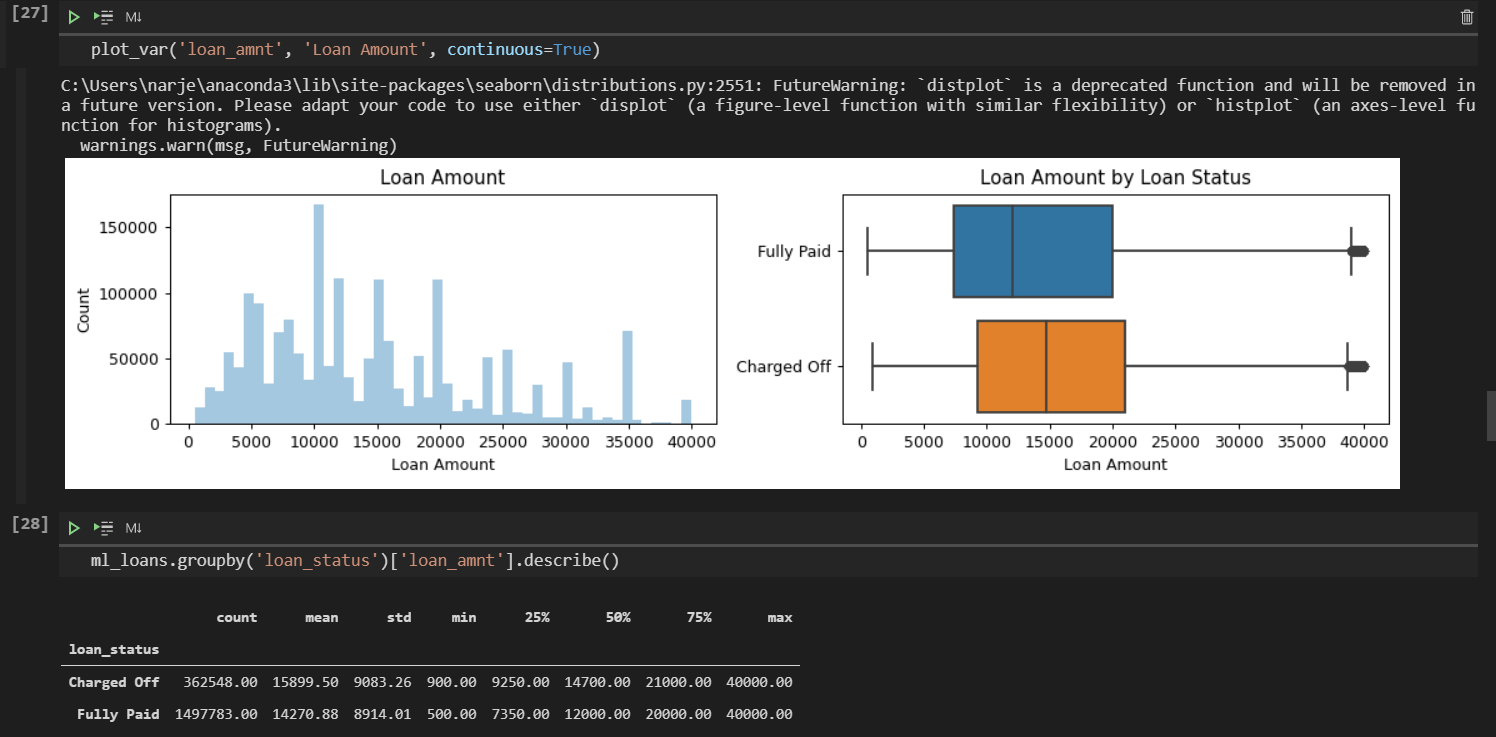


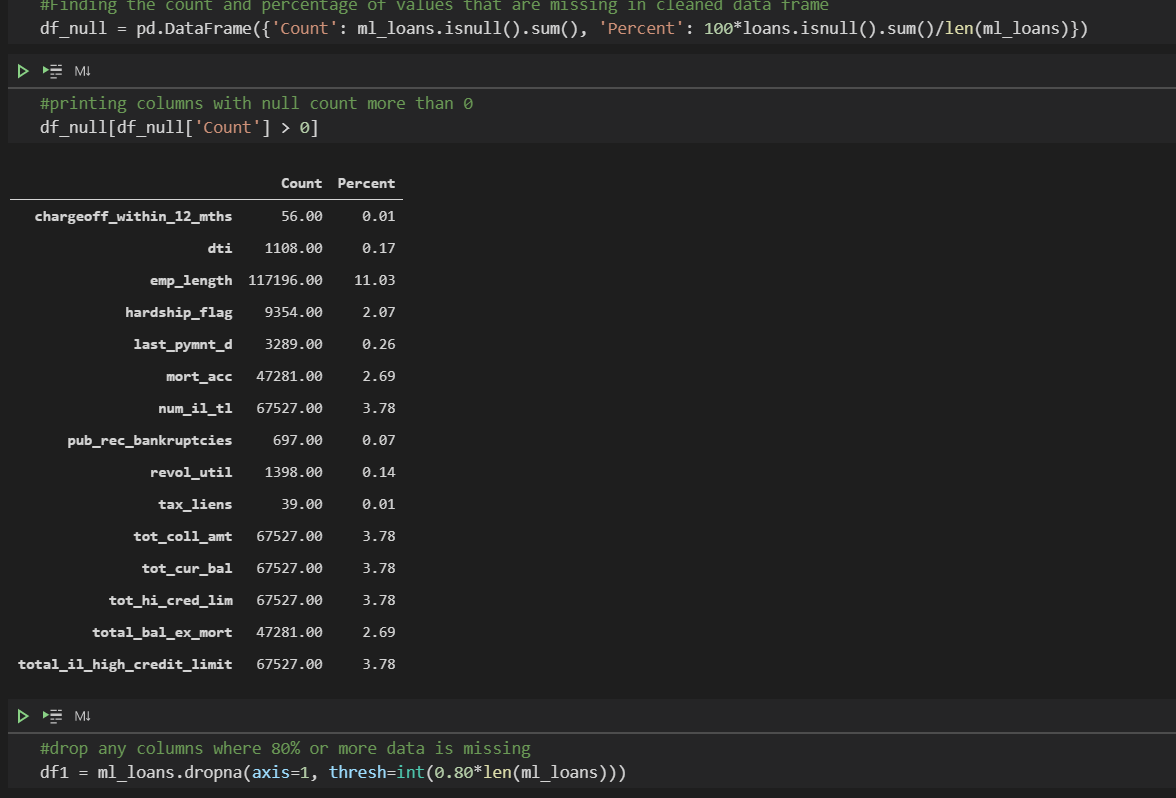
In this data set we have unbalanced data which mean 80% of data belong to Fully Paid class and %20 belongs to Charged off.

We need to check which features are best to be chosen for ML section.

