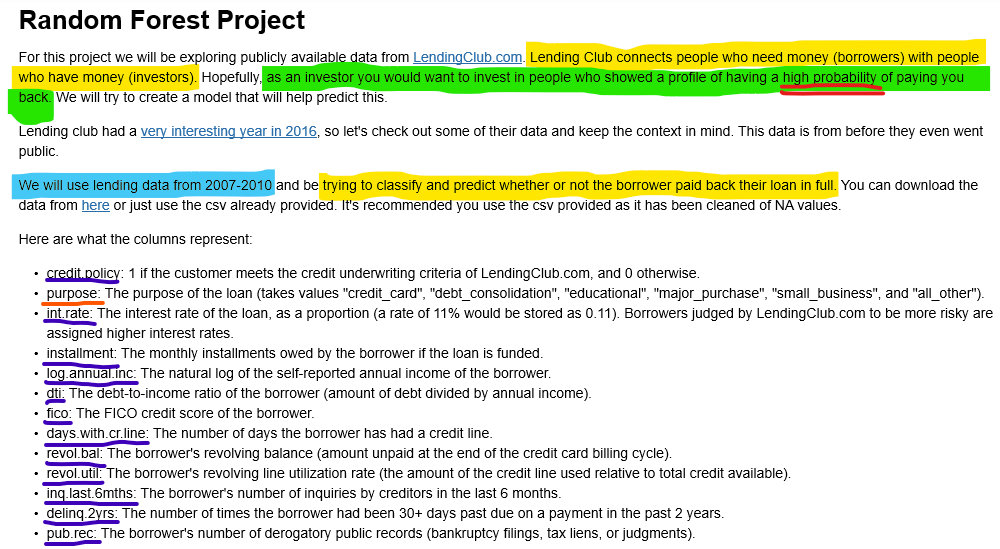
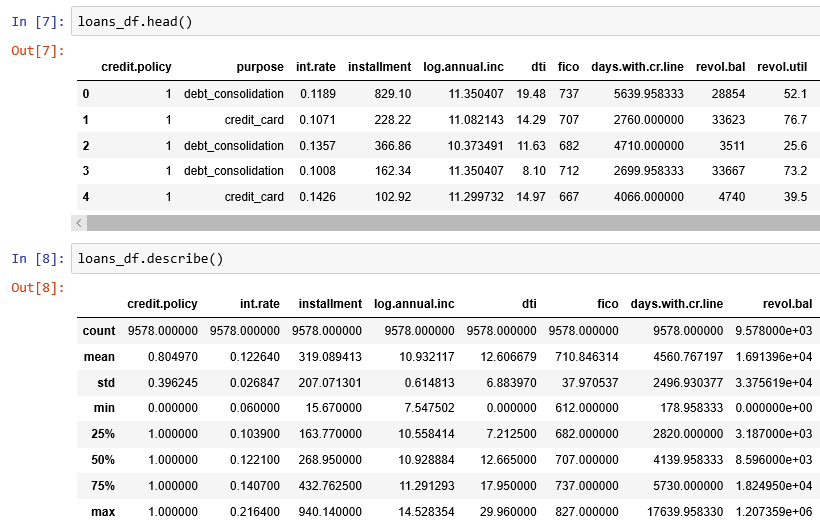
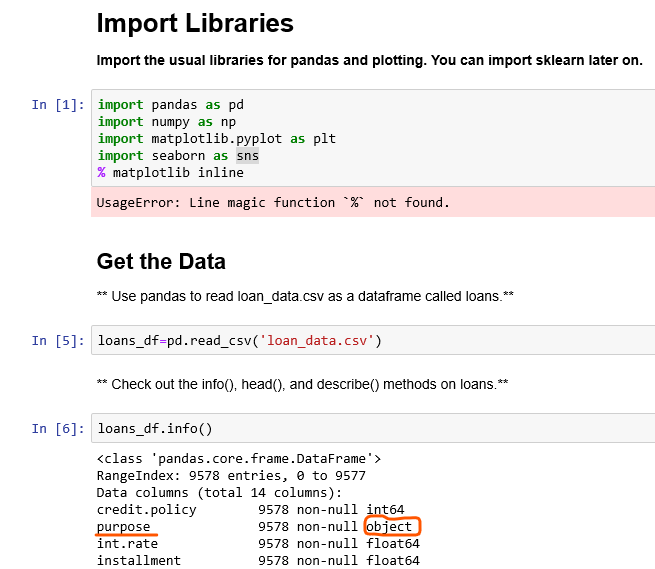
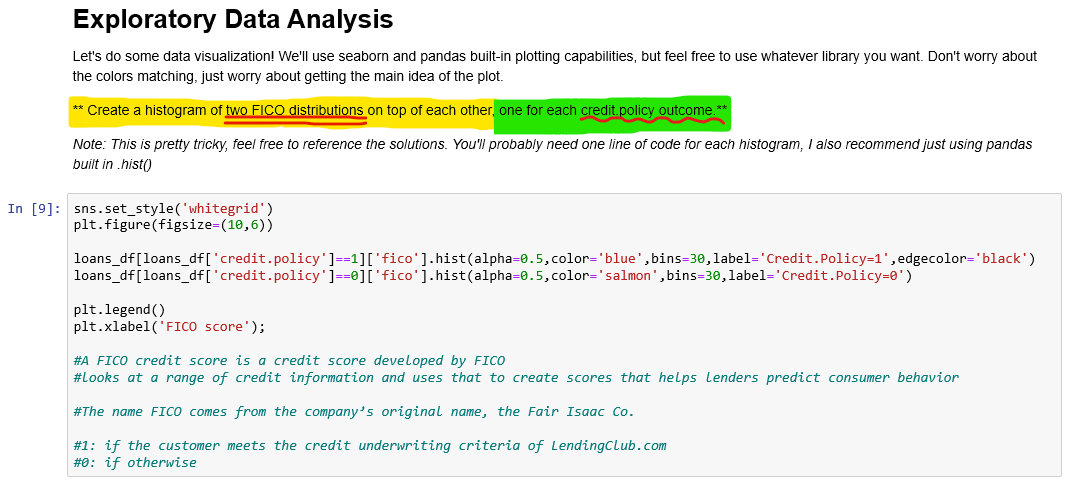
**Decision Trees and Random Forests**



