**Classification Analysis of Default Payment by Credit Card Clients**

**Final Project of IST 707 Data Analytics**

**Rucha Kadam**

**Introduction:**

Data analysis has increased the profitability of various industries. If the loss or profit of a certain product can be analyzed beforehand, it would help many companies to direct their marketing strategies to achieve a higher profit margin. In this report I will be explaining the classification analysis of default payment by credit card customers. Default payment is failure to pay the credit card dues on time. Banks would certainly want to know in advance which kind of client would be paying their dues on time and who wouldn’t. By knowing this, they can either make appropriate offers on interest rates to the various categories of customers. This will ensure the banks a certain amount of profit than being completely blind about it. In this report we will see information about the dataset, preprocessing steps for cleaning the dataset, building models and analyzing the insights obtained from the models. For the data analytics I have used Python on Jupyter Notebook.

**About the dataset:**

The dataset consists of information on default payments: If the payment was made on time or not, demographic factors: age, sex, marital information, credit card details, history of payment and bill statements: previous and current. The dimensions of the dataset are 25 columns and 30000 rows. This dataset was taken from Kaggle.

**Research Questions:**

1. Which variables are the strongest predictors of default payment?
2. How does the probability of default payment vary by categories of different demographic variables?

**Classification Models:**

I have used four different types of classification models: Decision Tree, Random Forest, Ada Boost, Gradient Boost. Later in this report we will see which model gives us better accuracy to determine the answers to the above-mentioned research questions.

**Audience:**

This analysis can be helpful for companies which give out credit cards or loans like banks and investment firms.

**Data Preprocessing:**

There are no null values in the dataset. The description of the dataset provided on Kaggle does not provide any information on data in Education column labeled as category 0, category 5 and category 6. I have replaced these labels as label 4 which is described as other in the dataset.

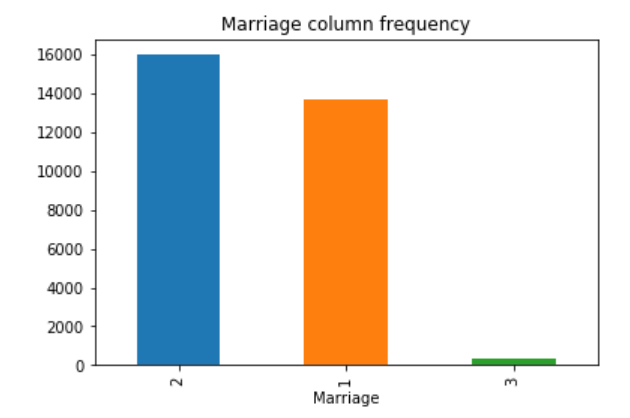
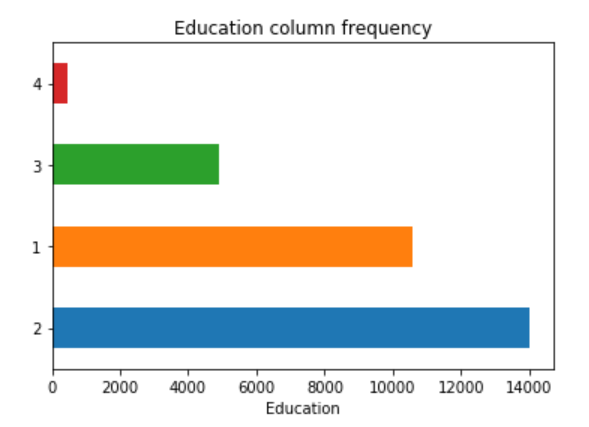
The column Marriage has label 0 which is again undocumented in the dataset description. So, I have replaced all the label 0 values with label 3, which according to description are others, where1 equals to married and 2 equals to single person.

