Cristhian Salvania Tolentino

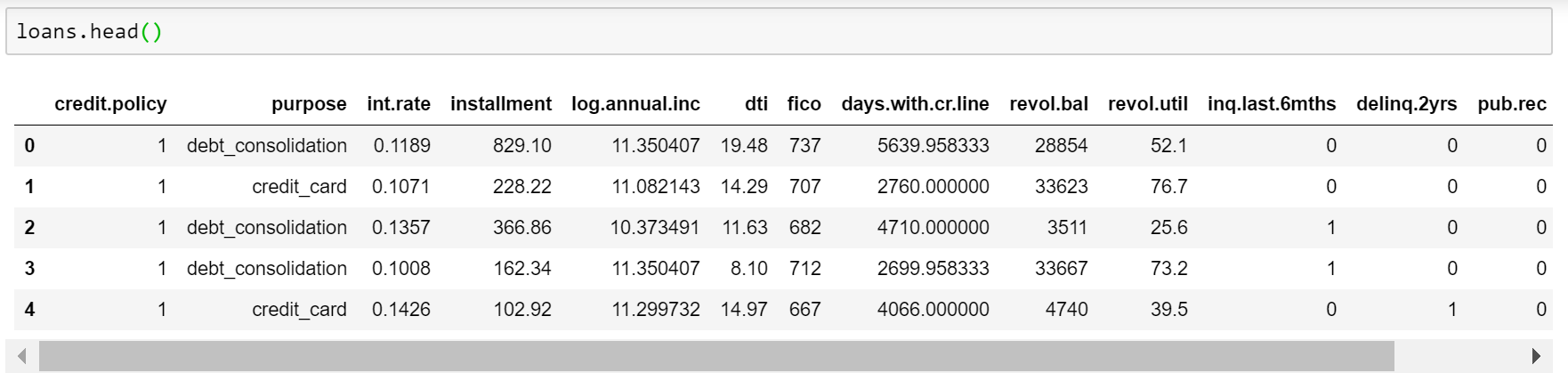
Professor Celepcikay

CS4319-21823

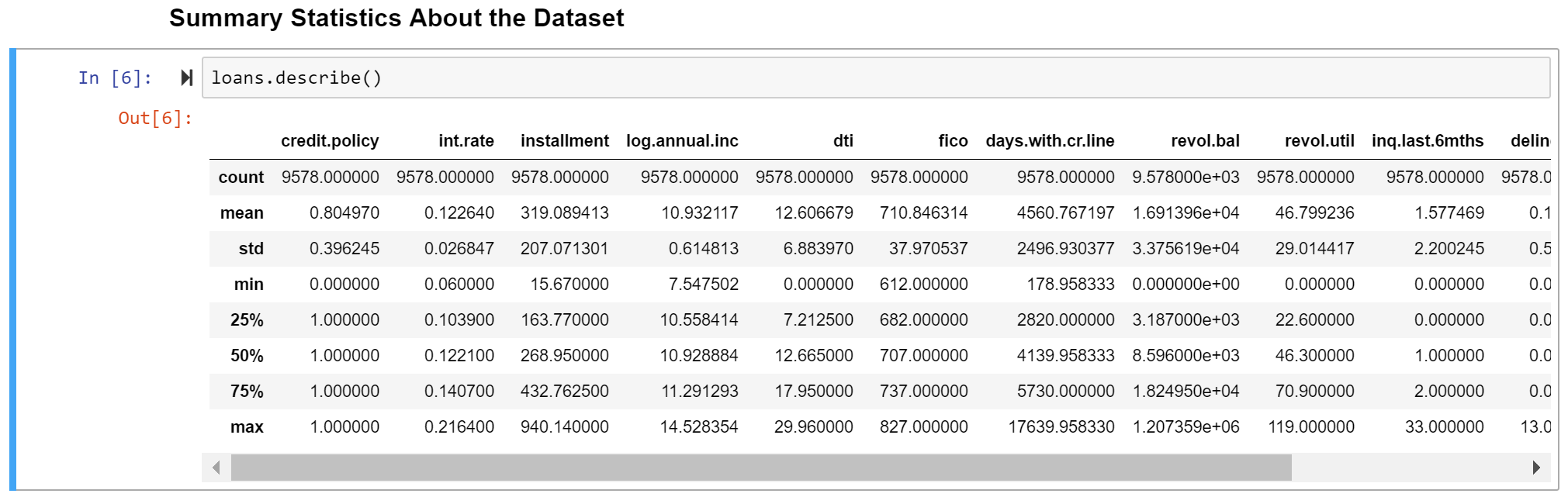
May 5, 2019

Final Project Part 1: Lending Club Decision Tree

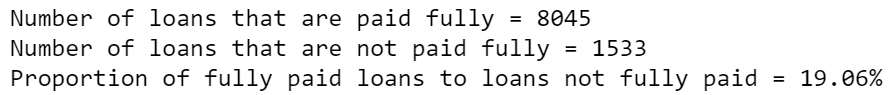
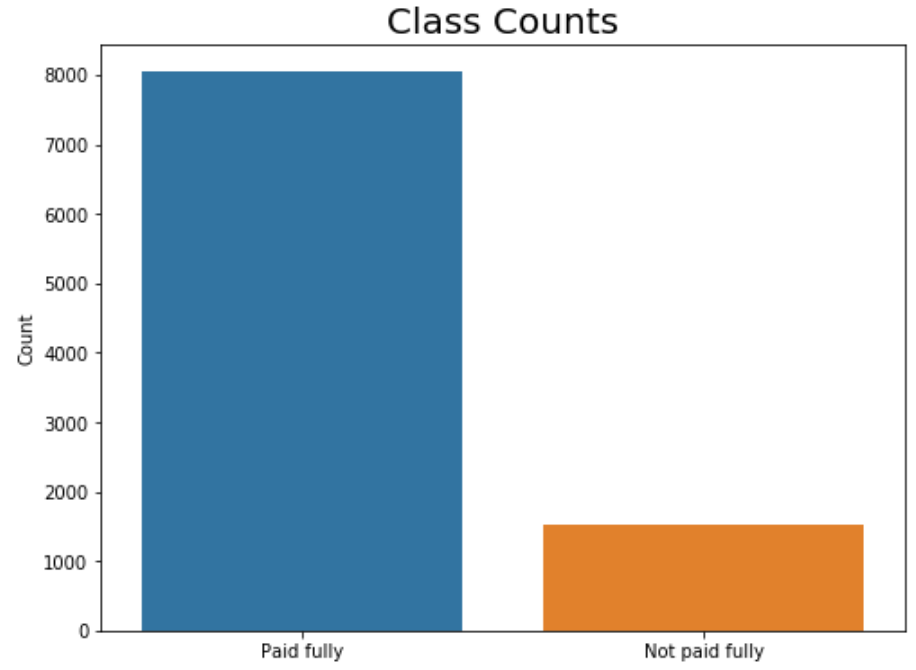
Lending Club connects people who need money (borrowers) with people who have money (investors). As an investor, I want to invest in people who showed a portfolio of having a high probability of paying me back. In this project, I have created different types of data visualizations to help me better understand their portfolio in order to create a model to help me classify and predict whether or not the borrower paid back their loan in full.



