Census Income Prediction

Faisal Aldhuwayhi - Faisal Alsumait - Abdullah Alsaleh
438102142 - 438104293 - 438105267

Introduction

The project aims to employ several supervised techniques to accurately predict individuals' income. The importance of this project lies in, for example, helping non-profit organizations evaluate their much-needed donation requests from different individuals.

Our approach is to use supervised machine learning techniques to train a model that can predict the income of an individual accurately based on a set of different features.

Task Definition

- The dataset that will be used is the **Census income dataset**, that was extracted from the machine learning repository (UCI), which contains about 32561 rows and 15 columns. The target variable in the data set is income level, which shows whether a person earns more than 50,000 per year or not based on 14 features containing information on age, education, education-num, gender, native-country, marital status, final weight, occupation, work classification, gender, race, hours-per-week, capital loss, and capital gain.

- **A classifier** is used, since the target variable has binary classes (income level greater than 50,000$ or not), which indicates a classification problem. This is because the target is having only values of 0 (<=50k $) and 1 (>50k $).

Algorithm Definition

- Multiple supervised machine learning algorithms have been used for the project:

1. **Logistic regression**, despite its name, is a linear model for classification rather than regression. In this model, the probabilities describing the possible outcomes of a single trial are modeled using a logistic function.
2. **Random forest** is a meta estimator that fits several decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.
3. **AdaBoost classifier** is a meta-estimator that begins by fitting a classifier on the original dataset and then fits additional copies of the classifier on the same dataset but where the weights of incorrectly classified instances are adjusted such that subsequent classifiers focus more on difficult cases.

Experimental Evaluation

- The data has been split into **training and testing** parts of the features and the label, with a test size of 20%, and with a random state to get the same randomness with the next runs.

- **Cross-validation** has been applied between different models to select the most suitable ones.

- **Random-Forest classifier** is found to have a better accuracy score compared to Logistic regression and AdaBoost classifier.

- Since there is an **imbalance in the classes** of the classification. This problem is solved through Oversampling, which is a technique used to modify unequal data classes to create balanced data sets.

Experimental Evaluation

- **Accuracy metric** has been chosen for the evaluation of the models. Which is the ratio between the number of correct predictions and the total number of predictions.

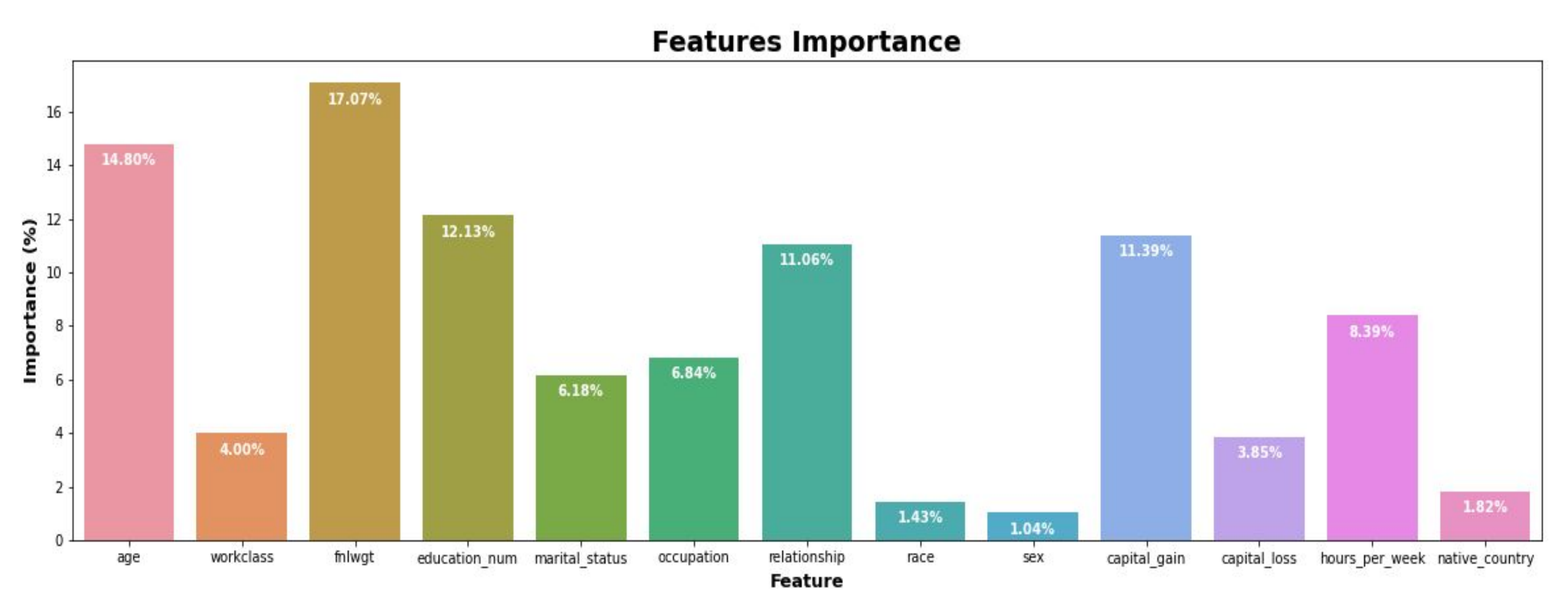
- **F1-score** has been used as one of the metrics in the experiment, it can be interpreted as a harmonic average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0.

- **Feature selection** has been applied by taking the most important features, using the ability of the model to show the features that have the most predictive power.

Results

Random forest Classifier	Accuracy score	93%
	F1-score	93%
AdaBoost Classifier	Accuracy score	86%
	F1-score	67%
Logistic Regression	Accuracy score	84%
	F1-score	65%

Results



Features Importance
Importance (%)
Feature
0
2
4
6
8
10
12
14
16
14.80%
4.00%
17.07%
12.13%
6.18%
6.84%
11.06%
1.43%
1.04%
11.39%
3.85%
8.39%
1.82%
age
workclass
fnlwgt
education_num
marital_status
occupation
relationship
race
sex
capital_gain
capital_loss
hours_per_week
native_country