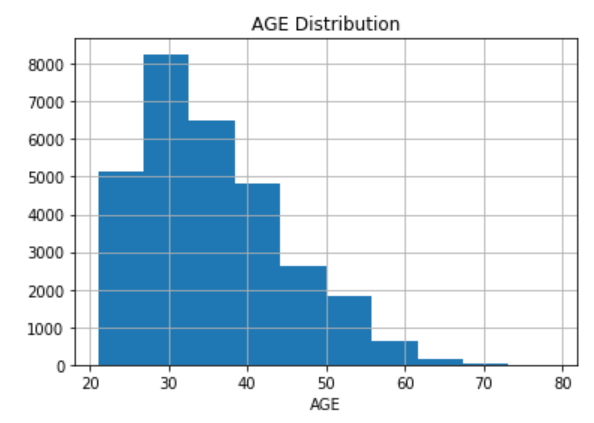
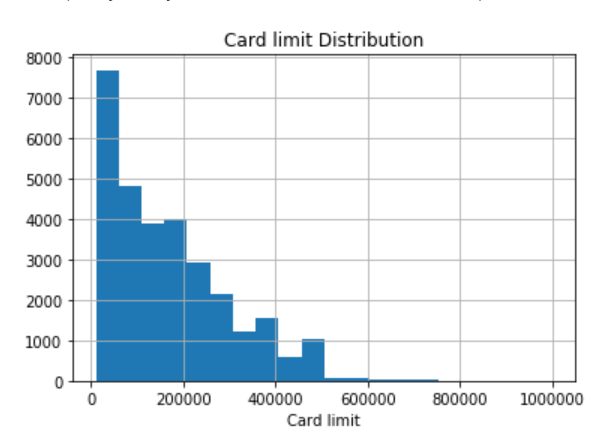
The payment delay description has data labeled as 0, -1 and -2 but there is no description about it in the dataset. Hence, I have replaced -1 and -2 with 0 and we will consider it as the dues were paid by the customer.

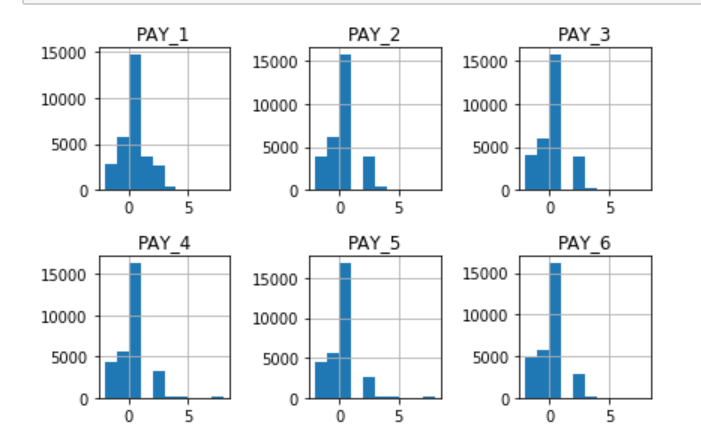
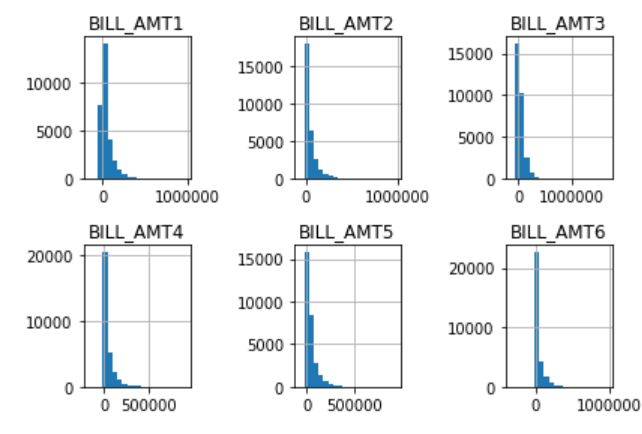
I have also created a feature importance function using .feature\_importance\_ function to create a table of columns and their respective scores. This will further be used in the classifiers ahead.

**Descriptive Analysis:**

To understand how the data set is divided, I have tried to plot the various frequency plots of all the columns. I have performed frequency distribution to understand the value counts in the all the variables and to understand the difference between them as this can also be a factor that has to be considered when applying the variables to the classifier models.