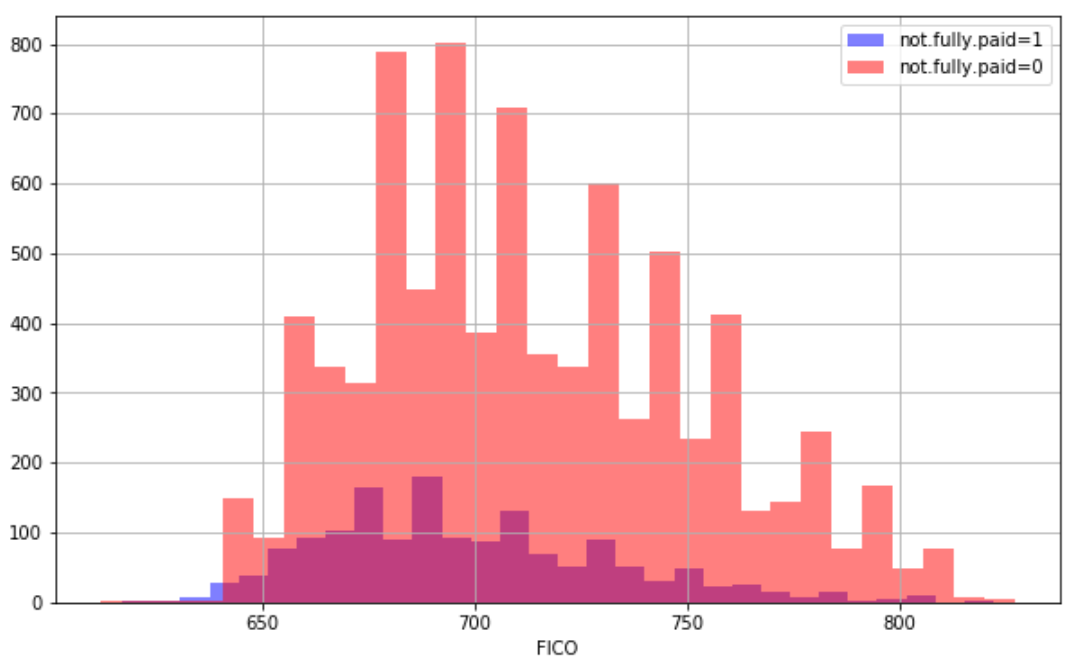
Each row represents an entry for a loan: type, continuous numerical attributes with floating point values, and one class, which tells us whether the loan is fully paid or not.



The dataset contains a lot of attributes, so I will briefly describe a few of the columns. 75% of the total loans meet the credit underwriting policy, have an interest rate of 14.07% and have an installment amount of $432.76, and are fully paid. The borrower’s lowest FICO credit score is 612 and his/her highest FICO credit score is 827.

