ADA BOOSTING EXPERIMENT REPORT

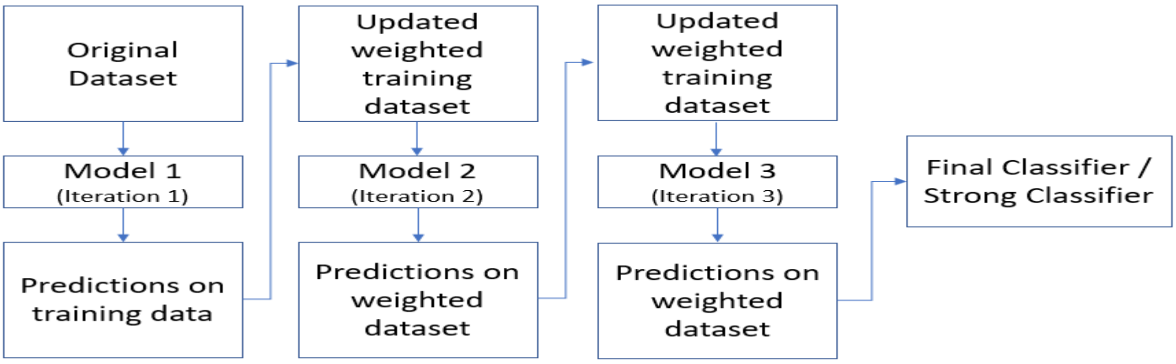
GANESHREDDY ANNAPAREDDY

50442295

ganeshre@buffalo.edu

1 ADA BOOSTING

Adaptive Boosting is an ensemble learning technique that employs an iterative process to improve weak classifiers by learning from their errors. With AdaBoost, decision trees with one level are the most often employed algorithm. Another name for these trees is Decision Stumps.


Original Dataset
Model 1 (Iteration 1)
Predictions on training data
Updated weighted training dataset
Model 2 (Iteration 2)
Predictions on weighted dataset
Updated weighted training dataset
Model 3 (Iteration 3)
Predictions on weighted dataset
Final Classifier / Strong Classifier

"The above given figure shows us an overview on how the ada boosting works"

Each iteration of the process involves training the data sample and adjusting the classifier weights to ensure accurate predictions of odd observations. In this instance, a basic classifier that accepts weights from the training set is a level Decision Tree.

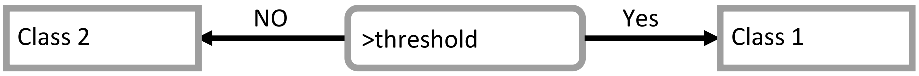
Following are the steps for ADA Boosting Algorithm:

- STEP A: A weight will be given to each data point. All of the weights are equal at first. The weighting formula is as follows:

$$w(x_i , y_i) = 1/N, i=1,2,3,.....n$$

where N is the number of data points and w is the weight.

- STEP B: Utilizing the weighted samples and the training data, a weak classifier (decision stump) is created. For each characteristic, a decision stump is made, and the Gini Index of each tree is calculated after that. We shall start with the tree with the lowest Gini Index.



- STEP C: In this phase, the performance of the chosen stump is assessed, and its weight is assigned. The "Amount of Say" is used to evaluate the performance. The equation is as follows:

  $$(1/2)\log((1\text{-Total error})/(\text{Total error}))$$

  Where Total error is the total sample weights for the incorrectly categorized data points. In order to change the weights for the subsequent model, it is required to evaluate the stump performance. The next model will also produce the same results if the weights are not updated.

- STEP D: Correct forecasts receive less weight while incorrect guesses receive more weight. This will enable the following model to concentrate more on identifying incorrect samples. The weights are modified in this phase using the following formula. Positive feedback will be given for incorrectly categorised data.

  $$\text{NEW SAMPLE WEIGHT} = \text{SAMPLE WEIGHT} * e^{\text{amount of say}}$$

  The amount of say will be negative for correctly classified data

  $$\text{NEW SAMPLE WEIGHT} = \text{SAMPLE WEIGHT} * e^{-\text{amount of say}}$$

- STEP E: The weights now need to be normalized by dividing each weight by the sum of weights after they have been updated. Thus, a new dataset is created. Create a fresh dataset and check to see whether the mistakes have decreased. For this, we will delete the "sample weights" and "new sample weights" columns, and then we will split our data points into buckets based on the "new sample weights."

- STEP F: The sample that has that value within its bounds is chosen. A random integer between 0 and 1 is chosen. A copy of the fresh sample is added to the fresh dataset.

