**BDA564 End-to-end Big Data Analytics**

**ASSIGNMENT 1**

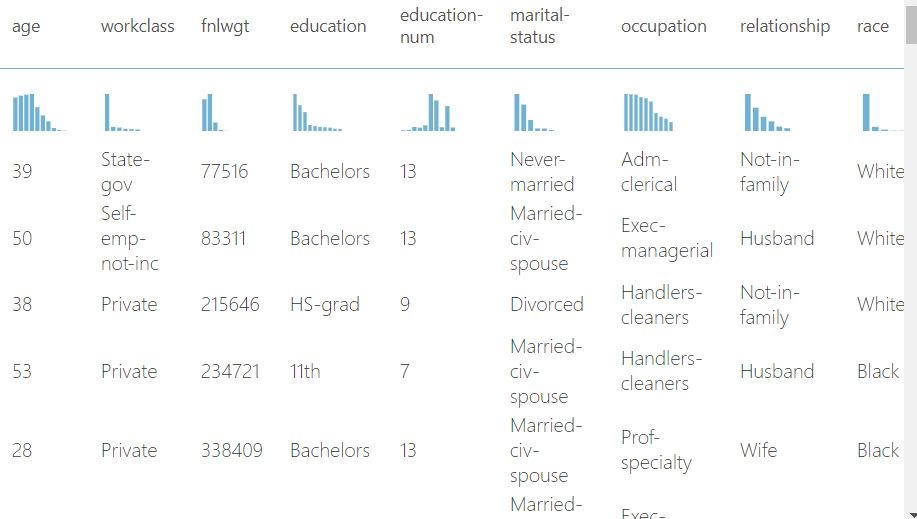
**Part-A:** Please explain the dataset in terms of its general importance, size, features,

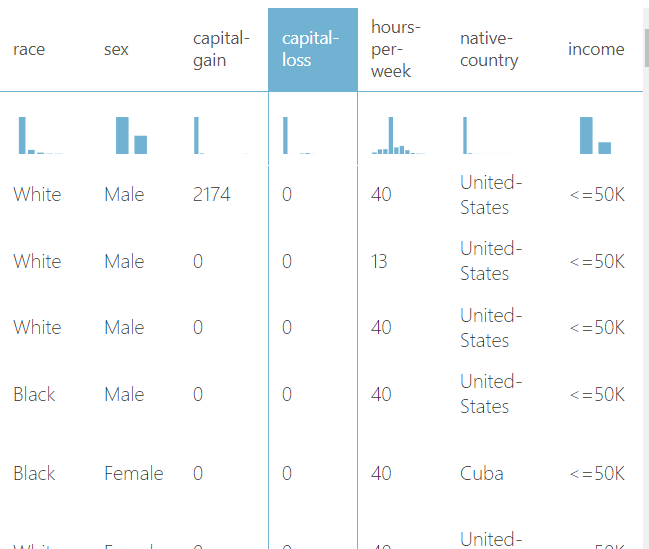
and main use. (10 points)

- There are 15 columns and 32561 rows in dataset.

- Prediction task is to determine whether a person makes over 50K a year.

- Due to visualization of dataset:





**Attribute Information:**

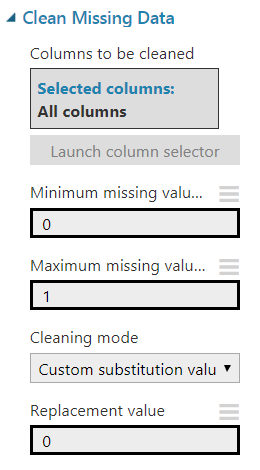
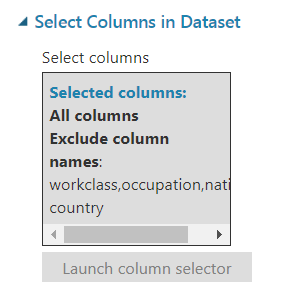
Credit by <https://archive.ics.uci.edu/ml/datasets/adult>

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.

**Part-B:** Please explain the preprocessing (data preparation) stage briefly. What

happens in the *clean missing data* and *select column in dataset*. Why are these steps

relevant for the data processing? (10 points)

On clean missing data step; all missing values of all columns are replaced with 0. After that, select columns in dataset step workclass, occupation,native-country columns are excluded. Because these columns had big amount of missing values before cleaning data step. So, after cleaning, all missing data become zero and useless.

