Case Study 2: Lending Club

Team 4: Emily Strong and Raksha Kaverappa

Main Code Repository: <https://github.com/erstrong/INFO-7390-ADS-Fall-17-TeamNo.4/tree/master/Assignment%202>

**Part 1A: Data Wrangling**

*Code:*

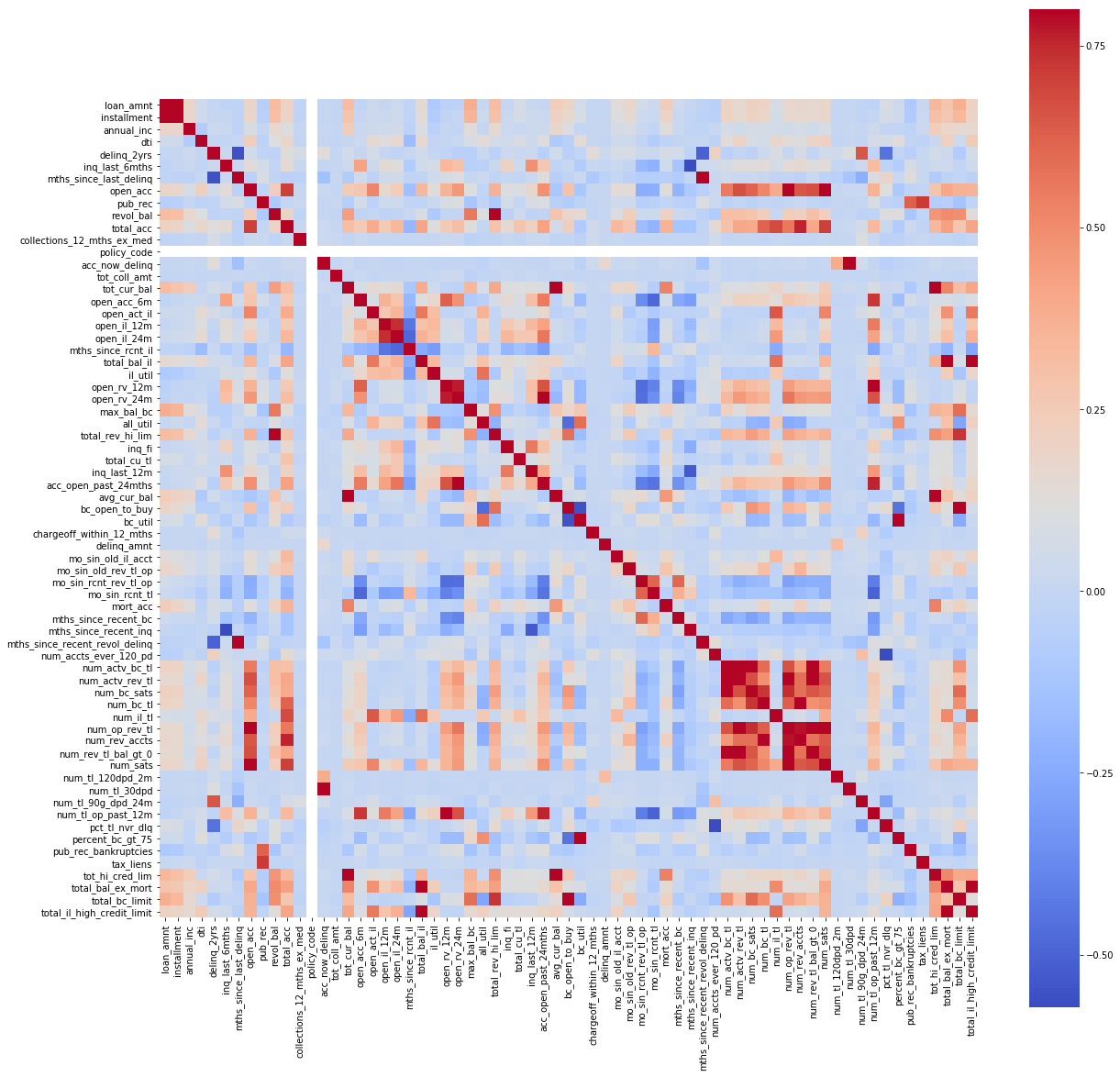
[ADD LUIGI NOTES]

To scrape the approved loan data:

* We created a session with Lending Club investor credentials.
* We scraped from a hidden div on the page the names of the zip files that need to be accessed, and then read the csvs in each zip file into a data frame.
* We then merged the data frames and added a timestamp and the set name.
* We dropped all columns with >50% missing data.
* We removed columns that are computed by Lending Club after the loan application based on the data dictionary.
* We used a correlation heat map (shown below) and the data dictionary to eliminate redundant columns.
* We did missing value substitutions using mean, median, and max for annual\_inc, mths\_since\_recent\_inq and mths\_since\_last\_delinq.
* We dropped any rows with missing values for the remaining features, cumulatively removing 6% of the data.
* We reformatted numeric features present as strings, and created dummy variables for the categorical features.