- STEP G: All the aforementioned steps are repeated using the new dataset.

  Repeating the aforementioned processes will result in a low training error.

2 EXPERIMENT

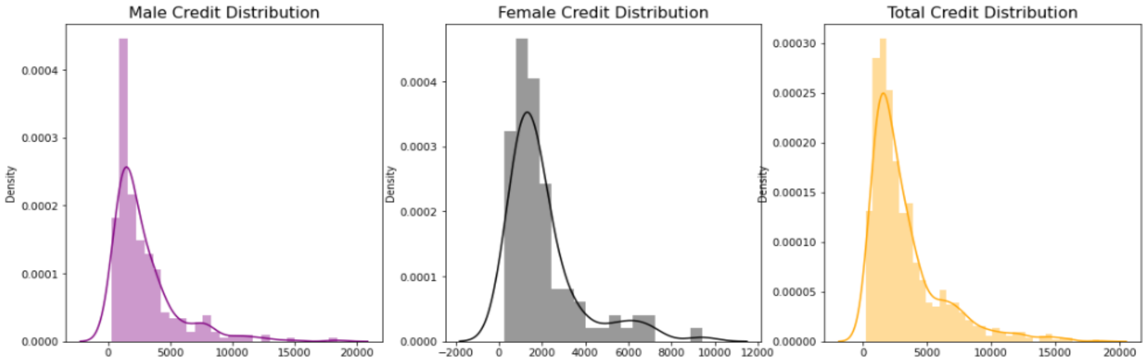
Dataset :

For the random forest experiment I have used the German credit data from Kaggle. It has the data which consists of Gender, age and various other things in German language . The dataset has The 1000 columns and 20 rows excluding the status and I have translated what ever is in German to English. Which are "laufkont":"status","laufzeit":"duration","moral":"credit_history","verw":"purpose","hoehe":"amount","spar kont":"savings","beszeit":"employment_duration","rate":"installment_rate","famges":"personal_status_sex", "buerge":"other_debtors","wohnzeit":"present_residence","verm":"property","alter":"age","weitkred":"othe r_installment_plans","wohn":"housing","bishkred":"number_credits","beruf":"job","pers":"people_liable","t elef":"telephone","gastarb":"foreign_worker","kredit":"credit_risk”.

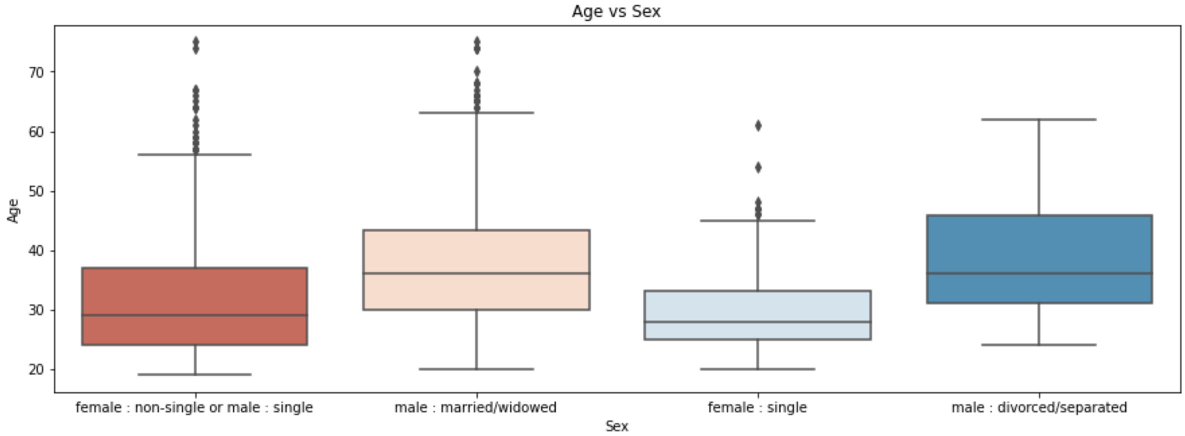
Column	Description
status	1 : no checking account2 : . < 0 DM 3 : 0<= . < 200 DM 4 : .>= 200 DM / salary
duration	Tells us what the duration was
credit_history	Tells us what kind of credit history was there
Purpose	Explains about the purpose
Amount	Shows us what the amount was
Savings	Shows us how much they have in the savings
Employment duration	Gives an idea on how long were they employed if not are they unemployed
Instalment rate	Tells about the instalment rates
Personal sex status	Male or female
Other debtors	Are there any other people supporting them to be the debtors
Present residence	Where they are staying
Property	What kind of property do they have
Age	Briefs about the ages
Other instalment plans	If they have stores or not.
Housing	If they own the house or they are tenants
Number credits	Explains how many credits they have
Job	What kind of job they have or they don't have any
People liable	How many people are liable
Telephone	If they have a telephone or not
Foreign worker	If they are a foreign worker or are they German
Credit risk	If they are under any credit risk or not which is showed by 0's and 1's