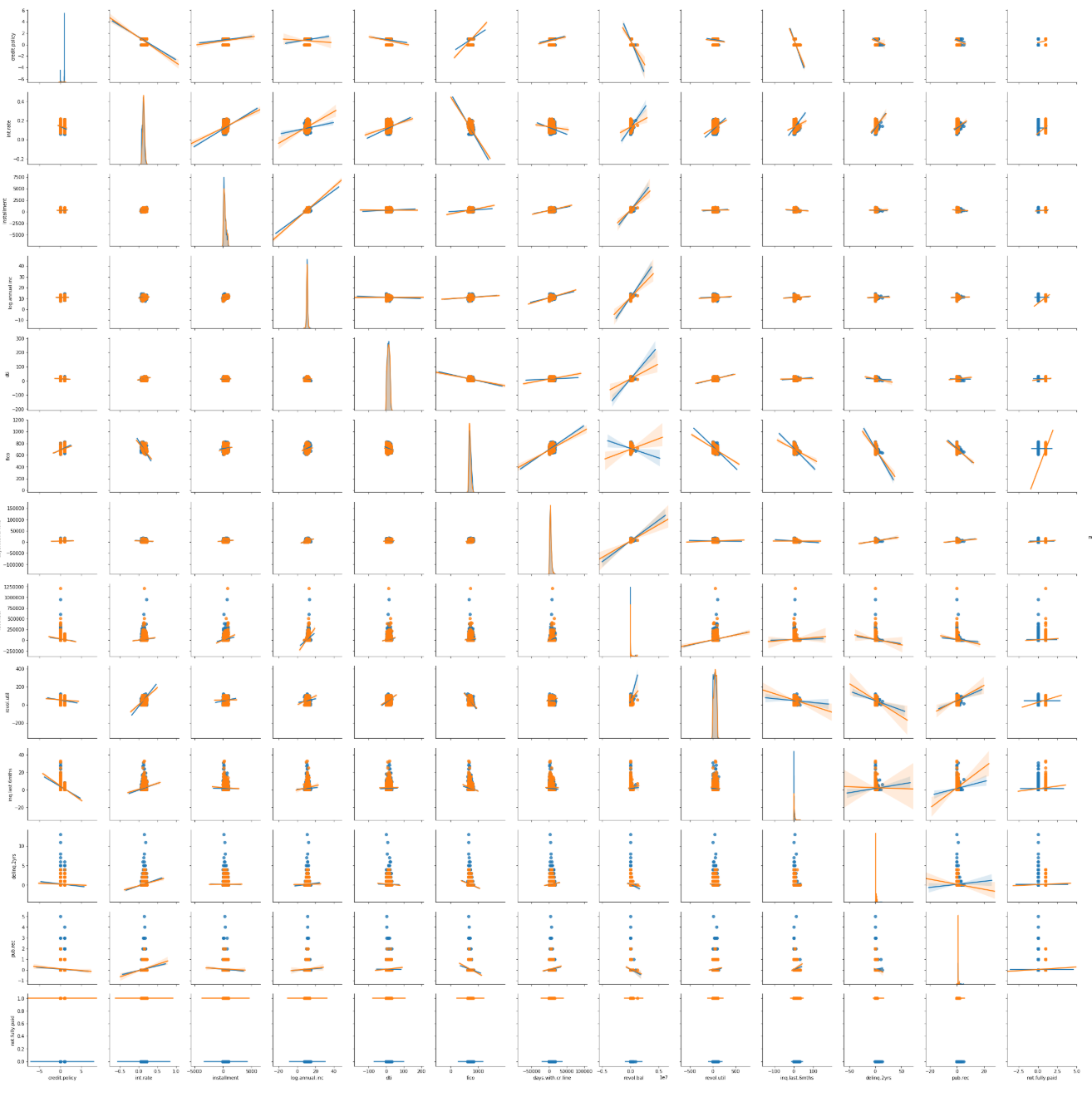


About 84% of the loans have been fully paid and 16% have not been paid fully.

****

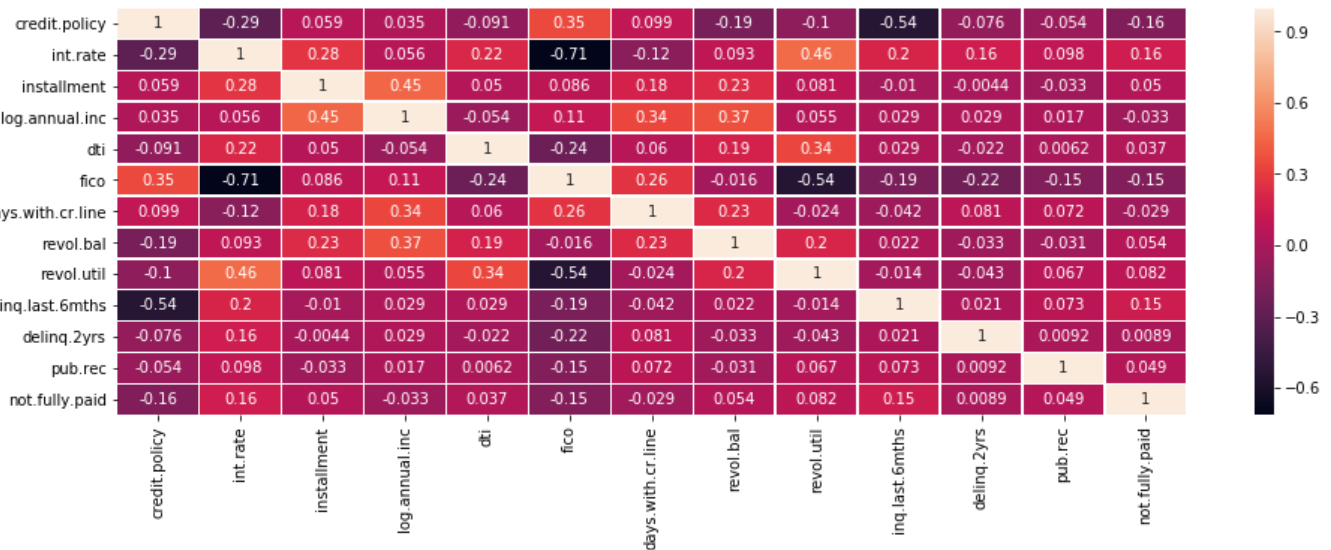
This histogram shows the joint distribution of values of fico and not.fully.paid attributes. As we can see, there are a little over of 800 not fully paid loans when the borrower has a credit score of about 690.