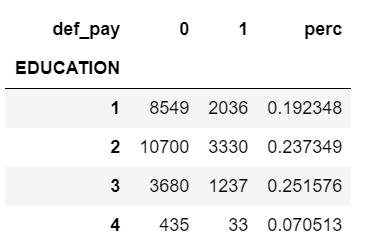
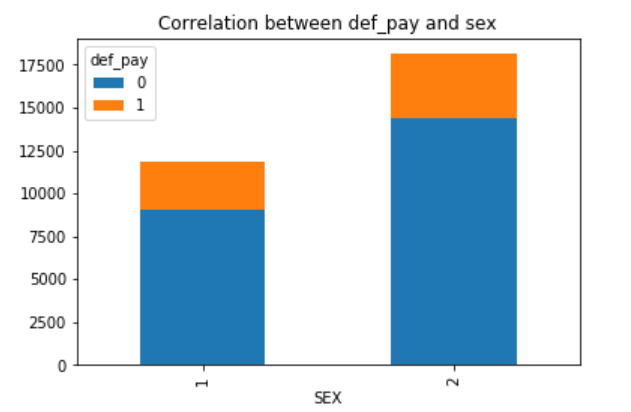
 

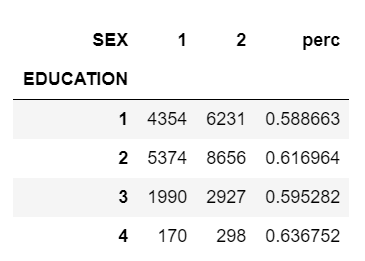
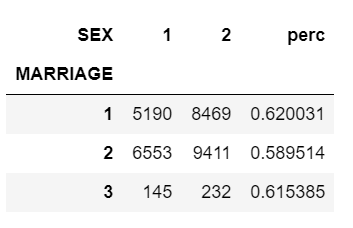
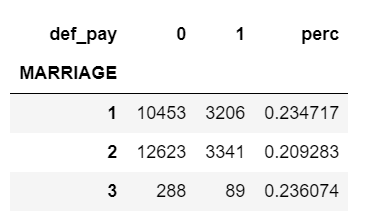


In the marriage column, we can see that the label 3 has the least number of counts than the rest while in education variable, label 4 has the least counts. We can see from the age distribution graph that the number of people using credit cards is more between 25 to 40 age group. Older people seem to be less interested in issuing credit cards. The card limit distribution graph shows that very a smaller number of people have high limit while maximum people have card limit upto 100000. Note that the data set is Taiwanese. Hence is the currency. Not much information can be obtained from the bill amount variables and the previous payment variable. We will increase the number of bins in the pre-processing steps ahead for some of the columns to get a clearer view.

**Finding the correlations between the variables:**

In this step, I have tried to see how relevant each feature is. For this I have applied correlations.





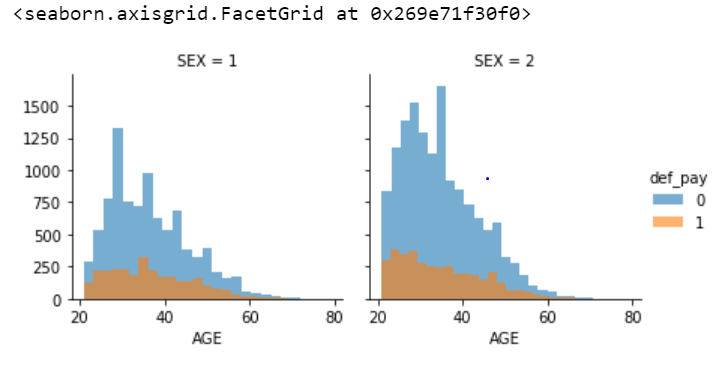
In the first graph we see the correlation between sex variable and def\_pay variable. In def\_pay, 0= default, 1 = no default and 1 = male, 2 = female. From the graph we can see that number of defaults by female are more than males and men are most likely to default in the next month than the females. When I correlated education and def\_pay, I found that higher is the education, more is the probability of defaulting in the next month. Correlation between marriage and def\_pay shows that married females and single men have higher possibility of defaulting a credit card bill than the others. The correlation between sex and education shows that both genders with higher education: university and graduate education show higher possibility of defaulting a credit card payment.