3 OBSERVATION

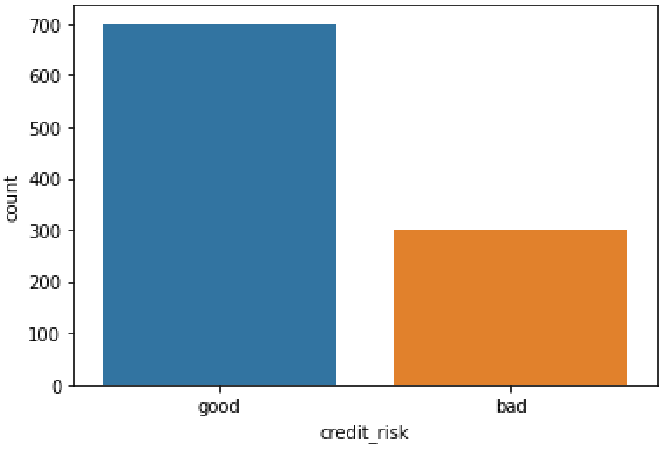
From the analysis that I have done first I got the bar graphs which shows us how the credit distribution is in between males, females and overall total credit distribution . As the size of the dataset isn't small, I am not removing any columns or rows from the dataset. Below figure shows us on how the distribution has



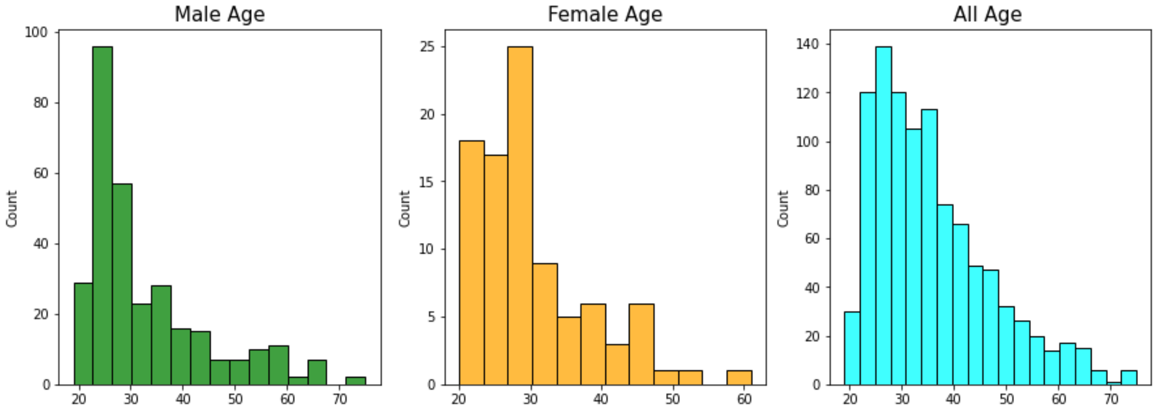
From the above graph we can see that the male's have taken more amount that that of females, males graph is around 20,000 where as the females is around 12,000 maximum. But the overall amount is around 20,000 because there are males and females who are married, divorced and single as well and their amount is also weighing in the graph. The below graph shows us that .