**Scatterplot of All Pairs of Attributes**



**Correlation Coefficient Matrix**

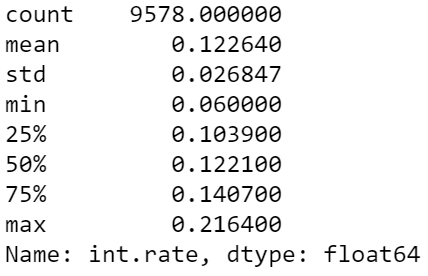
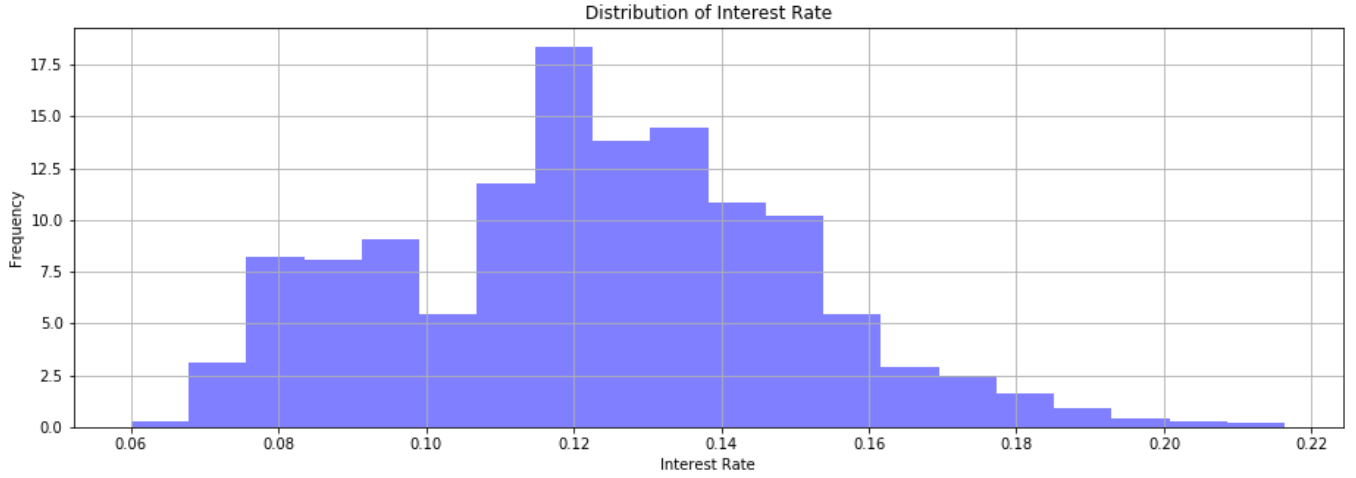
This correlation coefficient matrix is displayed as a color-encoded matrix, which clarifies what attribute pairs have the strongest relationships and the weakest relationships. Focusing on the bottom half of the table and below the diagonal line of the perfect ones, we can see that there are three strong negative correlationships colored in black squares: credit.policy and inq.last.6mths at -0.54, fico and revolu.util at -0.54, and int.rate and fico at 0.71.



Although (credit.policy, inq.last.6mths), (int.rate, fico), and (fico, revol.util) have the highest correlations of all the total pairs of attributes, they are still not the best attributes to predict whether the borrower will pay off their loans. Because we are building a decision tree based on ‘not.fully.paid’ attribute, this is the dependent variable and thus, we’ll be focusing its relationships with the attributes that it has the strongest with.

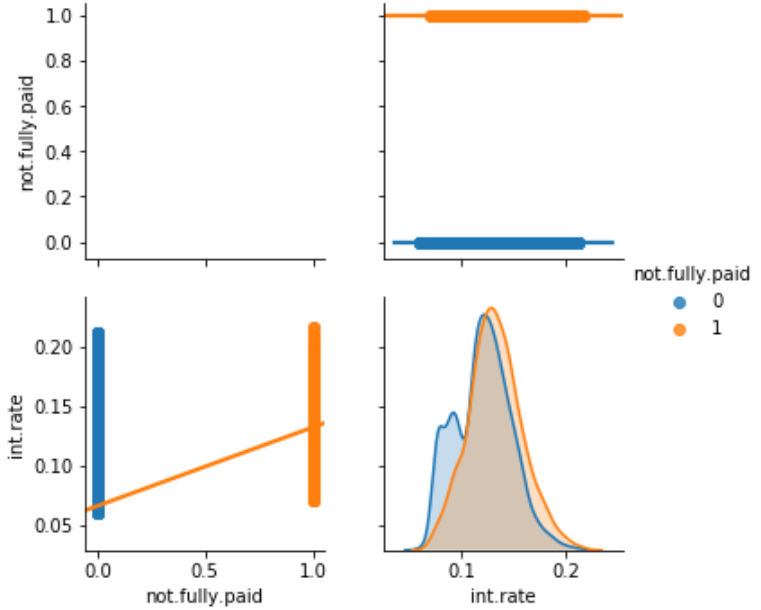
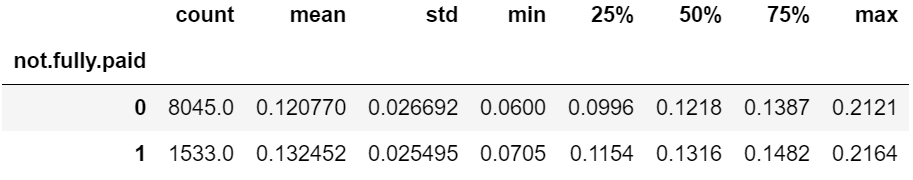
**int.rate**

Interest rates range from 6.00% to 21.64% with a mean of 12.26%. We can see that 25% of total loans have interest rates lower than 10.39%, while 50% have interest rates between 10.39% and 14.07%.



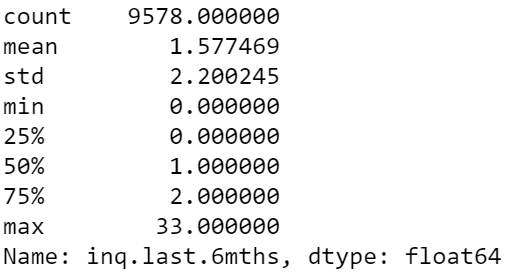
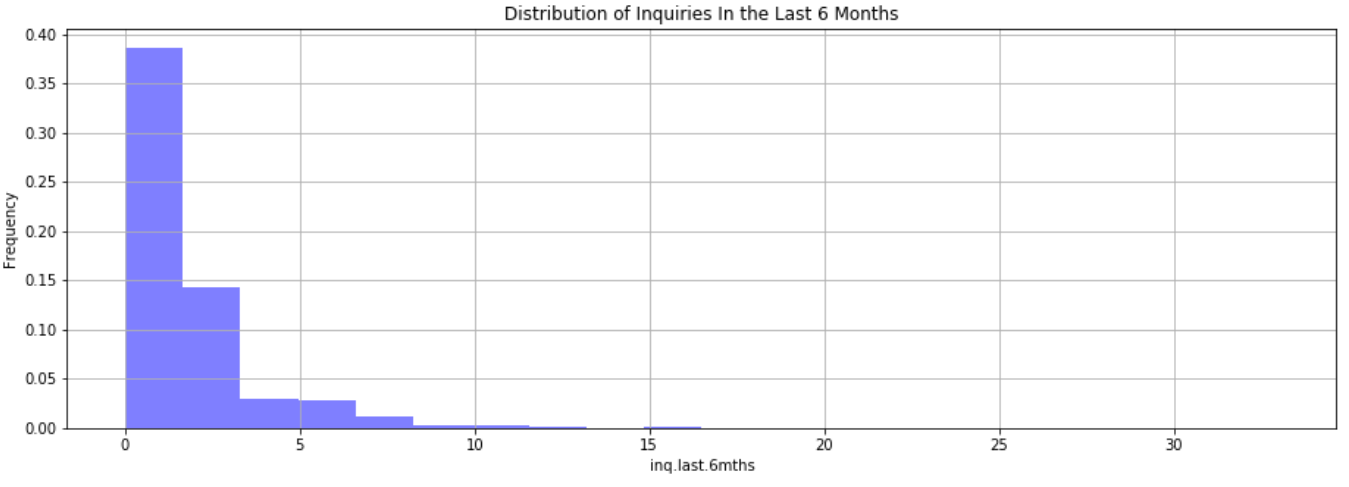
**Relationship between not.fully.paid and int.rate**

