Census Income dataset: Using multiple machine learning models

Analyzing the Data:

This data was extracted from the [1994 Census bureau database](http://www.census.gov/en.html). **The prediction task is to determine whether a person makes over $50K a year or not.**

Using the python language and several visualizations, I have attempted to fit 4 machine learning models and find the best model to describe the data.

Data Preprocessing:

The dataset contained null values, both numerical and categorical values. The categorical values were both nominal and ordinal. The data had redundant columns as well.

Since the missing values were represented by ‘?’ , they were replaced by NAN values and removed after detection. The dependent column, ‘income’ which is to be predicted has been replaced with 0 and 1 and hence convert the problem to a dichotomous classification problem. There was one redundant column, ‘education.num’ which was an ordinal representation of ‘education’, which is removed above.

Now that unnecessary data points and redundant attributes have been removed, it is necessary to select the set of attributes really contributing to the prediction of the income.

Correlation:

To check the **correlation between a binary variable and continuous variables**, the **point biserial correlation** has been used. After appropriate application of the test, ‘fnlwgt’ has been dropped which showed negative correlation.

For feature selection, all the numerical columns are selected except ‘fnlwgt’. For categorical variables, chi-square estimate is used. Chi-square estimate is used to measure the correlation between 2 categorical variables.

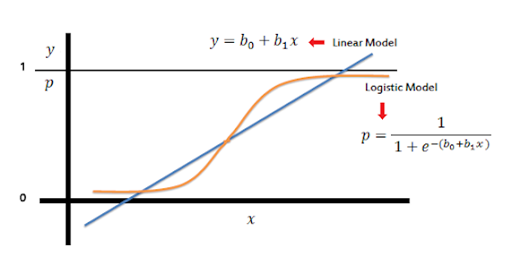
First, the categorical variables are encoded or rather dummies are generated and the numerical values are **normalized** to be between [0,1]. **It’s simply a case of getting all your data on the same scale: if the scales for different features are wildly different, this can have a knock-on effect on your ability to learn** (depending on what methods you’re using to do it). Ensuring standardized feature values implicitly weights all features equally in their representation.

Model:

As mentioned above, 4 models are shown below. The training and testing is divided in 80–20 for logistic and naive bayes whereas 70–30 for decision tree and random forest.

Logistic Regression:

The foremost model to predict a dichotomous variable is logistic regression. **The logistic function is a sigmoid function, which takes any real input t, and outputs a value between zero and one.** It gives the probability.

****

After fitting the model, we find the model accuracy. I generated the confusion matrix and it does somewhat good. We will compare all the models in the end.

Naïve Bayes Classifier:

A **naive Bayes classifier** assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature, given the class variable. Basically, it’s “**naive**” because it makes **assumptions that may or may not turn out to be correct.**

Decision Tree:

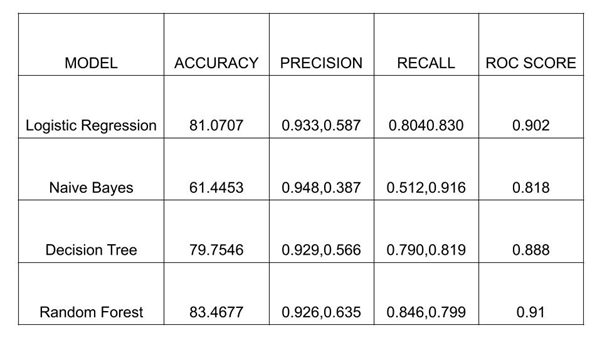
A **decision tree is a branched flowchart** showing multiple pathways for potential decisions and outcomes.**The tree starts with what is called a decision node,** which signifies that a decision must be made. From the decision node, a branch is created for each of the alternative choices under consideration.

Random Forest:

Random Forests are a combination of tree predictors where each tree depends on the values of a random vector sampled independently with the same distribution for all trees in the forest. **The basic principle is that a group of “weak learners” can come together to form a “strong learner”.**

I have used one model accuracy measure to form a comparative study between all models. To construct the ROC curve the following code is thus.

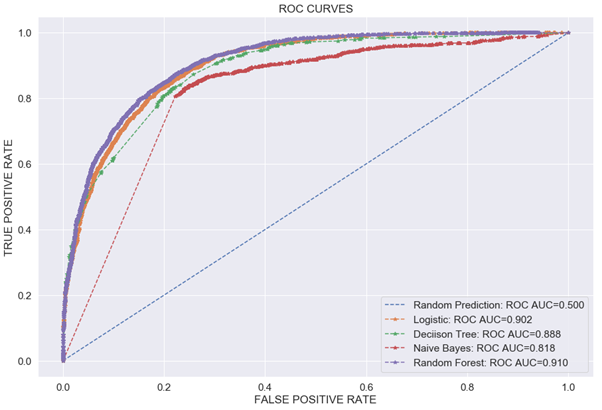
A comparative study of the above models with respect to accuracy, precision, recall, ROC score is computed together for better decision.



From the table above, random forest gives the best accuracy and ROC score.

All the ROC curves is shown below.

AUC – ROC Curve:



**Random forest covers the maximum area and hence is a better model.** I have not tried neural networks on this problem as there were only 30K plus data points I felt it would overfit the data. To further improve, more complex ensemble methods can be used. Also, according to Ockham’s Razor “the simplest explanation is most likely the right one”