**...**

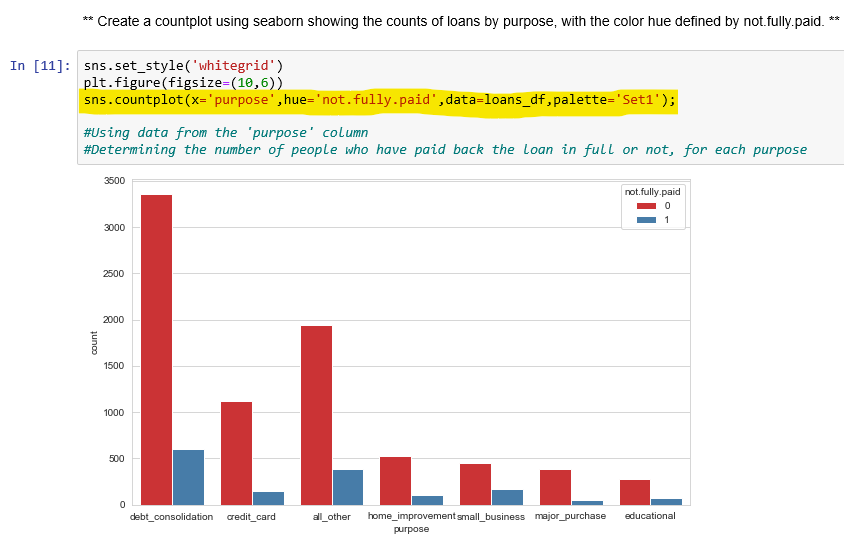
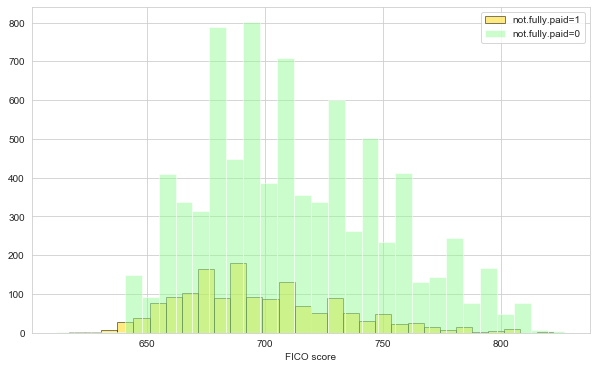
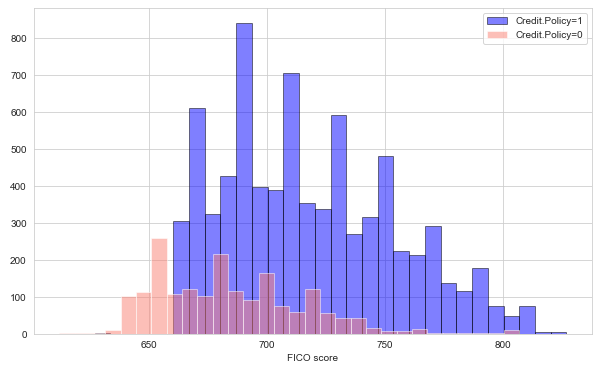
**...**