According to the summary statistics and linear regression model, not fully paid loans tend to have much higher interest rates than fully paid loans.



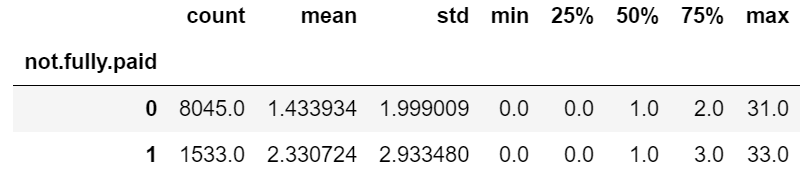
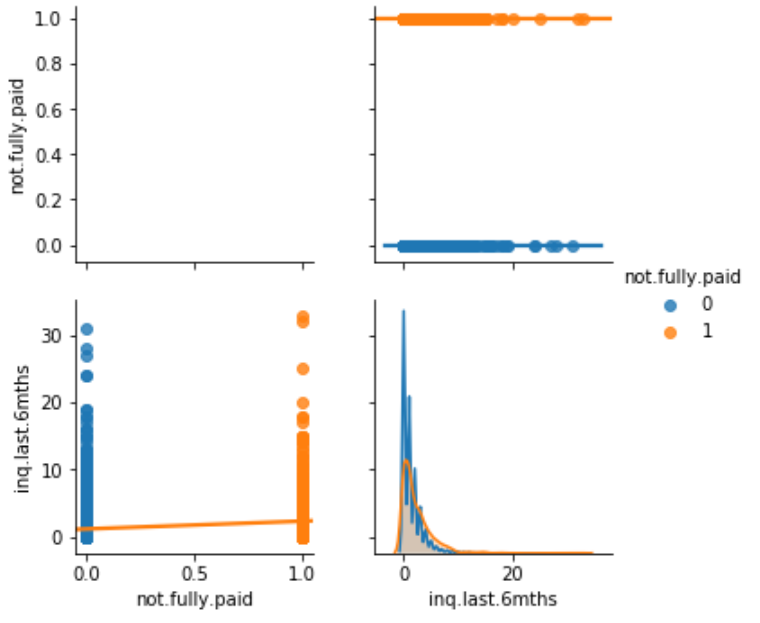
**inq.last.6mths**

Inquires in the last 6 months range from 0 to 33 inquiries.



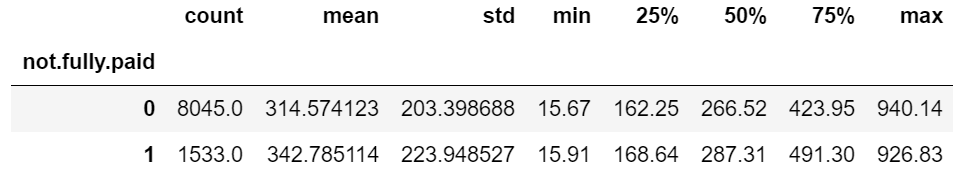
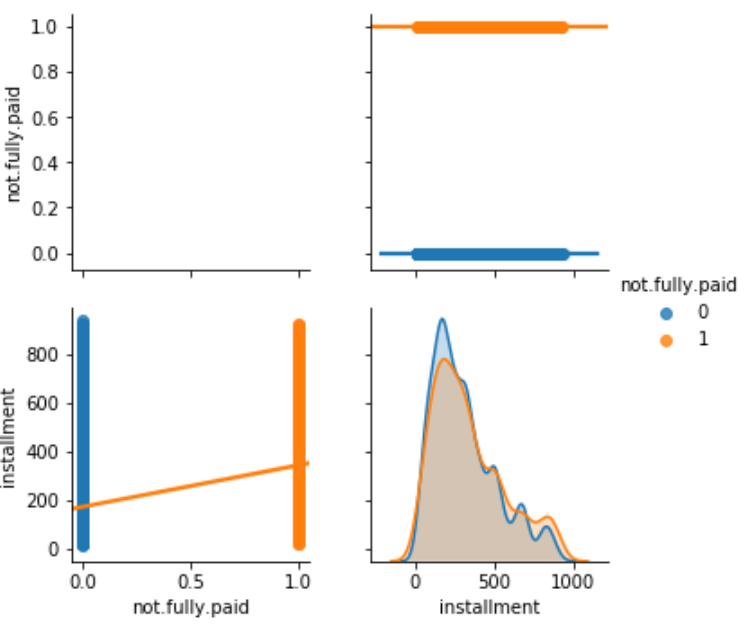
**Relationship between not.fully.paid and inq.last.6mths**

Not fully paid loans tend to have 0.9 more inquires in the last months.