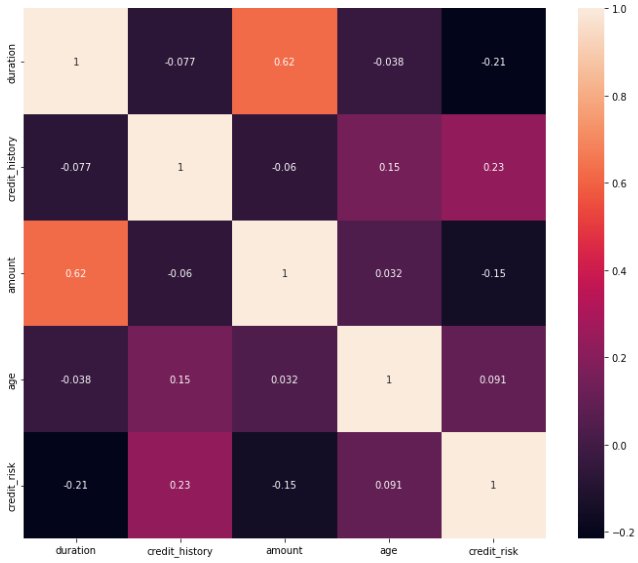
***Risk*:** The graph on the right hand side shows gives us an overview of the credit risk which explains about if the person will be in a credit risk or not to buy anything. Here I got the result as 70 who come under a good category which directly falls under risk, while the rest 30 percent falls on the bad category which doesn't fall under the risk factor for the banks to determine each individual on their ratings for a loan .

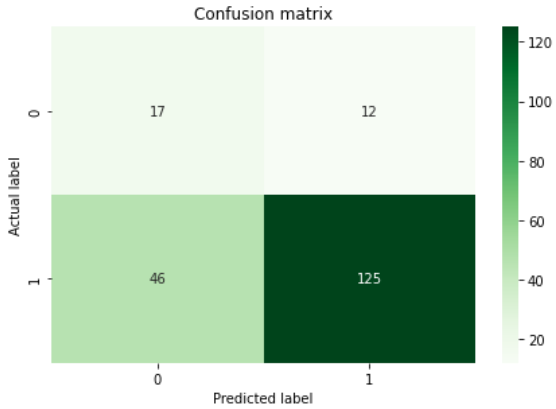


***Age Dependency*:** The below graphs gives us a clear vision on the age and gender dependency. Here most of the male's are more likely to take out a loan in their mid twenties, where as the females are more likely to take more loans near to their thirties. In the all age group people are more likely to take a loan in their twenties to their forties and as the age goes by the percent of people taking out a loan goes down for all age groups and genders as they might get financially stable by having job's or their business.

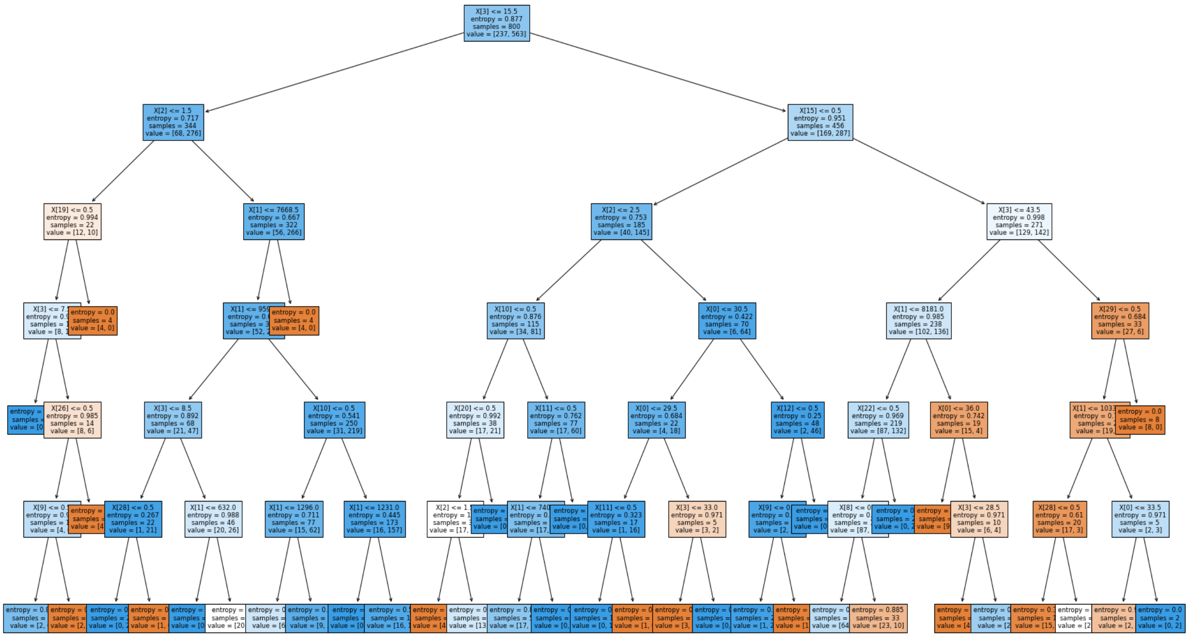


It can be observed that a person with more years of work experience are at less risk to the bank, but most people in the records aren't having a decent amount of work experience. About half of people who have little checking account are considered as bad rating in risk. Other aspects that were categorical, such Purpose, Housing, Job, and Sex, were transformed to categorical variables and were properly encoded. From the corelation matrix we can see that the lesser the duration the credit risk also less, the grater the duration the greater the risk is.

