Loan Fulfillment

A report by Jorge Londono

Problem Statement

- Bank loans that are unsecured can be very risky for lenders. If the borrower does not fully pay the loan, the lender can end up losing money.

- It is very important for lenders to predict whether a loan applicant will be able to fully pay the loan or not.

- What are the best predictors we can find to foresee the if someone will default or charge off the loan?

Data

Origin of Data

- This data was provided by a kaggle user who collected the data from all of the Lending Club loans from 2007 to 2018.

- The data contains over 150 variables for each loan and its corresponding borrower.

Types of Variables

Categorical Variables:

- Loan Terms
- Application Types

Continuous Variables:

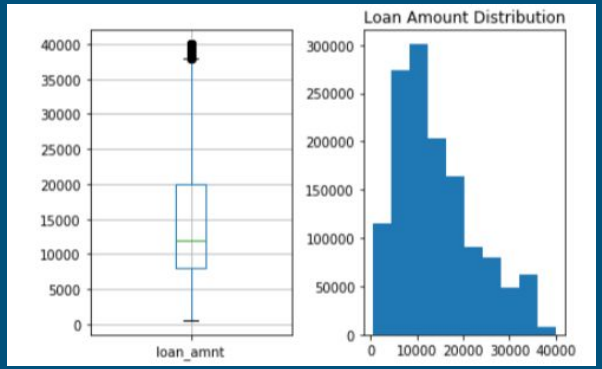
- Loan Amounts
- Annual Income

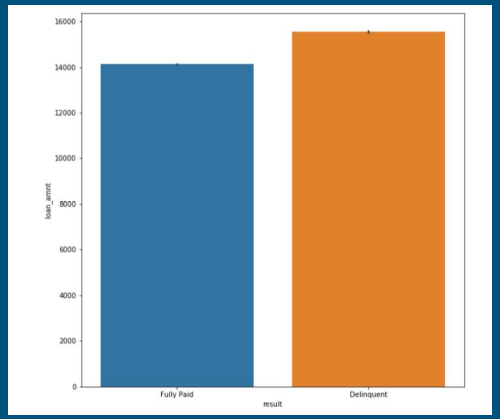
Data Wrangling

- Data is so large that it needs to be loaded in by chunks
- Loans that are not Fully Paid, Default, or Charged Off are removed.
- We will create a new column called result which will have 1 of 2 inputs: Fully Paid, or Delinquent.
- Our Data Frame has a total of 1,346,111 observations

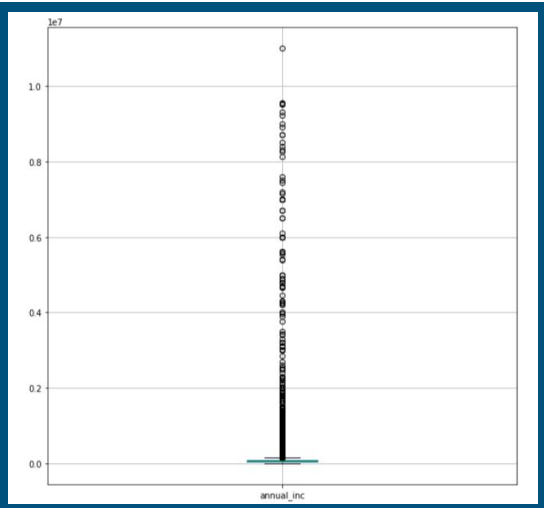
	id
result	
Delinquent	269360
Fully Paid	1076751

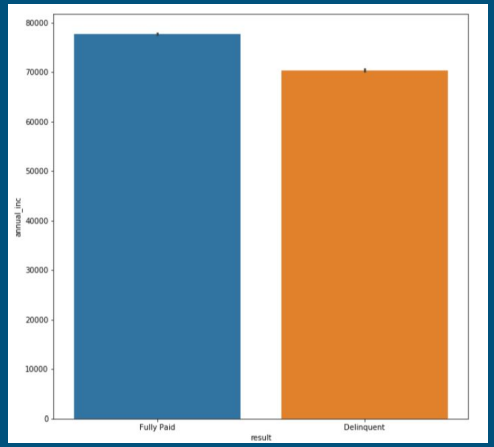
Exploratory Data Analysis - Continuous Variables