**Categorical data**

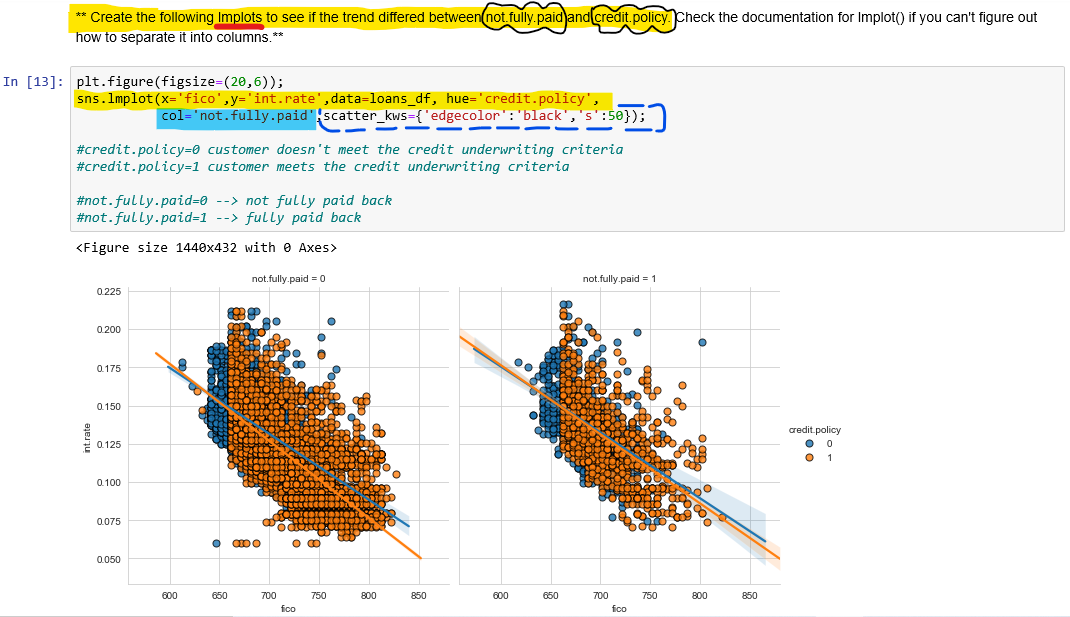
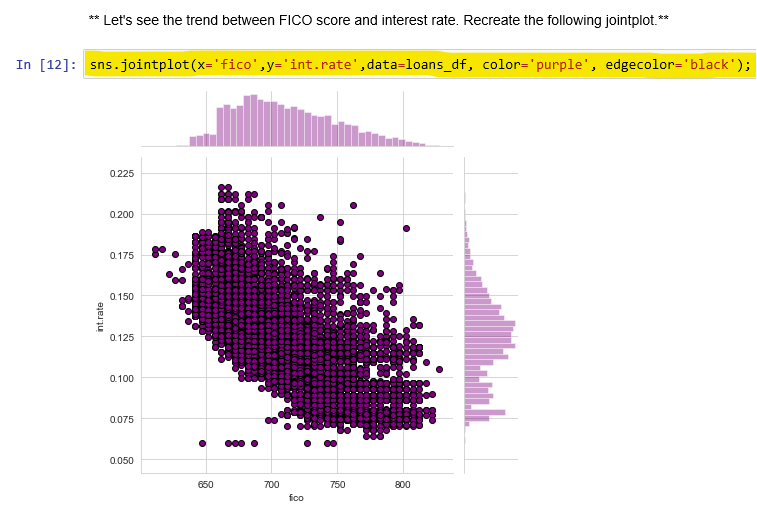


Data on the x-axis

Histogram data



* More customers **did not pay back** the loan in full than did
* This outperformance was most prominent for debt consolidation loans