How data looks like for 2 mentioned classes

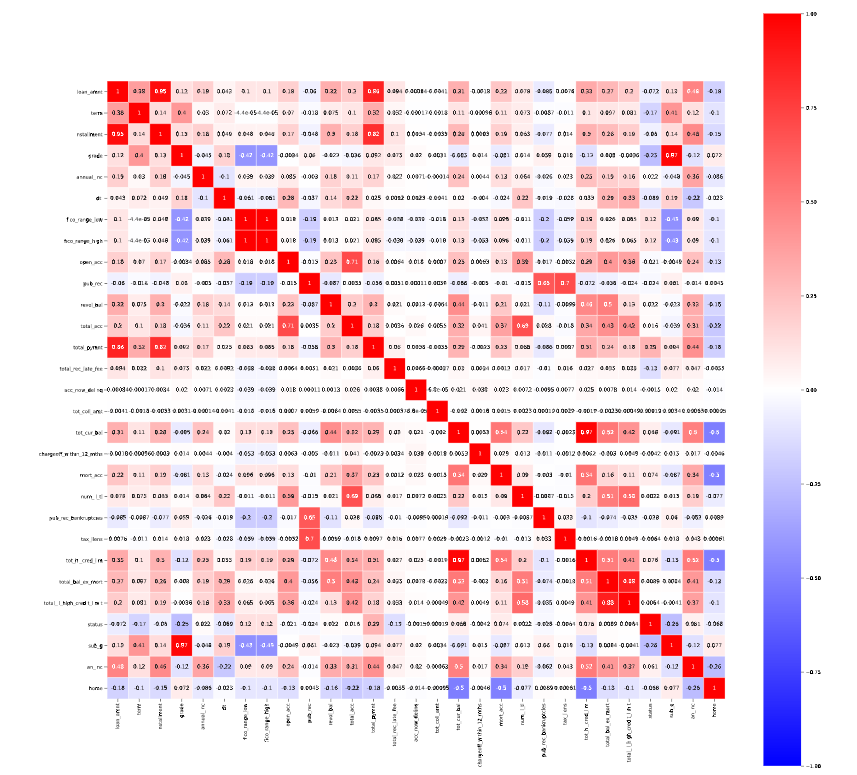


Find and drop any columns where 80% or more data is missing



Finding the correlation between variables

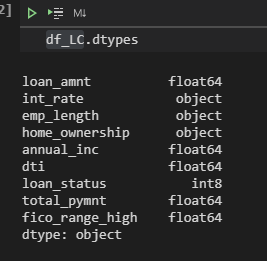
We will now look at the correlation structure between our variables that we selected above. This will tell us about any dependencies between different variables and help us reduce the dimensionality a little bit more



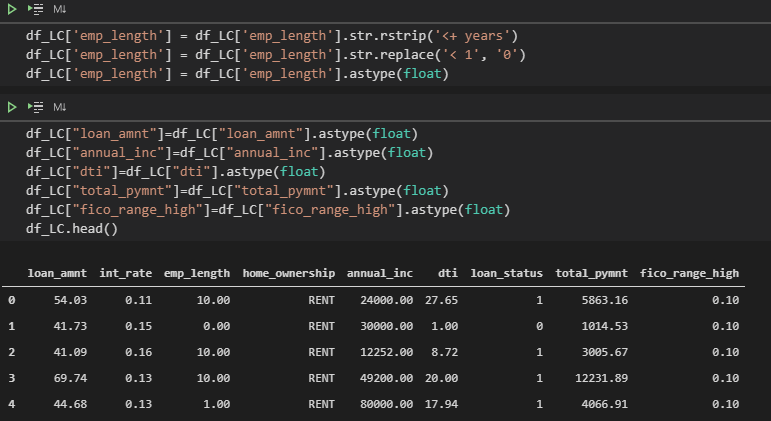
It can be seen from the plot above that loan amount and installment have a very high correlation amongst each other. This is intuitive since a person who takes a large sum of loan would require extra time to repay it back. Also, interest rate, sub grade and grade have a very high correlation between them. This is obvious since interest rate is decided by grades once the grades are decided, a subgrade is assigned to that loan (leading to high correlation).

Choose only features they provide valuable info:

