US Census Income Level Predictor

Karan Gupta

Problem and Objective

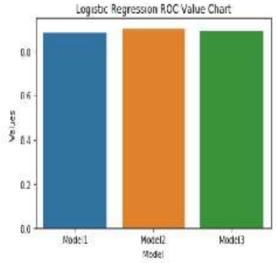
- The objective is to predict income level of US citizens from census data and bin it in two categories i.e. above 50k and below 50k
- The problem is to decide which features best help classify income level of citizens into those categories using classification algorithms

Data Description

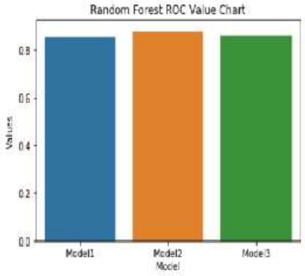
- The data is collected from UCI's ML repository and each citizen is described by 41 variables that affect his/her income level
- Total Rows: 199,524
- Number of Numerical Columns: 10
- Number of Categorical Columns: 24

Prediction Performance

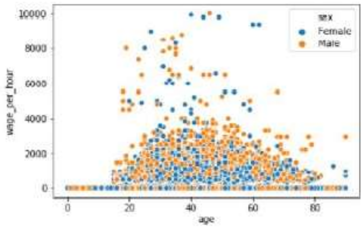
Logistic Regression

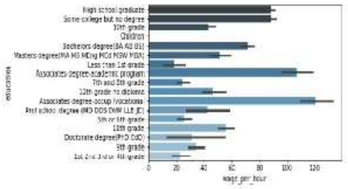


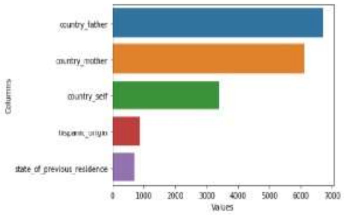
Random Forest

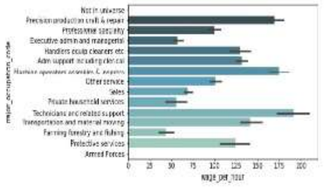


Data Exploration









Machine Learning Models

Model	Features	Algorithm	ROC Value
Model 1	Class, Education, Wage per hour, Industry Code	1. Logistic Regression 2. Random Forest	0.890840 0.863477
Model 2	Race, Sex, Employment, Tax, Household, Citizenship, Capital Gained, Capital Lost, Stock Dividends, Business, Veterans_benefit , Weeks worked annually, Age	1. Logistic Regression 2. Random Forest	0.905607 0.884434
Model 3	Education, Wage per hour, Industry Code, Occupation Code	1. Logistic Regression 2. Random Forest	0.900420 0.866227

Inferences

- Based on the performance of above models we identify that logistic regression has a better performance than Random Forest.
- Additionally, Model 2 has a good set of input features which help in classifying income level with a better prediction accuracy.

Conclusions

- For predicting incomes of citizens, we found that their financial features such as capital gains, capital losses, stock dividends and several others have a higher feature importance than others
- We can utilize this project in successfully classifying incomes given a set of features to optimally describe a user

Data Source: http://archive.ics.uci.edu/ml/machine-learning-databases/census-income-mld/census-income.data.gz