Adult Salary Prediction By Using Classification

Ghanim AlGhanim

Omran Fallatah

Abstract

The goal of this project was to use classification to predict the salary bracket of adults(>$50k or <$50k). We worked with the data from https://www.census.gov/ (https://www.census.gov/) , leveraging feature Selection, feature engineering, dummy features, SMOTE. Then, we built Logistic Regression, KNN, Decision Tree and Random Forest models. We concluded by comparing between the accuracy of each model

Design

This project originates from the UCI machine learning repository. Classifying adult's salary accurately via machine learning models would enable employers to improve their job offers to new hires and give accurate salary bonuses to their employees. Also, it would enable employees to have a clear vision on where they lie with regards to their skills and experiences.

Data

Adults Salary dataset contains 48,000 data points and 15 features for each data point. A few feature highlights include Age, Gender, Work Class, Education and Marital Status. Our target feature is salary.

After cleaning the data, removing outliers, applying feature engineering, replacing NaN's with mode and dummy variables we ended up with 48,000 data points and 28 features.

Algorithms

Data manipulation and cleaning.

- Removed outliers.
- Mapped categorical f
  eatures into numerical.
- Replaced NaN values with mode.
- Applied dummy variables on Occupation feature.
- Dropped unnecessary columns.

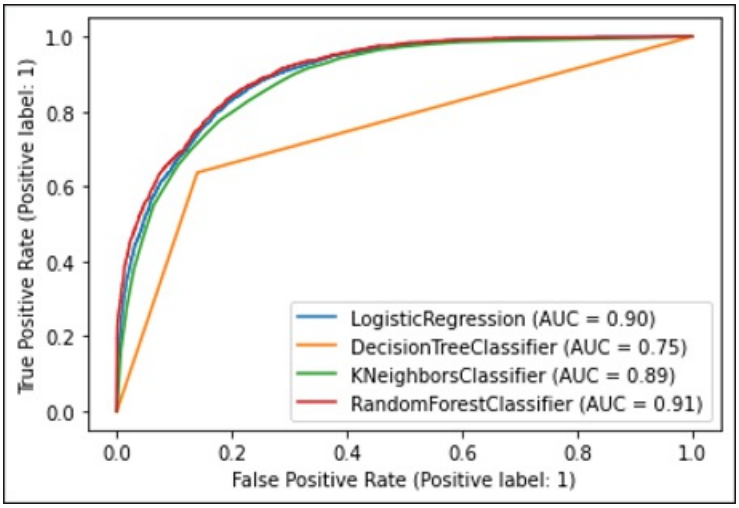
Models

Logistic regression, k-nearest neighbors, and random forest classifiers were used before settling on random forest as the model with strongest cross-validation performance.

Model Evaluation and Selection

We split into 80/20 train and test respectively. The training dataset has 39073 data points and the test dataset has 9769 data points after the test/train split. All scores reported below were calculated with 10-fold cross validation on the training portion only.

Algorithm	Accuracy	Precision	Recall	F-1 Score	ROC-AUC Score
Logistic Regression	0.7920	0.5566	0.8404	0.6697	0.90
Decision Tree	0.8138	0.5881	0.6390	0.6125	0.75
K-Nearest Neighbors	0.7957	0.5321	0.8293	0.6482	0.89
Random Forest	0.8560	0.5645	0.8468	0.6775	0.91



Tools

- Data manipulation and cleaning : Pandas , Numpy.
- Plotting : Seaborn, Plotly and Matplotlib.
- Modeling : Scikit-learn.

Communication

In addition to the slides and the visuals included in the presentation, we will submit our code and proposal.