**lmplot():** **Linear Regression plot**

Parameter that the data will be compared to

**col**

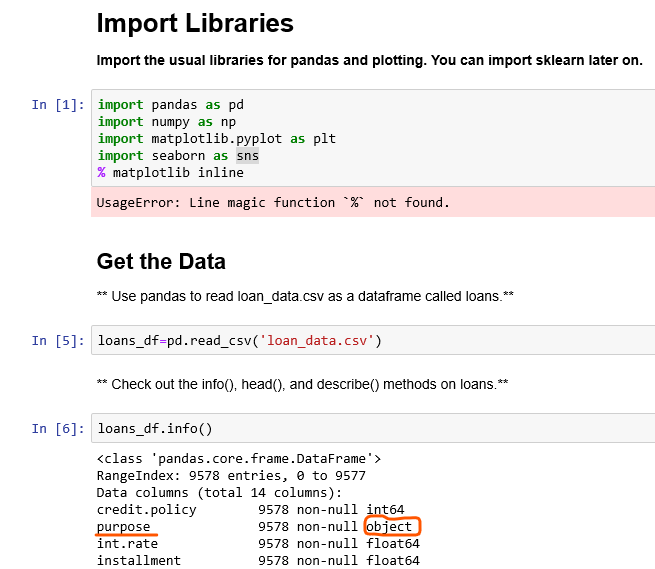
* Creates 2 copies of this graph, separated into 2 columns
* **col1**=not paid back fully (0)