**** ****

**Relationship between not.fully.paid and installment**

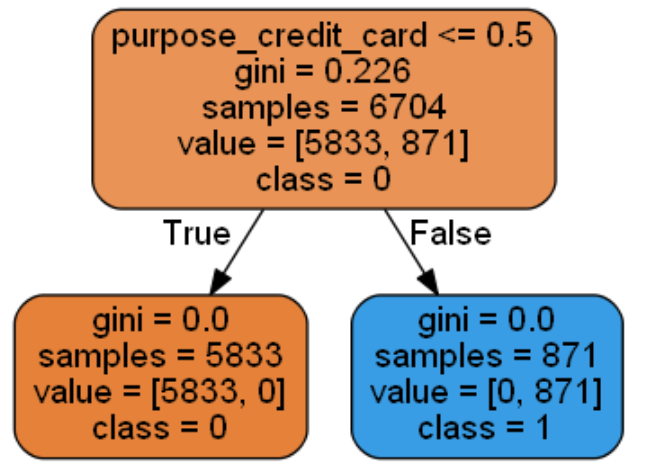
On average, loans that are not fully paid have $28 higher installments than fully paid loans.

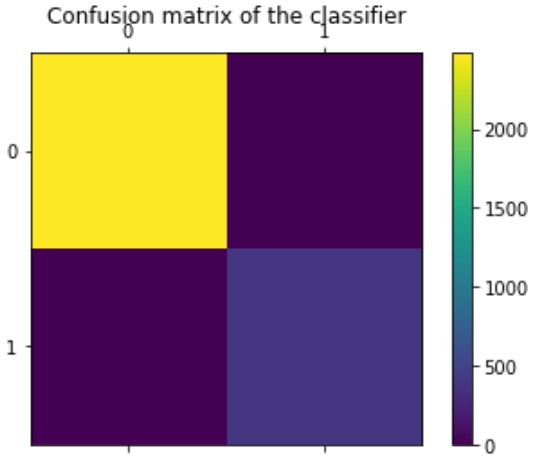
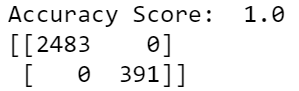
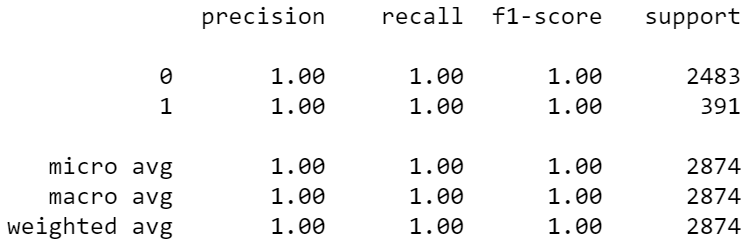
**** ****

**Decision Tree Model**

After data analysis, I have concluded which features will be included in the decision tree. The model will have 17 independent variables: credit.policy, int.rate, installment, log.annual.inc, dti, fico, days.with.cr.line, revol.bal, revol.util, inq.last.6mths, pub.rec, purpose\_credit\_card, purpose\_debt\_consolidation, purpose\_educational, purpose\_home\_improvement, purpose\_major\_purpose, and purpose\_small\_business; while the dependent variable is not.fully.paid.

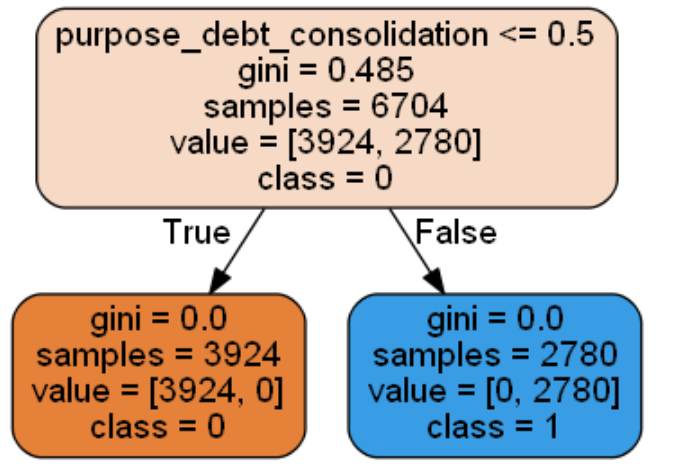
I have generated decision trees based on loan purposes. The first decision tree is based on samples tested with 6,704 credit card loans:

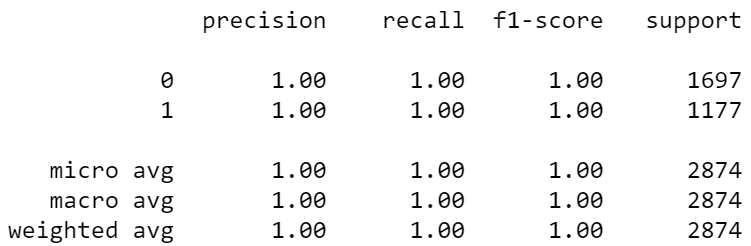
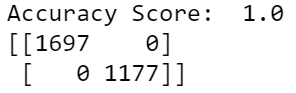
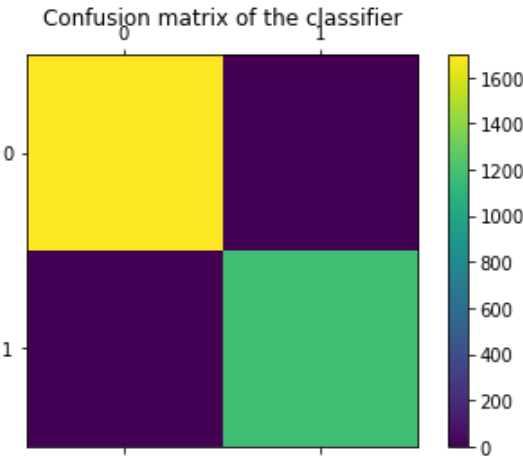


The accuracy score and precision are both calculated as 1, which means that this decision tree on credit cards has accurately and precisely predicted 5,833 are fully paid and 871 are not fully paid.