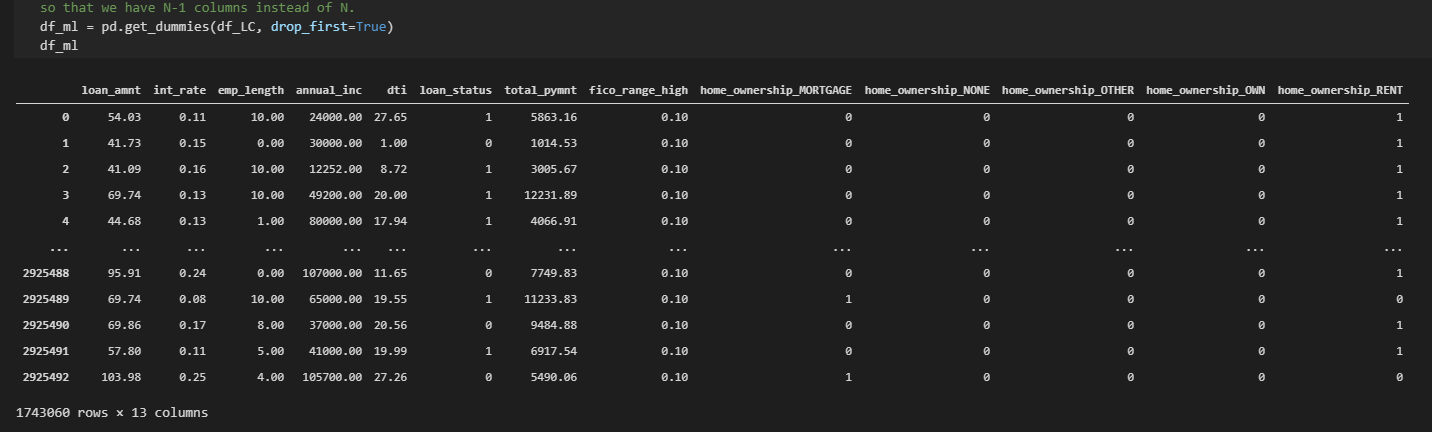




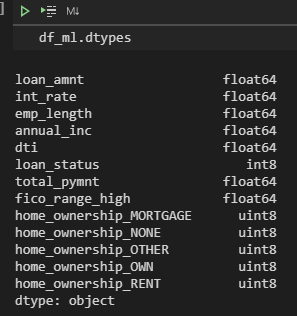
First we need to transfer different data types to numerical data types.



categorical variables for the analysis and the machne learning algorithms doesn't take categorical and string variables directly, we have to creat dummy variables for them it would be wrong in our analysis since a lot of these variables have multiple categories. Just using weights can cause discrepencies in the algorithm. Instead, we will one hot encode these so that we have a 1 wherever that category turns up and 0 otherwise. This will also create seperate columns for each level of category. Also, we'll be dropping one of the categories so that we have N-1 columns instead of N.



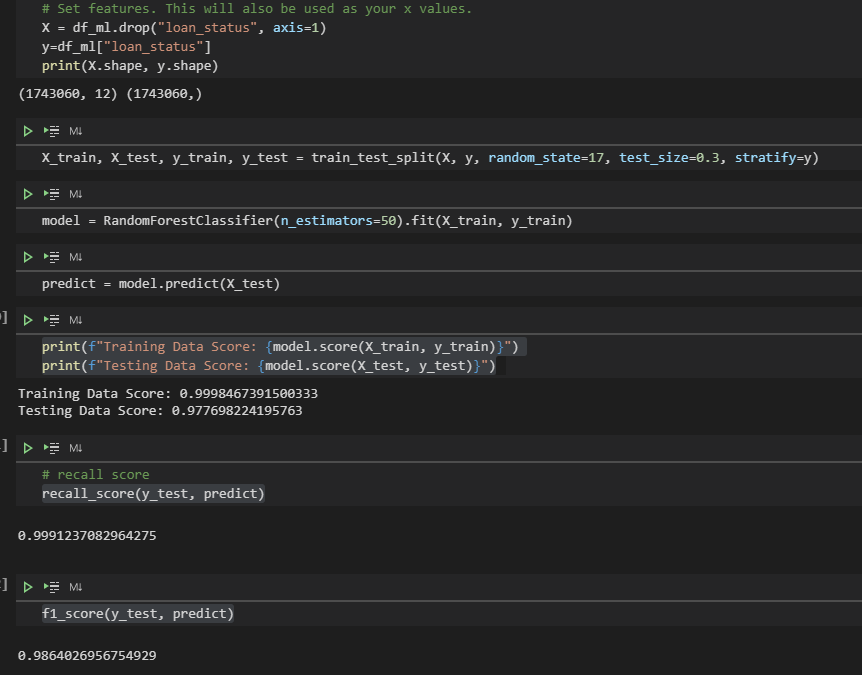
Final data types:



Using Random Forest Calssifier:

We decide to use Random forest specifically for this inbalanced data set and check model performance.

**Imbalanced data**



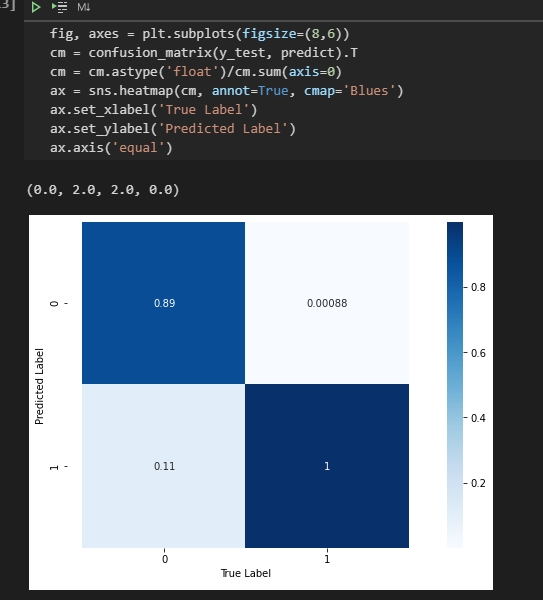
**Performance metrics for imabalanced data sets:**

Confusion Matrix: a table showing correct predictions and types of incorrect predictions.