**Preprocessing steps for classification:**

1. In the first step, we look at the variables of repayment status and bill amount. From the analysis, we can say that

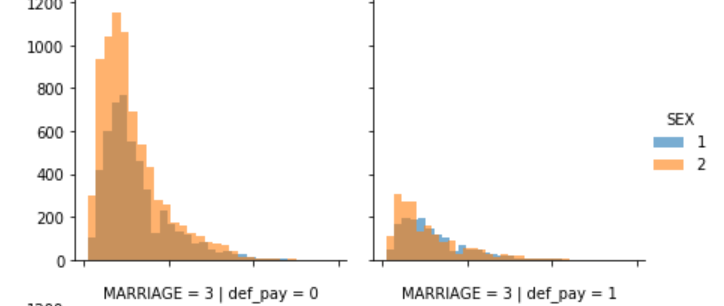
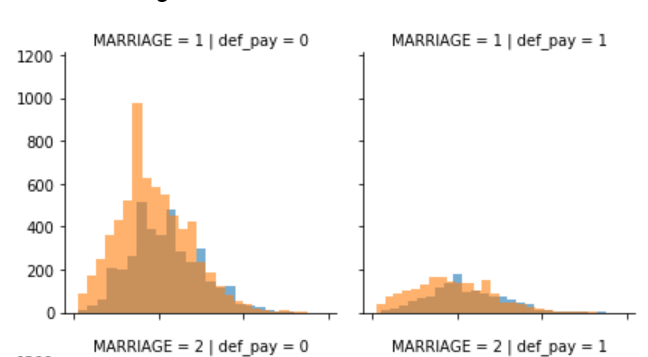
* If we have a BILL of X, we pay Y
* The month after I must pay X-Y + X', being X' my new expenses, I pay Y'
* The month after I must pay X+X' - Y - Y' + X'', I pay Y''
* So, on so forth. On top of that I may or may not have months of delay.
* It seems that if by September I have a bill too close to my limit, I generally fail. However, I can already see some dramatic exceptions.
* Moreover, some clients that joined our dataset at a later month: they have 0 in BILL and PAY AMT for a while and then they start.

1. Next, using the sns library facegrid function, I have tried to compare the relationship between sex, age and def\_pay. We get the following output.



As we can see that as age increases, in both genders, the default payment decreases, or we can say that people become more responsible with age. Also, we can see more variations in the female graph then for the males, but there seems to be a similar pattern.

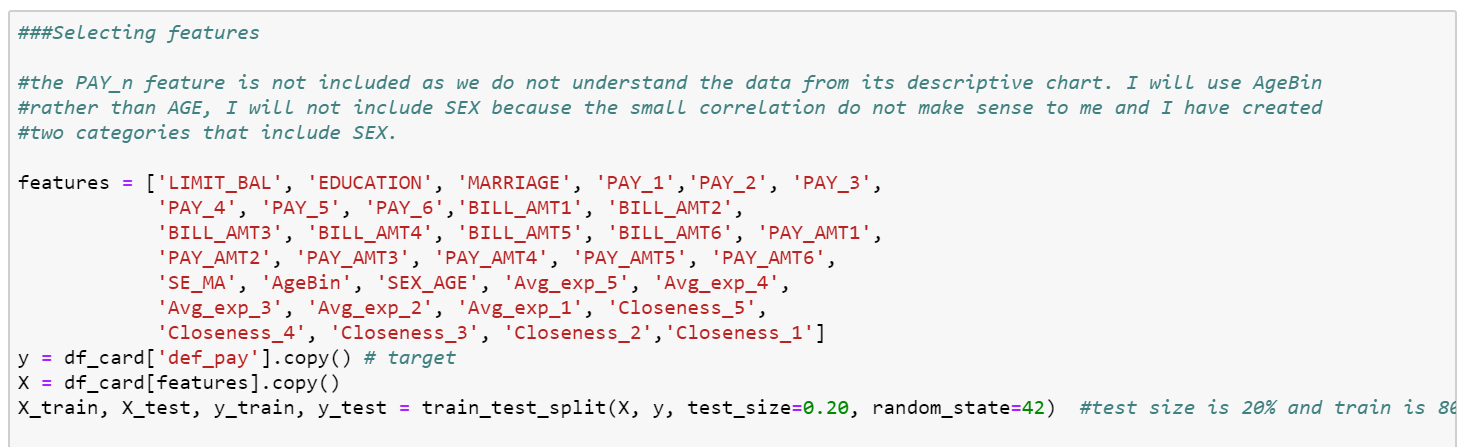
1. Next, using the sns library facegrid function, I have tried to compare the relationship between sex, marriage and def\_pay. We get the following output. We can see that variable age compared with other variables gives more descriptive results.



