1) Prepare a classification model using Naive Bayes

for salary data

Data Description:

age -- age of a person

workclass -- A work class is a grouping of work

education -- Education of an individuals

maritalstatus -- Marital status of an individulas

occupation -- occupation of an individuals

relationship --

race -- Race of an Individual

sex -- Gender of an Individual

capitalgain -- profit received from the sale of an investment

capitalloss -- A decrease in the value of a capital asset

hoursperweek -- number of hours work per week

native -- Native of an individual

Salary -- salary of an individual

**Soln/Inference:**

# Exploratory data analysis

* We can see that there are 30161 instances and 14 attributes in the training data set.
* We can see that there are 15060 instances and 14 attributes in the test data set.

# Explore categorical variables

* There are 9 categorical variables.
* The categorical variables are given by workclass, education, maritalstatus, occupation, relationship, race, sex, native and Salary.
* Salary is the target variable.

# Explore Numerical Variables

* Summary of numerical variables
* There are 5 numerical variables.
* These are given by age, educationno, capitalgain, capitalloss and hoursperweek. All of the numerical variables are of discrete data type.

# Encode categorical variables

* We can see that from the initial 14 columns, we now have 102 columns.
* We now have training and testing set ready for model building. Before that, we should map all the feature variables onto the same scale. It is called feature scaling.

# Feature Scaling

* We now have X\_train dataset ready to be fed into the Gaussian Naive Bayes classifier.

# Check accuracy score

* Model accuracy score: 0.7995
* Here, y\_test are the true class labels and y\_pred are the predicted class labels in the test-set.

# Check for overfitting and underfitting

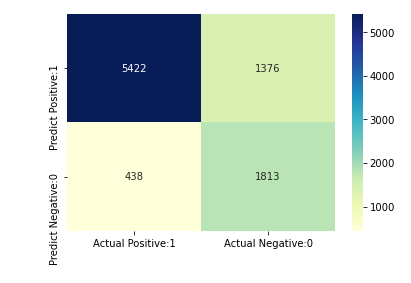
* Training set score: 0.8023
* Test set score: 0.7995
* The training-set accuracy score is 0.8023 while the test-set accuracy to be 0.7995. These two values are quite comparable. So, there is no sign of overfitting

# Compare model accuracy with null accuracy

* Null accuracy score: 0.7582
* We can see that our model accuracy score is 0.8023 but null accuracy score is 0.7582.
* So, we can conclude that our Gaussian Naive Bayes Classification model is doing a very good job in predicting the class labels.

# Confusion matrix

* True Positives(TP) = 5422
* True Negatives(TN) = 1813
* False Positives(FP) = 1376
* False Negatives(FN) = 438



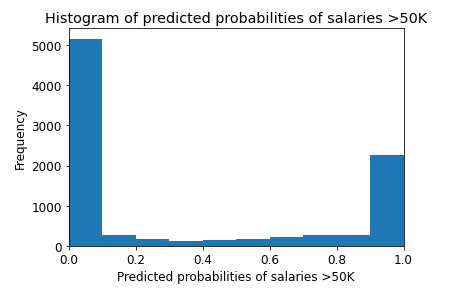
# Calculate class probabilities

* **Observations**
* In each row, the numbers sum to 1.
* There are 2 columns which correspond to 2 classes - <=50K and >50K.

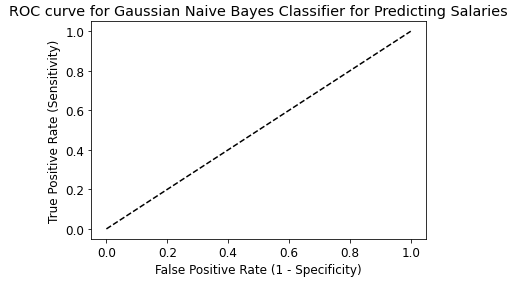
\* Class 0 => <=50K - Class that a person makes less than equal to 50K.

\* Class 1 => >50K - Class that a person makes more than 50K.

* Importance of predicted probabilities
* We can rank the observations by probability of whether a person makes less than or equal to 50K or more than 50K.
* predict\_probabillity process
* Predicts the probabilities
* Choose the class with the highest probability
* Classification threshold level
* There is a classification threshold level of 0.5.
* Class 0 => <=50K - probability of salary less than or equal to 50K is predicted if probability < 0.5.
* Class 1 => >50K - probability of salary more than 50K is predicted if probability > 0.5



# ROC - AUC



* ROC AUC : 0.8902

# Interpretation

* Cross validated ROC AUC : 0.8923

# k-Fold Cross Validation

* Average cross-validation score: 0.8018