**Part-C:** Please explain the algorithm (Two-class Boosted Decision Tree) used in this

model very briefly. Why do you think that this algorithm might have been chosen? (10

points)

The aim of the model is trying to guess if a person’s income is over 50K or not. So, this is a binary classification problem. Because of the problem structure, two-class boosted decision tree is appropriate.

Due to azure machine learning studio’s website (<https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/two-class-boosted-decision-tree> ), two-class boosted decision tree algorithm is an ensemble learning method in which the second tree corrects for the errors of the first tree, the third tree corrects for the errors of the first and second trees, and so forth. Predictions are based on the entire ensemble of trees together that makes the prediction.

**Implementation details**

1. Start with an empty ensemble of weak learners.
2. For each training example, get the current output of the ensemble. This is the sum of the outputs of all weak learners in the ensemble.
3. Calculate the gradient of the loss function for each example.

\* In a binary classification model, the log-loss is used, much like in logistic regression.

1. Use the examples to fit a **weak learner**, using the gradient just defined as the target function.
2. Add that weak learner to the ensemble with a strength indicated by the learning rate, and if desired, go to Step 2.
3. The tree-building algorithm greedily selects the feature and threshold for which a split minimizes the squared loss with regard to the gradient calculated in Step 3. The selection of the split is subject to a minimum number of training examples per leaf.

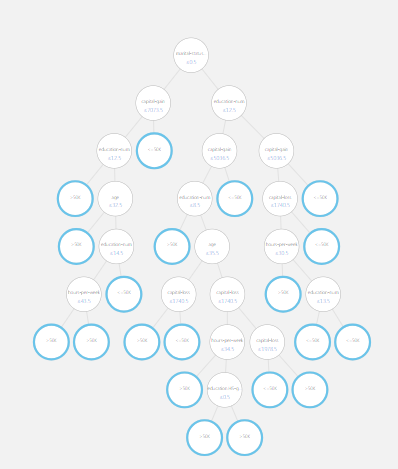
The algorithm repeatedly splits until it reaches the maximum number of leaves, or until no valid split is available.

The output of the algorithm is an untrained binary classification model.

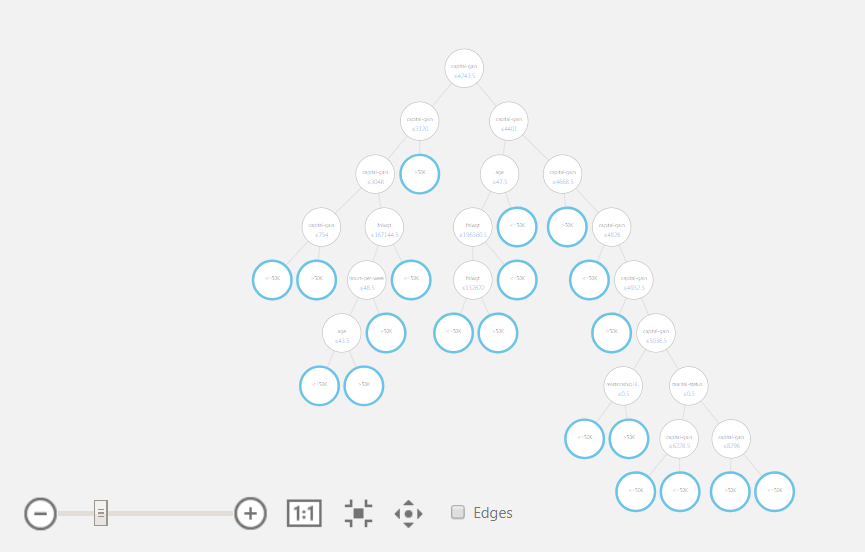
**Part-D:** Please comment on the trained model. What is the output? In the visualization part of train model, what are those trees? (10 points)

Boosted algorithm constructs 100 trees due to output of train model node.