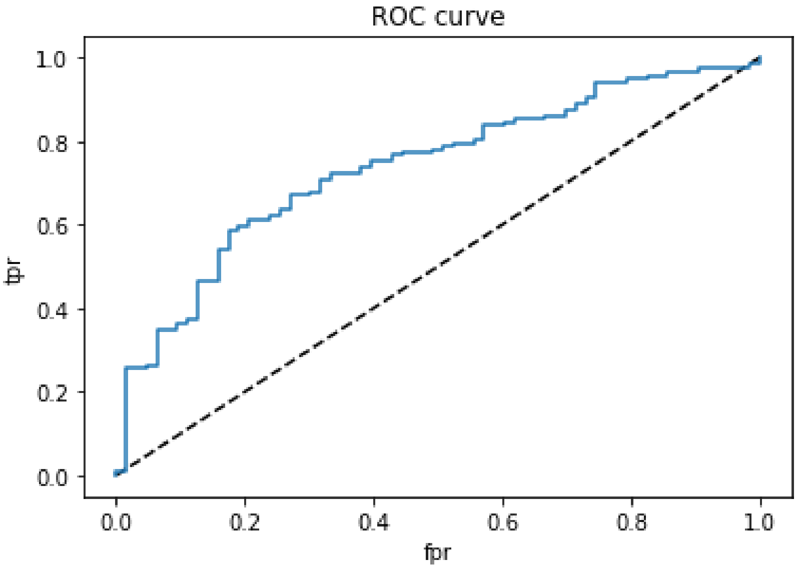




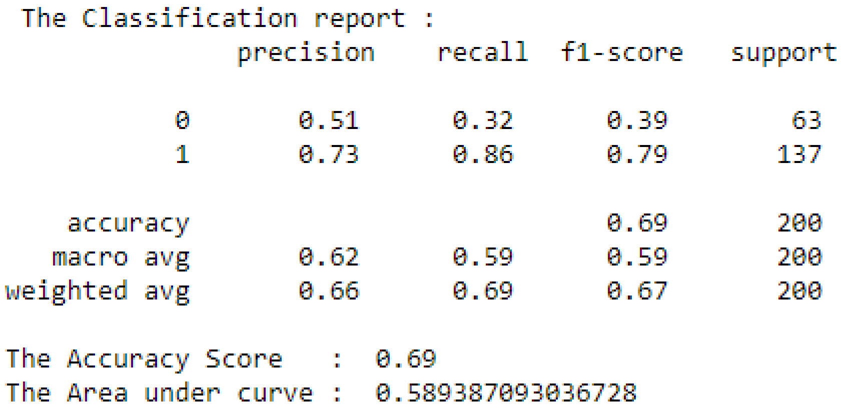
Mostly greater credit history tends to have less risk than that of low credit history. Greater amount tends to have a risk than that of the smaller amount. Risk and Purpose are unrelated concepts. The correlation matrix allows us to confirm the aforementioned finding. The information is further divided into ADA, train, and test sets. The decision tree is used as the base estimator in a boosting model of level 1.



We utilize the classification report and the confusion matrix to assess the model because accuracy is not the appropriate metric. Plotting the ROC curve may also be used to demonstrate the model's effectiveness. The curve illustrates the model's sensitivity and specificity. The model performs better at differentiating the categories the more the curve deviates from the red line. The model has effectively performed on the dataset based on the plot, which shows that it is pretty near to the flat Line.



From the above graph we can see the blue line is far away from the flat line mostly except for the starting and the end. We can know the area under the curve in the classification report with all the averages, precision, and support. The area under the curve we got for the dataset I have taken is 0.589. The model has satisfactorily performed on the dataset based on the plot, where it is pretty near to the flat line.



```
 The Classification report :
              precision    recall  f1-score   support

           0       0.51      0.32      0.39        63
           1       0.73      0.86      0.79       137

    accuracy                           0.69       200
   macro avg       0.62      0.59      0.59       200
weighted avg       0.66      0.69      0.67       200

The Accuracy Score   :  0.69
The Area under curve :  0.589387093036728
```