1. Now, for better analysis, I have combined the data from two variables as follows (refer code file for details):
   1. Sex with marriage renamed as SE\_MA – and divided it as married man, single man, divorced man and same for woman
   2. Sex with Age renamed as SE\_AG – and divided into male and female and further into ages. For example: male in 20’s, female in 30’s etc. so that we get a wider range to compare.
   3. Bill amount and payment is renamed as Client – this is the comparision between the various bill amount and payment combinations (6).
   4. Bill amount and Balance limit renamed as closeness – This basically gives how far the bill is from the limit should matter. Since the result can vary a lot from one client to the other (the LIMIT\_BAL variable has a very wide range), I will again weight this difference on the LIMIT\_BAL feature.

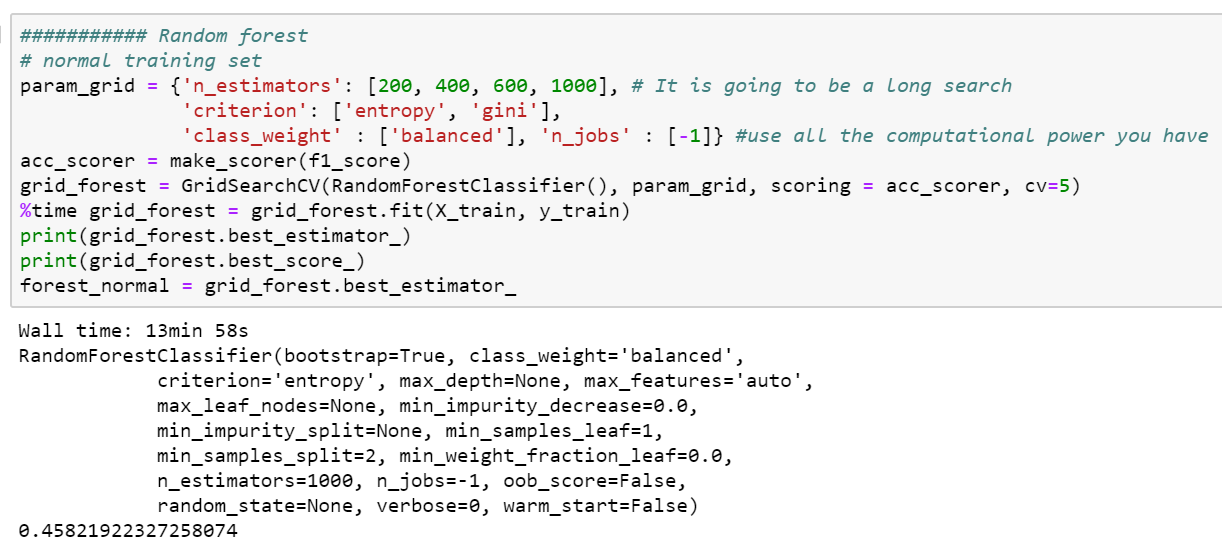
**Top 3 Machine Learning Models:**

Feature selection: I have used the variables given below in models: Random Forest, Ada Boost and Gradient Boost. Here you can also see that I have separated the training size to 80% of the data set and the test to 20%.