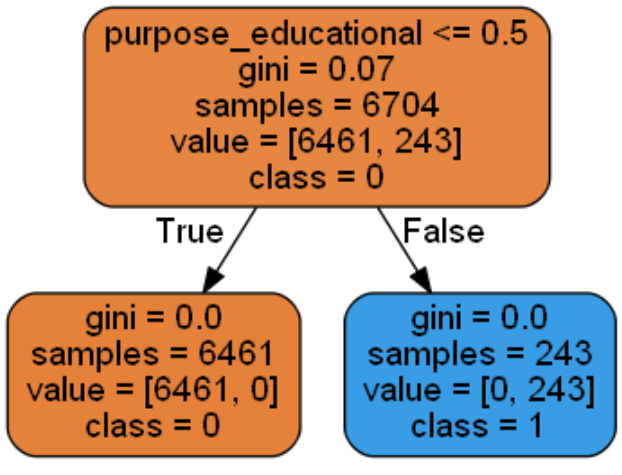
This decision tree is based on 6,704 samples of debt consolidation loans:

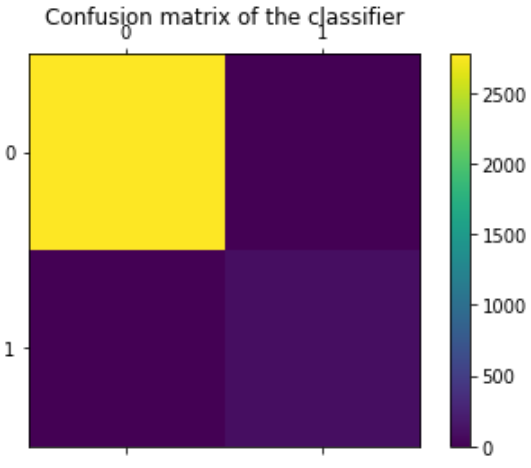
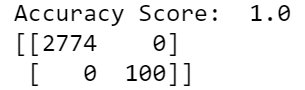
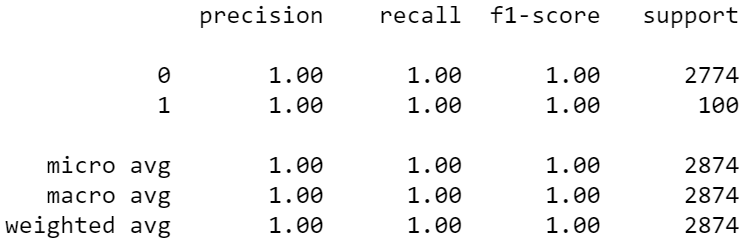




The accuracy score and precision are both calculated as 1, which means that this decision tree on credit cards has accurately and precisely predicted 3,924 are fully paid and 2,780 are not fully paid.

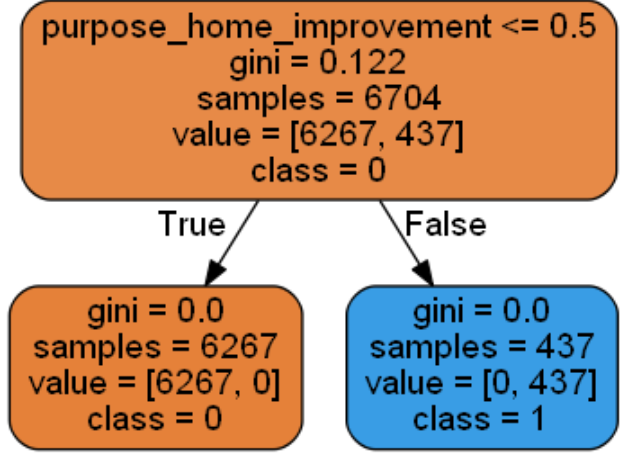
This decision tree is based on 6,704 samples of educational loans:

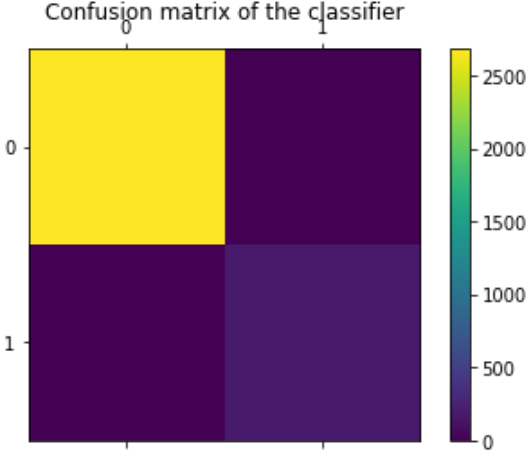
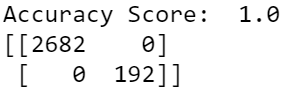
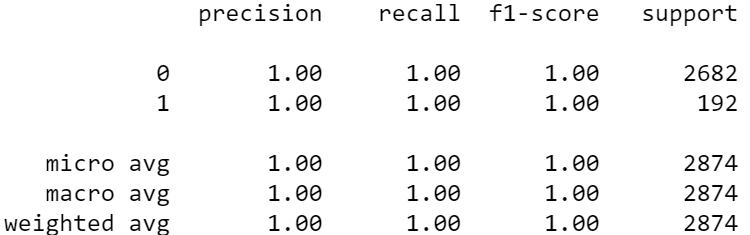


The accuracy score and precision are both calculated as 1, which means that this decision tree on educational loans has accurately and precisely predicted 6,461 are fully paid and 243 are not fully paid.

