PROBLEM STATEMENT

While it can be difficult to determine an individual's general income bracket directly from public sources, we aim that we can infer this value from other publically available features.

Our problem is a binary classification problem.

Also, what are the most effective features on ‘income’ that need our attention?

DATASETS AND INPUT

The dataset for this project originates from the UCI Machine Learning Repository. The dataset was donated by Ron Kohavi and Barry Becker, after being published in the article "Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid". You can find the article by Ron Kohavi online. The data we investigate here consists of small changes to the original dataset, such as removing the 'fnlwgt' feature and records with missing or ill-formatted entries.

EVALUATION METRICS

We will use the **Accuracy score** and **F-score** metrics since we don’t have a critical case for the False Negative or False Positive.

BENCHMARK MODEL

We didn't have any previous applied models. So, we will use an initial benchmark which predicts all of the data as 1 ("More than 50K$").

- Accuracy score: 24.78%
- F-score: 29.17%

Then, we will use another benchmark which is our model scores on the test data set before optimization.

- Accuracy score: 85.76%
- F-score: 72.46%

SOLUTION STATEMENT

After our investigation we noticed that the most important features which affect the total income are:

- Capital-gain
- Capital-loss
- Age
- Education level
- Hours of working per week

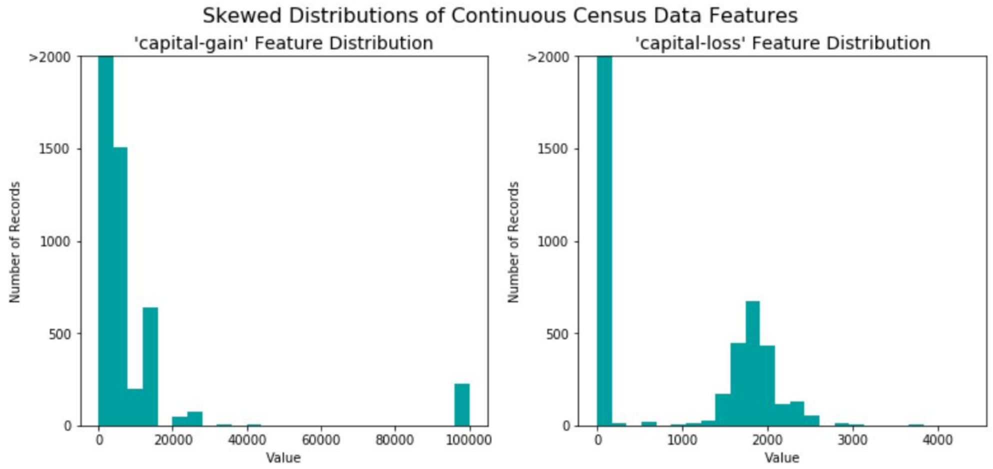
We choose the **AdaBoost** algorithm to predict if the total income is more or less than **50K$**.
Which achieved:

- Accuracy score: 86.64%
- F-score: 74.32%

PROJECT DESIGN

Data quality

1. The data are cleared from any missing values.
2. The total rows of data are 45222 with 11208 individuals whose income is greater than 50k$. So, our data is imbalanced.
3. There are skewed distributed features which are **Capital-gain** and **Capital-loss**.



4. The data values vary in different distributions.
5. There are many categorical features.

Data pre-processing

1. We applied a logarithmic transformation on the skewed features.
2. For numerical features we used MinMaxScaler for normalization.
3. We applied One-hot encoding for categorical features using DummyVariables.
4. Randomly, the data has been split into train and test datasets, using 20% of total data as a test set.

Model selection

We did an initial comparison between 3 models and chose the **AdaBoost** algorithm for our problem.

Model Tuning

Finally, we used the GridSearch algorithm to optimize the output and for the hyperparameter tuning.

Metric	naive predictor	Unoptimized Model	Optimized Model
Accuracy Score	24.78%	85.76%	86.64%
F-score	29.17%	72.46%	74.32%