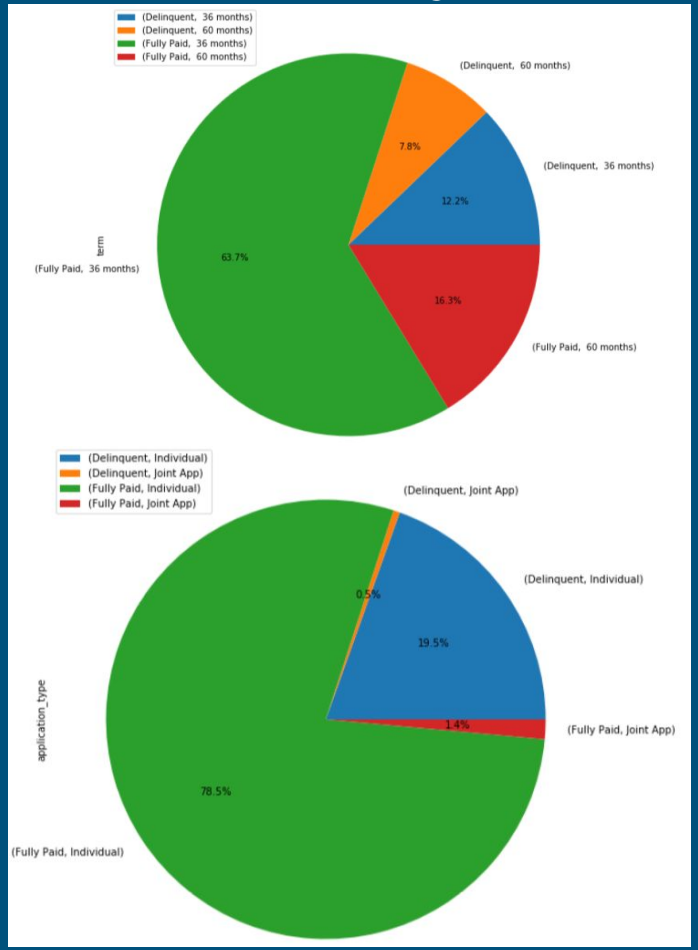
- The first row of graphs show the total loan amount.
- The data skews to higher loan amounts.
- The average loan amount is greater for delinquent accounts.





- The second row of graphs show the annual income of borrower
- The data skews to higher annual income
- The average annual income is greater for fully paid accounts accounts.

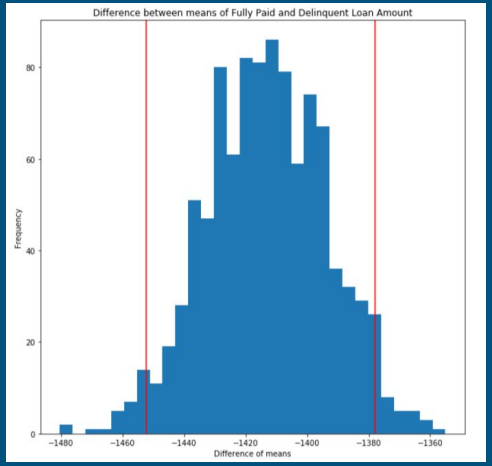
Exploratory Data Analysis - Categorical Variables



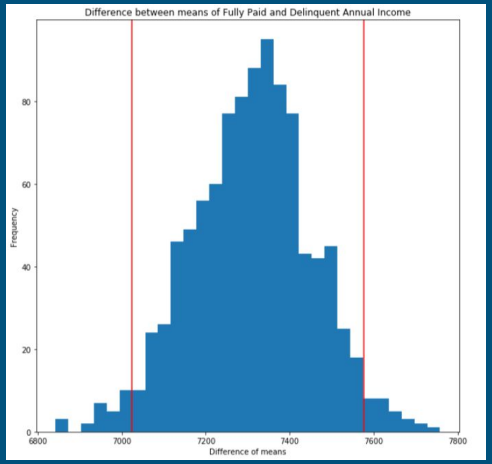
- From this pie chart we can see that the proportion of delinquent loans versus fully paid loans is ls larger when there are 60 month payment terms than 36 month payment terms.