**col2**=paid back fully (1)

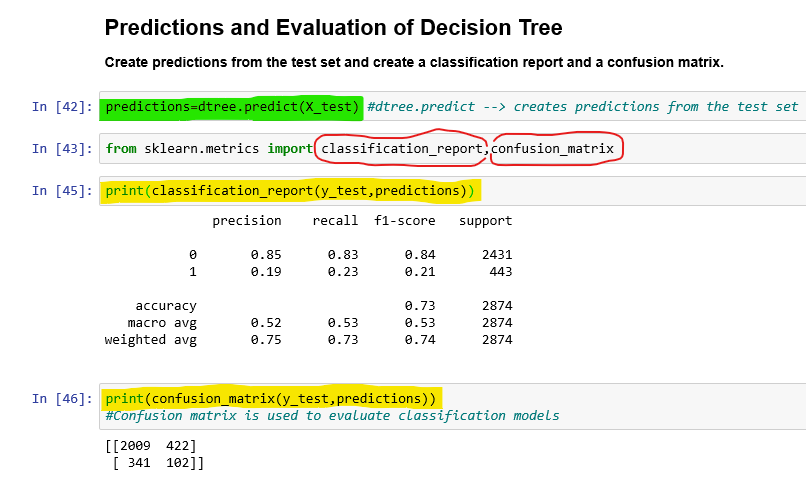
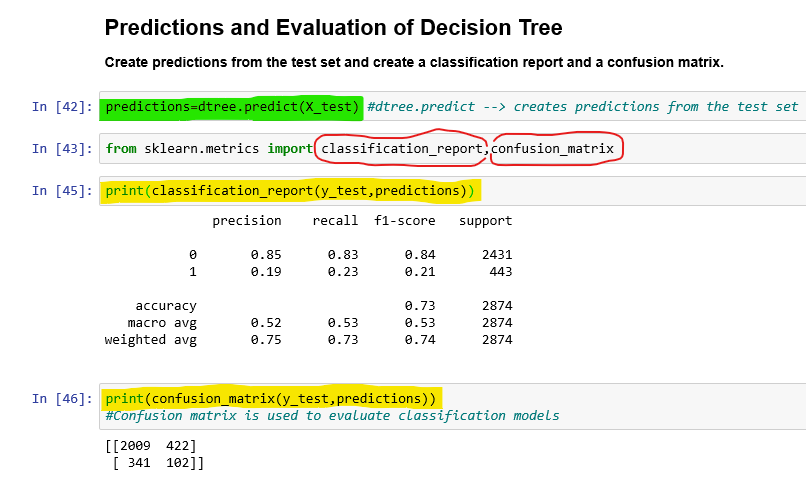
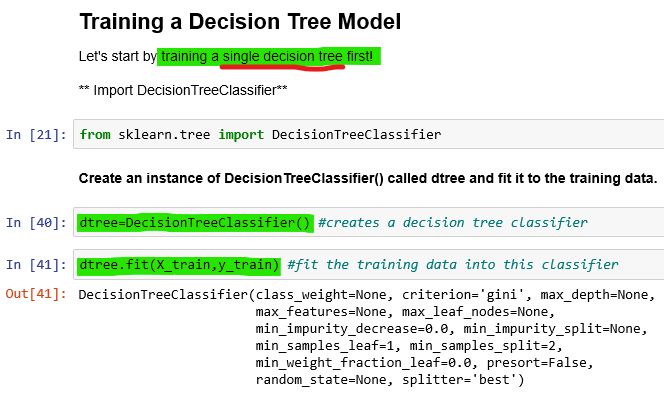
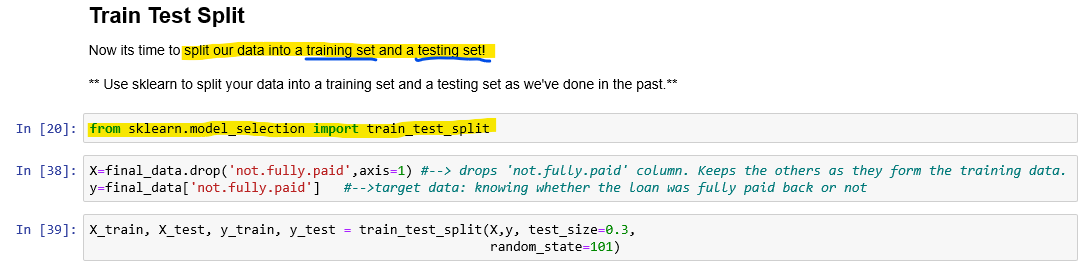
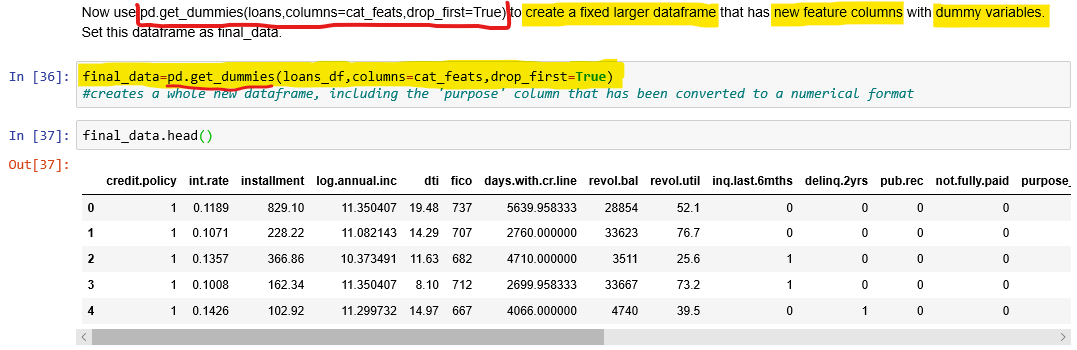
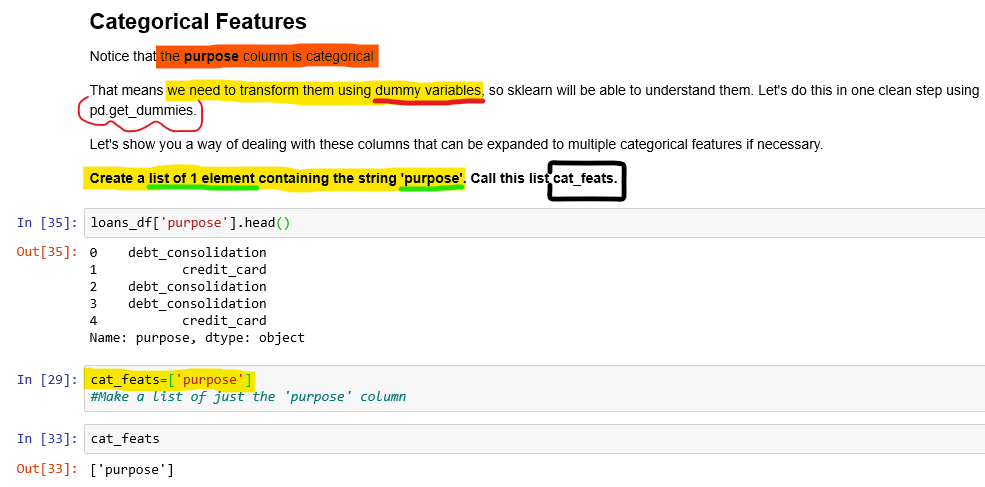
**Setting up the Data**

Let's get ready to set up our data for our Random Forest Classification Model!