Precision: the number of true positives divided by all positive predictions. Precision is also called Positive Predictive Value. It is a measure of a classifier’s exactness. Low precision indicates a high number of false positives.

Recall: the number of true positives divided by the number of positive values in the test data. Recall is also called Sensitivity or the True Positive Rate. It is a measure of a classifier’s completeness. Low recall indicates a high number of false negatives.

F1: Score: the weighted average of precision and recall.

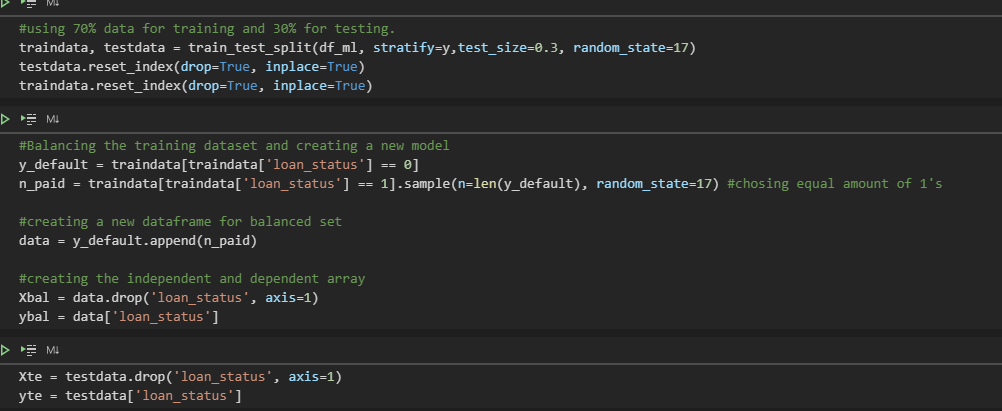


confusion matrix:

The diagonal elements represent the number of points for which the predicted label is equal to the true label, while off-diagonal elements are those that are mislabeled by the classifier. The higher the diagonal values of the confusion matrix the better, indicating many correct predictions.

Balanced data:

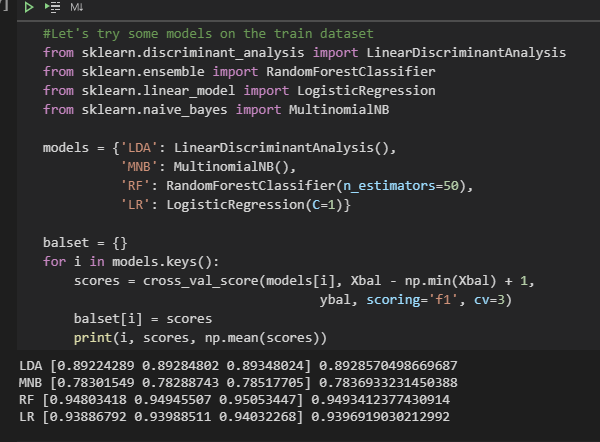
Used equal number of training data for both loan status 0 and 1.



Results after using balanced data:



**Compare models:**

****

Linear Discriminant Analysis

A classifier with a linear decision boundary, generated by fitting class conditional densities to the data and using Bayes’ rule.

The model fits a Gaussian density to each class, assuming that all classes share the same covariance matrix.

The fitted model can also be used to reduce the dimensionality of the input by projecting it to the most discriminative directions, using the transform method.

Naive Bayes classifier for multinomial:

The multinomial Naive Bayes classifier is suitable for classification with discrete features (e.g., word counts for text classification). The multinomial distribution normally requires integer feature counts. However, in practice, fractional counts such as tf-idf may also work.

