Classification Model to Predict Annual Income

Presented By

Group 7
Jing Tang
Panthea Saffarzadeh
David Graham
Maxim Smetanin
Ramila Mudarth
Sivi Rakaj

Abstract

Evaluate classification models to determine the likelihood that an individual would earn an income of $50,000 a year annually more or less given certain demographics and social features

Overview

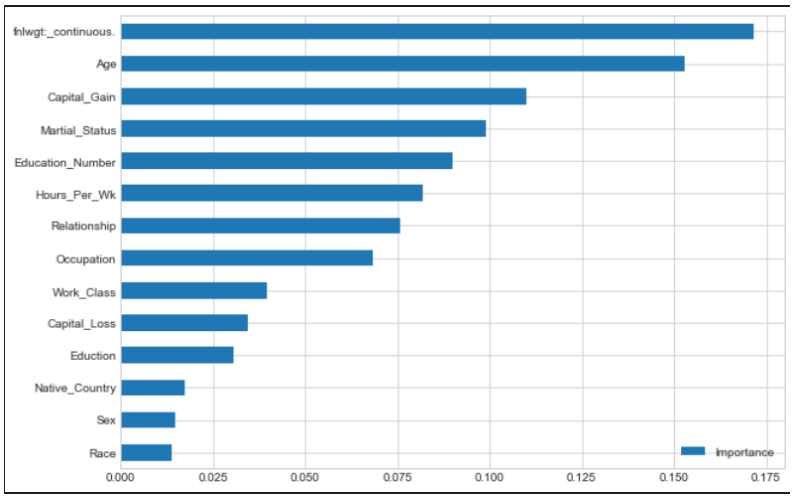
There is a hypothesis that demographics and socioeconomics status determines an individual's annual income. Using US census data from 1994 we attempted to derive various predictive modeling to assess the possibility of an individual's income, given their social status and demographics, to have the ability to earn over or under 50K. Through feature importance analysis we can determine if features like education, gender, age or immigration status influence an individual's income. The predictive modeling can then be applied within a community so as to focus on these features to stabilize socioeconomic disparity.

Data Exploration

The data set used is from a 1994 Census dataset which has 32,561 entries with a mixture of 70% categorical and 30% numerical data types. The Target Data is the income level i.e .>50 K or <50K. Thus making it a binary classification. The data set had no null values except some irregular data "?" in Work class, Occupation and Native Country. We replaced the "?" values with "Other". We re-engineered some features to broaden categorical classification. With Feature importance analysis, we determined the top five features - Final Weight(fnlwgt) , Capital Gain, Marital Status Education, and Age.

Data Analysis

Some of the high level observations were that it was a mildly imbalanced dataset of our target variable - Income Type. Frequency of distribution on the categorical variables inferred that Males- White'/'Asian-pac-islander' has a greater likelihood of earning >50K. Education data exhibited some disparity and it was inconclusive. Married Individuals had a great likelihood of earning >50k. With the numerical variables, 75% of Individuals in the >50K classification are within the age group of 36 -51 years, with a median at 44. 75% of Individuals in the <50K classification are with the the age range of 25 -46 with median at 34.



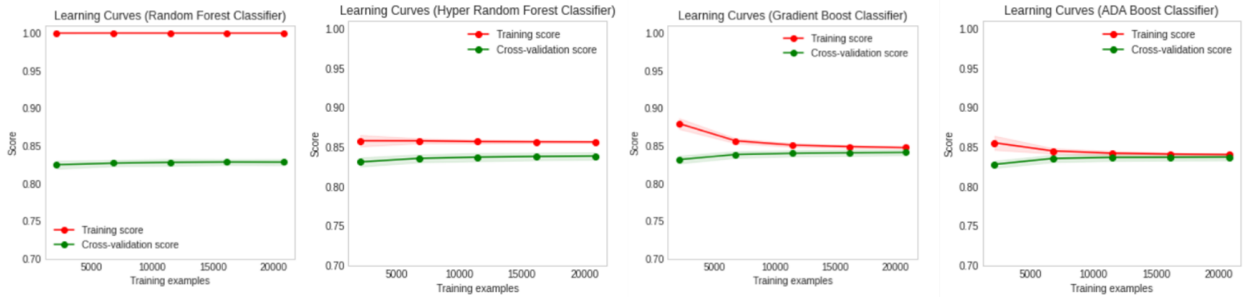
Features importance was calculated using Random Forest Classifier. Race, sex and native country have least contribution in income level.

Data Transformation Pipeline

Prior to modeling, the numeric and categorical data were scaled and transformed. Categorical Variables were transformed using OneHot Encoder. Numerical Variables were scaled using MinMax Scaler. The data was split into 80% Train and 20% Test prior to Modeling.

Data Modeling

The table below, provides the classification report for the different modeling classifiers we subjected our data through . The learning curves were run for Random Forest Classifier - Before and After Hyperparameter Tuning, Gradient Boost and ADA Boost.



Classifier Type	Accuracy Score	Target	Precision	Recall	F1 Score
Random Forest	83.63%	Less 50,000$:	0.86	0.88	0.89
		More 50,000$:	0.6	0.54	0.61
Random Forest Tuned	84.12%	Less 50,000$:	0.87	0.94	0.9
		More 50,000$:	0.73	0.55	0.63
Decision Tree Tunned	81.37%	Less 50,000$:	0.83	0.94	0.88
		More 50,000$:	0.69	0.41	0.52
ADA Boost	83.95%	Less 50,000$:	0.87	0.93	0.9
		More 50,000$:	0.71	0.57	0.63
Gradient Classifier	83.96%	Less 50,000$:	0.87	0.93	0.9
		More 50,000$:	0.71	0.56	0.63
Ensemble Voting	83.54%	Less 50,000$:	0.86	0.94	0.9
		More 50,000$:	0.72	0.52	0.61
Tensor Flow	83.53%	Less 50,000$:	0.88	0.91	0.89
		More 50,000$:	0.68	0.61	0.62
Wide & Deep Neural Network	83.45%	Less 50,000$:	0.87	0.92	0.89
		More 50,000$:	0.7	0.59	0.64

Conclusion

All the models were able to run with good performance with a 80 -20 split. The learning curves showcased that increase in observations would not improve the model's accuracy score but rather focus on fine tuning the parameters for the modeling. It can be inferred that the Random Forest Classifier - Tuned Hyperparameters, Gradient Classifier and ADABoost had the most accurate score for our Classification problem. The Decision Classifier with Tuning had the lowest performance. Some models have an F1 score of 0.9 so be confident that our models are efficient in prediction <50K.