The first tree is as follow:



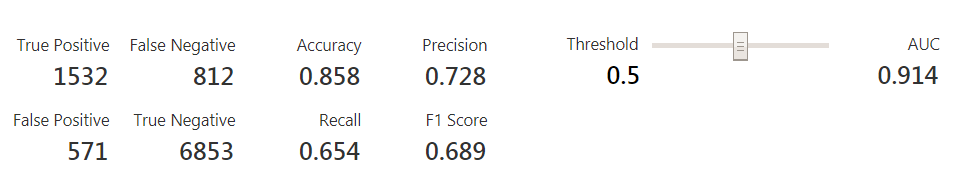
The final tree is as follow:



**Part-E:** What are your comments related to the outputs of score model and evaluate

model? You need to refer to the confusion matrix, accuracy, precision, recall. (10

points)



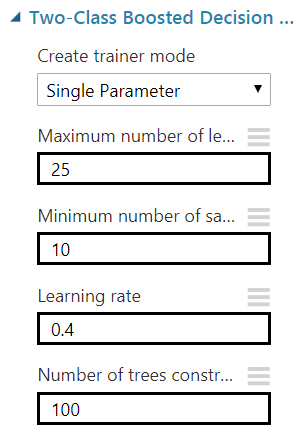
When threshold is 0.5; Accuracy is 0.858 which is reasonable for this dataset and AUC is, really high, 0.914. TP and TN values are also high, while precision is 0.72. Due to these metrics, there is no overfitting or underfitting situation and the model is acceptable and work well on this dataset.

**Part-F:** Please construct a parallel pipeline beginning from split module with the same

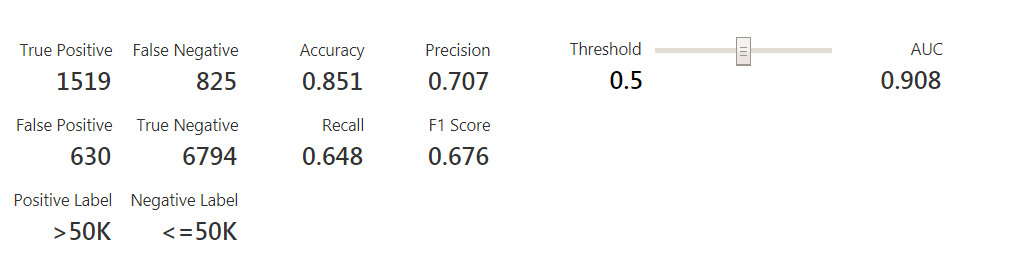
algorithm and test different parameters including maximum number of leaves, learning

rate, and number of trees constructed. (10 points)

Changed model:



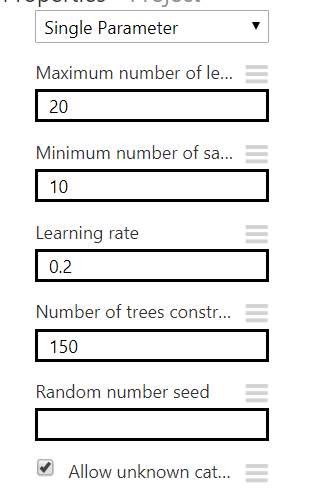
Output:

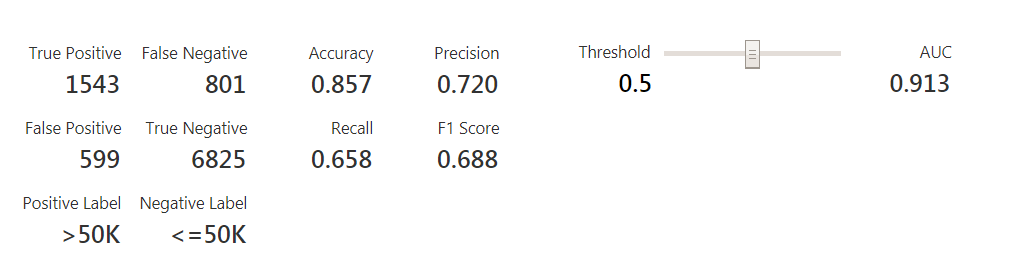


While learning rate rises 0.4 and max number of leaves increases 25; the output is worse than default values.

Because based on increase in learning rate means faster decisions.

Changed model 2:





When number of trees constructed increases 150, the results does not change hugely. So, 100 is a good option for this variable. Does not need any extra trial after 100 and this solution can improve execution time.

**Part-G:** Please suggest an alternative algorithm and provide a third pipeline for its

execution. You are free to choose any algorithm but it should be a classification

algorithm for two classes. (10 points)

Two-Class Decision Forest is an alternative algorithm for this problem. It also creates a two-class classification model.