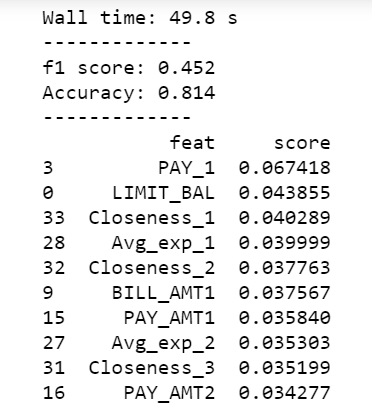
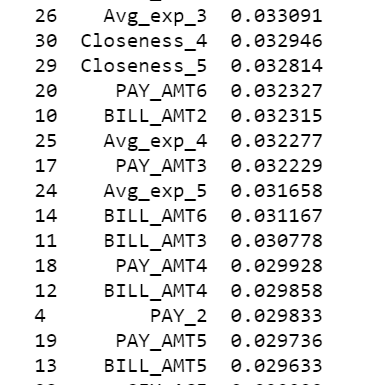
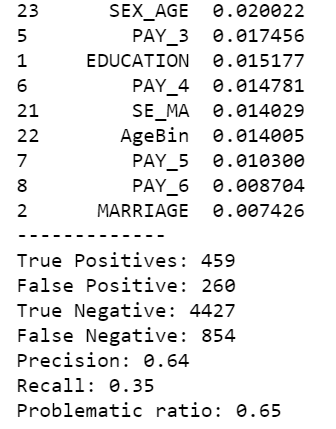


1. Random Forest:

Here we train the model and find out the best parameters which will help us get the optimum accuracy without the case of overfitting using the gridSearchCV function and print the best score.

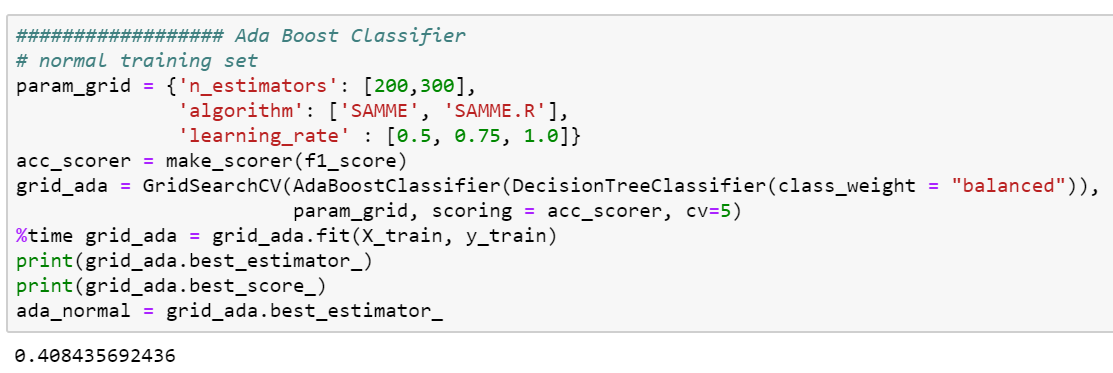


Output after applying the best model

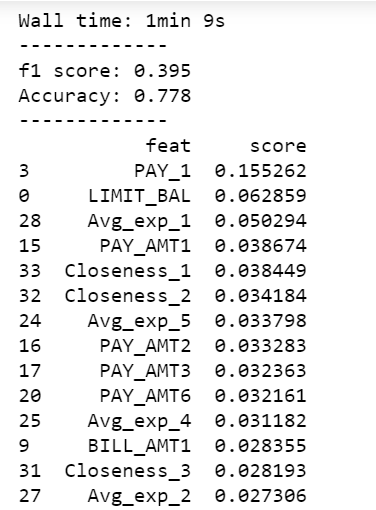
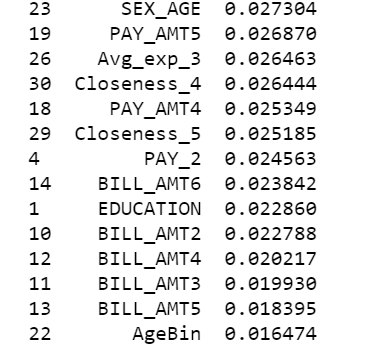
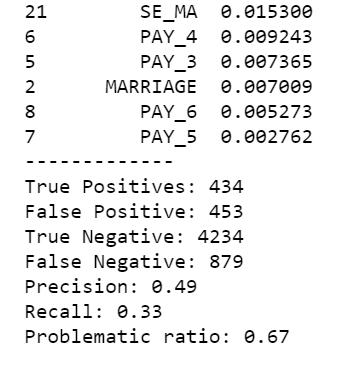
  