**Check loans.info() again.**



**Categorical data**



The number of customers predicted to have **not fully paid back**

The number of customers predicted to have **fully paid back**

Total number of observations

**Macro Average** (overall average):

e.g. Average precision of both outcomes:

**0.52**

**Accuracy:** Overall, how often has this model been correct?

**0.73**

**Weighted Average:** Accounts for the varying importance (number of observations) of each score before averaging them

**0.75**

**Precision:** The proportion of the customers that the model said didn’t pay back fully, actually didn’t (or vice versa)

**0.85**

**0.19**

**Recall:** The proportion of the customers that didn’t pay back fully, that were identified as such by the model (or vice versa)

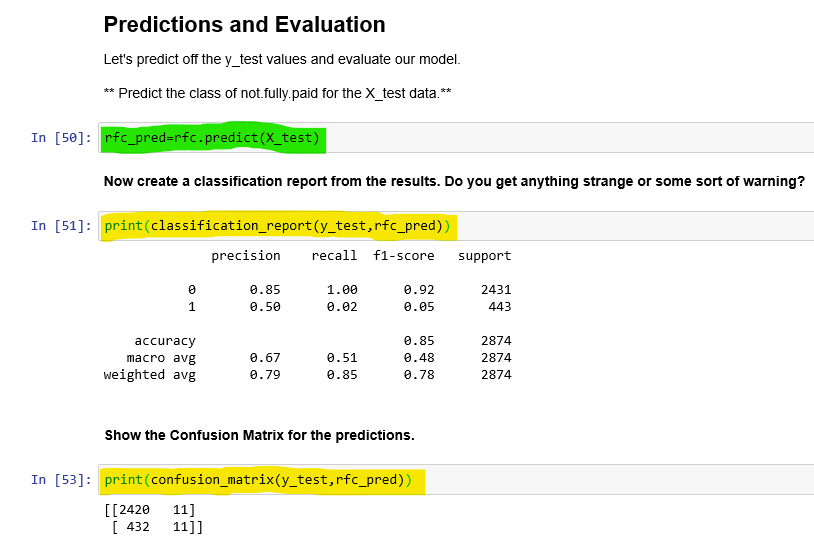
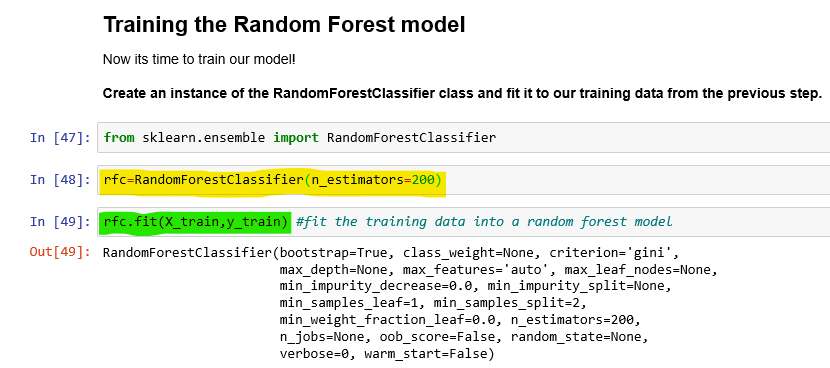
**0.83**

**0.23**

**F-score:** Measures accuracy by balancing **precision** and **recall**

**0.84**

**0.21**



**Weighted Average** is more accurate than the macro average

**0:** Didn’t pay back fully

**1:** Paid back fully

**What model performed better? the random forest or the decision tree?**

* The weighted averages of precision, recall and f1; performed better under the random forest model
* **F1 score:** for (0), the rfc value is more accurate but for (1), the single d-tree is a lot more accurate
* **Recall:** for (0), the rfc value is more accurate but for (1), the single d-tree is a lot more accurate
* **Precision:** for (0), both the rfc and d-tree values are the same, but for (1), the rfc value if is more precise than the d-tree value
* The overall accuracy of the random forest model was 85%, compared to the single decision tree’s 73%, meaning that the rfc’s predictions were correct 12% more of the time, therefore it can be concluded the random forest performed better overall