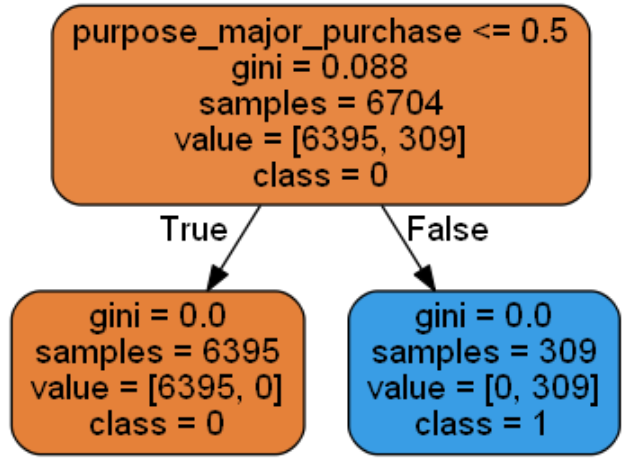
This decision tree is based on 6,704 samples of home improvement loans:

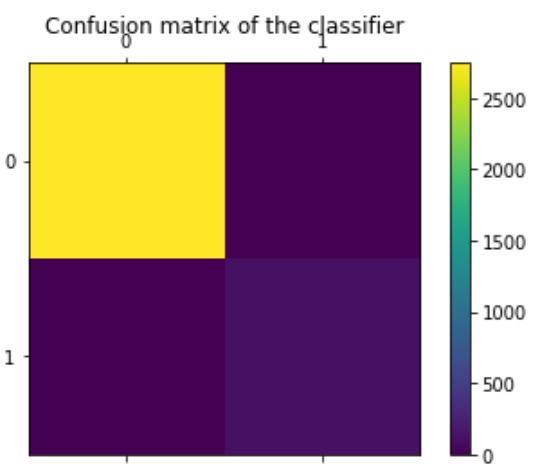
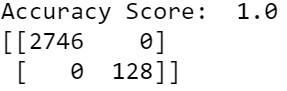
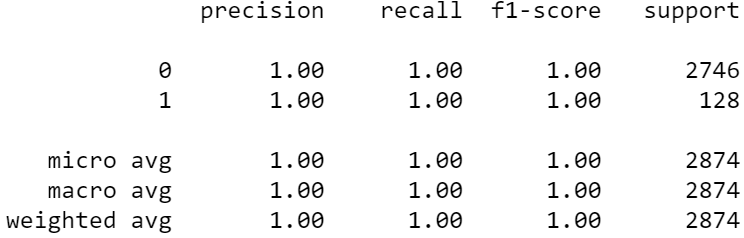


The accuracy score and precision are both calculated as 1, which means that this decision tree on home improvement loans has accurately and precisely predicted 6,267 are fully paid and 437 are not fully paid.

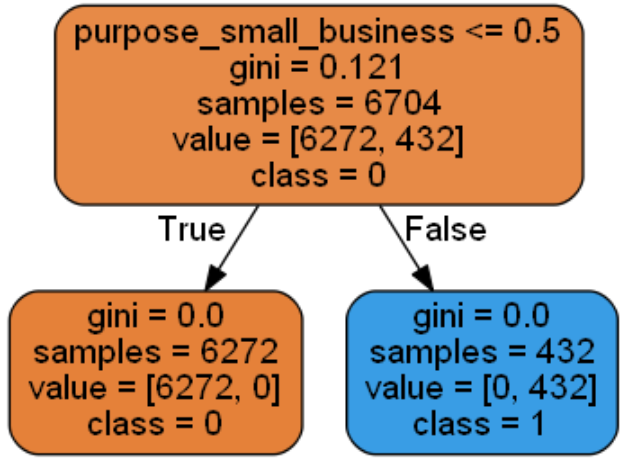
This decision tree is based on 6,704 samples of major purchase loans:

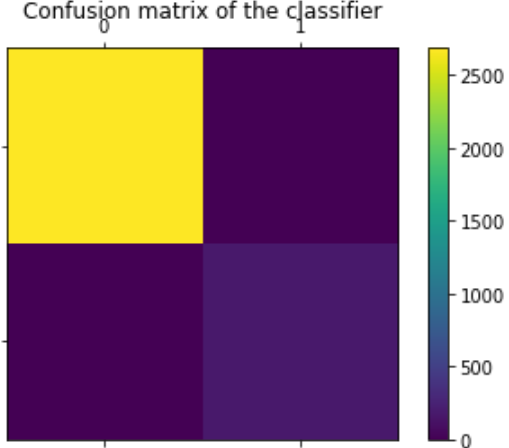
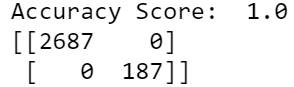
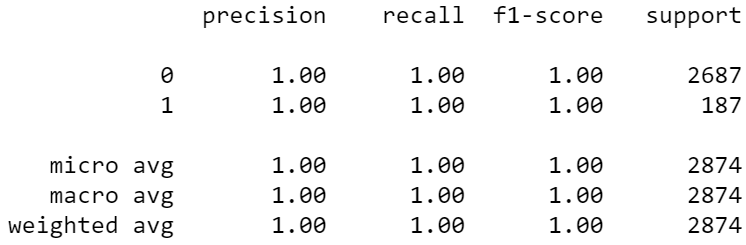


The accuracy score and precision are both calculated as 1, which means that this decision tree on home improvement loans has accurately and precisely predicted 6,395 are fully paid and 309 are not fully paid

This decision tree is based on 6,704 samples of small business loans:



The accuracy score and precision are both calculated as 1, which means that this decision tree on home improvement loans has accurately and precisely predicted 6,272 are fully paid and 432 are not fully paid.

**Decision Tree Based On Interest Rate**

Since 'not.fully.paid' has the strongest positive correlation with 'int.rate', I thought this will help me determine whether the borrower will pay back his/her loans or not. However, interest rate is a continuous attribute with floating point values, so the DecisionTreeRegressor() method had to be used and the decision tree definitely lost information when it was categorizing the attributes in different class labels. I even hardcoded for the class labels to show up, but the leaf nodes still did not show the class labels like the DecisionTreeClassifier did. Therefore, I did not continue to make decision trees for the rest of the attributes.

**Conclusion**

In this part, I have learned the overall process of Data Mining to predict whether or not the borrower will pay off a loan. After training and building decision trees based on purpose, I found that these models provided a somewhat informed prediction of the probability that a loan will be fully paid. Looking at all the decision trees, there are more fully paid loans than not fully paid loans. According to the Correlation Coefficient matrix, the most important attributes for predicting not fully paid are interest rate and inquiry in the last 6 months.