Jacob Maurer

CS-496 ML Final Report

**Data:**

This report concerns a dataset consisting of 48,842 samples. The features for each sample are:

Listing of attributes:   
  
>50K, <=50K.   
  
age: continuous.   
workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.   
fnlwgt: continuous.   
education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.   
education-num: continuous.   
marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.   
occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.   
relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.   
race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.   
sex: Female, Male.   
capital-gain: continuous.   
capital-loss: continuous.   
hours-per-week: continuous.   
native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

The data is classified by whether a person makes above or below $50,000. The immediate problem with this data is that not all the features are numerical or directly comparable with each other. I quantified some of the features such as the native-country by finding the average incomes of people with that characteristic. For example, the average income of people living in the United States is $59,160.

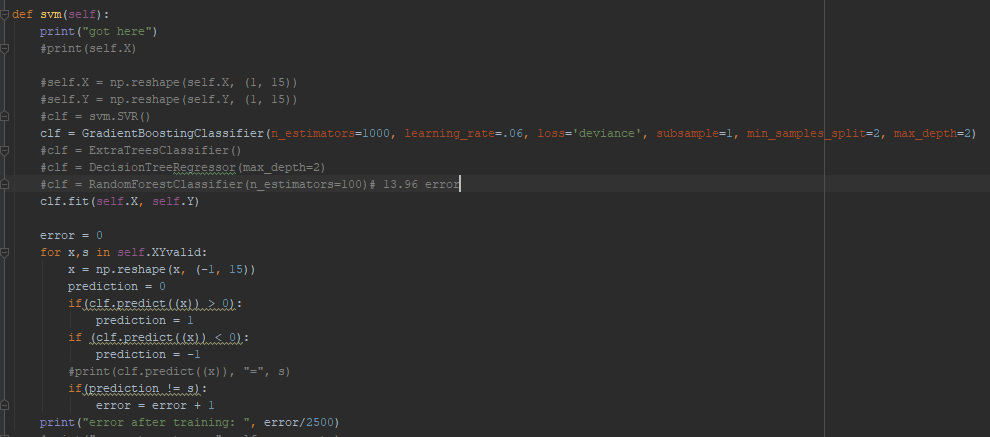
The feature vector is as follow:

x = np.array([1, x1, x2, x3, x4, x5, x6, x7, x8, x9, x10, x11, x12, x13, x14])

With the classification being ‘-1’ = ‘<50K’ and ‘1’ = ‘>=50K’

**Model:**

Here is the main method that trains and the model and validates the error:



I trained on 32,561 models and tested on 8140 samples. I received an error of .1271. I received an error of .1232 after training the model and testing on the 8140 samples being used as validation.

**Training data: 27,61 samples**

**Testing data: 2500 samples**

**Validation: 2500 samples**

**Accurary: 87.68%**

The model and specified parameters:

GradientBoostingClassifier(n\_estimators=1000, learning\_rate=.06, loss='deviance', subsample=1, min\_samples\_split=2, max\_depth=2)

**Other Models:**

I had attempted training on several other models, such as:

-Support vector machine with linear classification: **accuracy of 80.88%**

-Support vector regression: **accuracy of 75.76%**

-Extra Trees Classifier: **accuracy of 77.68%**

-Decision Tree Regressor: **accuracy of 83.4%**

-Random Forest Classifier: **accuracy of 86.0%**

**Parameters:**

**n-estimators:** The number of boosting stages to perform. Raising the number of n-estimators usually doesn’t cause over-fitting so a large number resulted in better performance. The error seemed to plateau at 1000 and above.

**Learning rate:** learning rate shrinks the contribution of each regression tree in composing the model. Seemed to perform best at n-estimators=1000 at learning\_rate=.06.

**Loss:** The loss in this case refers to the loss function. Can either be deviant or exponential. Deviant performed better in this case.

**Subsample:** a subsample less than 1 will lead to a reduction of variance and an increase in bias.

**Min\_samples\_split:** The number of samples required to split an internal node. Performed best at 2

**Max\_depth:** maximum depth of the individual regression estimators. Performed best at 2.

**Comparison:**

<https://towardsdatascience.com/logistic-regression-classifier-on-census-income-data-e1dbef0b5738>

accuracy: 85.0% using logistic regression classifier

<https://www.kaggle.com/overload10/income-prediction-on-uci-adult-dataset>

accuracy: 82.24% using random forest classifier

<https://yanhan.github.io/posts/2017-02-15-analysis-of-the-adult-data-set-from-uci-machine-learning-repository.ipynb.html>

accuracy: 79.04% using logistic regression

<https://pdfs.semanticscholar.org/3dd5/e9f335511efbb81d65f1d6d4995019f8b5fd.pdf>

accuracy: 86.29% using gradient boosting classifier

<https://www.google.com/search?q=NBTree&rlz=1C1CHBF_enUS811US811&oq=NBTree&aqs=chrome..69i57j0l5.4727j0j4&sourceid=chrome&ie=UTF-8>

Accuracy 85.93% using NBTree

My model achieved an accuracy of 87.68%, beating out the other models in the articles mentioned above.