- This pie chart shows us that around ⅓ joint applicants end up with delinquent loans while about ¼ individual applicants end up fully paying off the loans. This shows us that joint applicants are more likely to not fully pay their loan.

Bootstrapping



- Null Hypothesis: There is no difference in the mean loan amount of fully paid and delinquent groups
- Alternate Hypothesis: There is a difference in the 2 means
- Because this interval does not include 0, at 95% significance we can reject the null hypothesis



- Null Hypothesis: There is no difference in the mean annual income of fully paid and delinquent groups.
- Alternate Hypothesis: There is a difference in the 2 means
- Because this interval does not include 0, at 95% significance we can reject the null hypothesis

Algorithm Selection

- This is a classification problem since we are trying to predict a label with our features.

- Random Forest and Gradient Boost are two classification algorithms that will be optimal for this project as ensemble models usually have great results, and we do not have that many features to include.

Procedures

- Create a testing and training set
- Test different models and compute AUC scores
- Compare and select the best scoring model

Features

- Application Type
- Term
- DTI (deb to income)
- Fico Range Low
- Home Ownership
- Annual Income
- Loan Amount

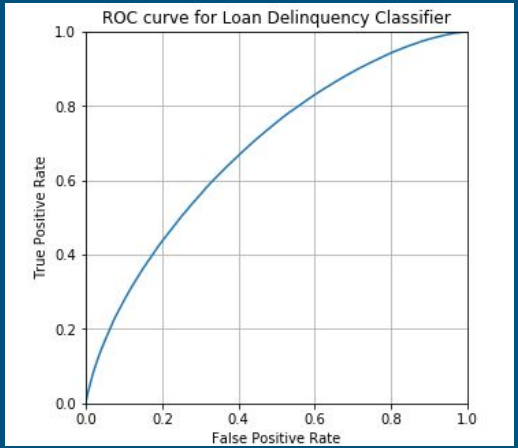
Results

Model	AUC Score
Test 1 Algorithm: Random Forest Classifier Parameters: Default Transformation: None	0.638
Test 2 Algorithm: Random Forest Classifier Parameters: Default Transformation: PCA	0.557
Test 3 Algorithm: Random Forest Classifier Parameters: n_estimators: 100 Transformation: None	0.640
Test 4 Algorithm: Gradient Boost Classifier Parameters: learning_rate: 0.5 Transformation: none	0.688

The best performing algorithm is Test 4: Gradient Boost with Learning rate set to 0.5

Results - Cont.

ROC Curve



Confusion Matrix

```
[[426489   3670]
 [103566   4570]]
```

Other Metrics

	precision	recall	f1-score	support
0.0	0.80	0.99	0.89	430159
1.0	0.55	0.04	0.08	108136
accuracy			0.80	538295
macro avg	0.68	0.52	0.48	538295
weighted avg	0.75	0.80	0.73	538295

Recommendations

- Joint applications are highest predictor of loan delinquency. Scrutinize these applications more.
- Fico score and debt to income ratio are also great predictors of borrowers not being able to fully pay the loan.

	Feature	importance
5	application_type_Joint App	0.381939
3	fico_range_low	0.235659
2	dti	0.123480
0	annual_inc	0.072181
1	loan_amnt	0.053636

Areas of Further Study

- Using application data to figure out who is most likely to be denied.
- Looking at small business loan defaults and what types of ventures are likely to have financial problems and why.