1. Ada Boost:

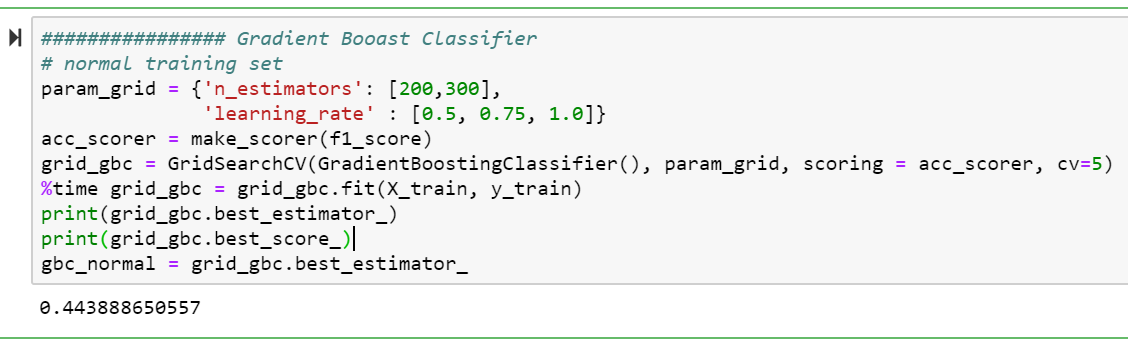
Here we train the model and find out the best parameters which will help us get the optimum accuracy without the case of overfitting using the gridSearchCV function and print the best score.



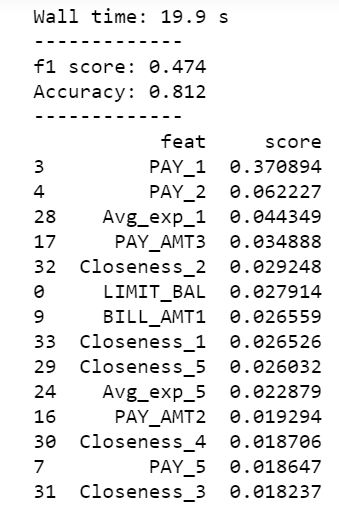
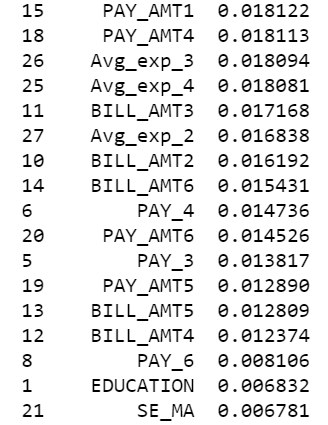
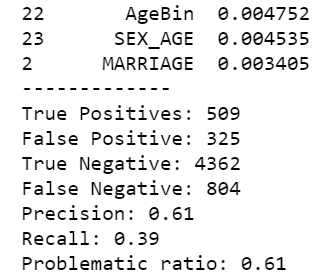
Output after applying the best model

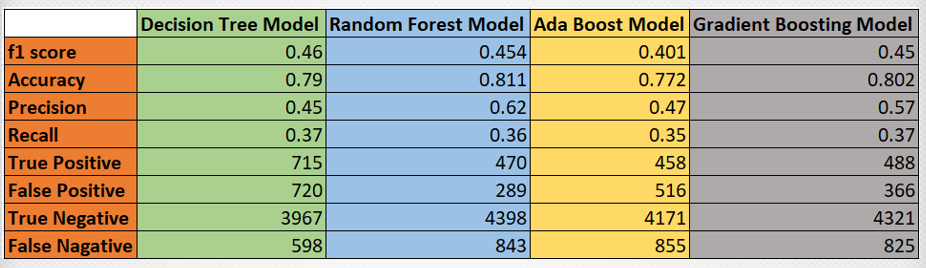
1. Gradient Boost Classifier



Output after applying the best model

**CONCLUSION:**



From the given models, we can say that random forest model gives us the best accuracy at 81.1% with the moderately better f1 score and precision. Also, the featured parameters used for the model can help us determine if an individual might default the payment or not.