To scrape the declined data:

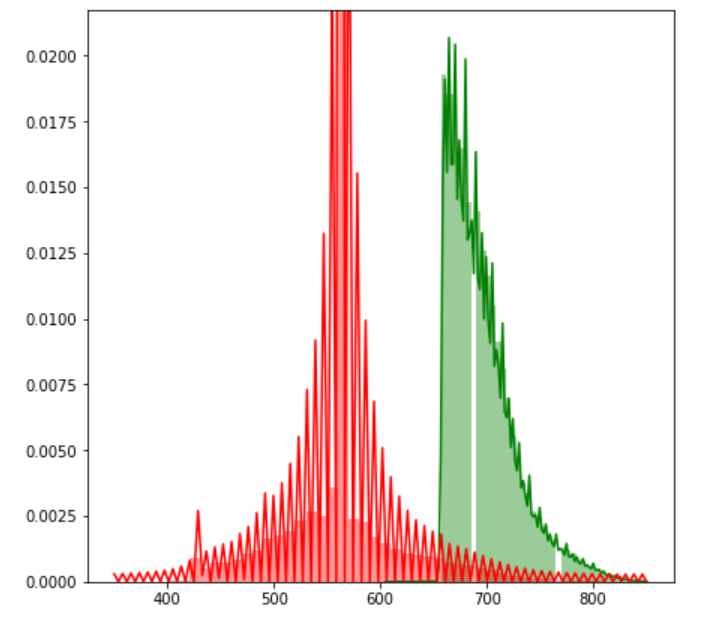
* We followed a similar process but without the authenticated session.
* The only feature with >50% missing data was Risk Score, with 58%. We chose to keep it because of the critical
* For Risk Scores, we converted Vantage scores (applications dated on or after November 6, 2013) to use the same scale as FICO scores.
* We dropped rows with missing location information, and used mean and mode substitutions for the remaining features.

****

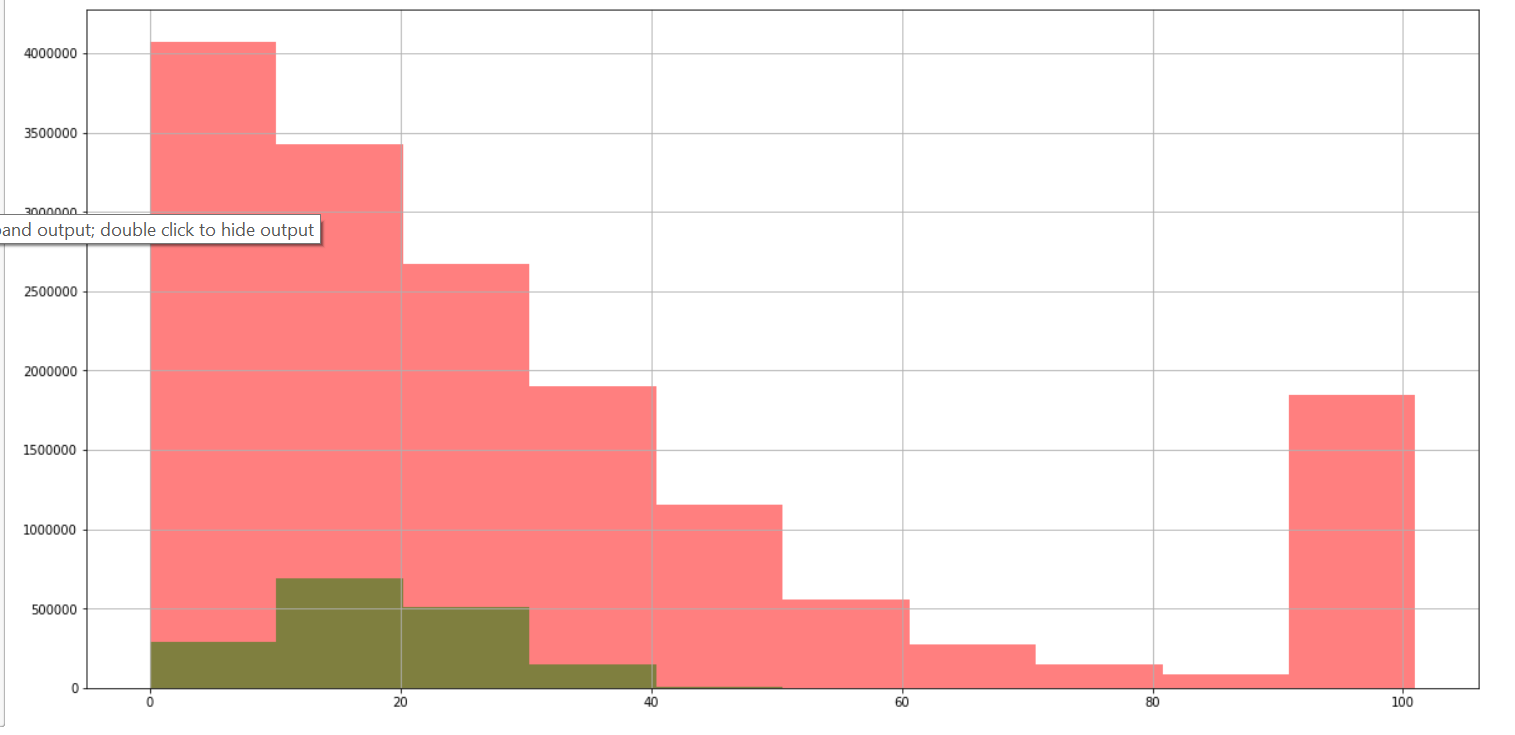
**Correlation Heat Map for Loan Features with <=50% missing data**

**Part 1B: Exploratory Data Analysis**

* For the first part of our analysis, we made a comparison study of the FICO score distributions of the accepted and declined datasets. We observed that most of the accepted FICO scores were below 650 approximately. The graph is as shown below:

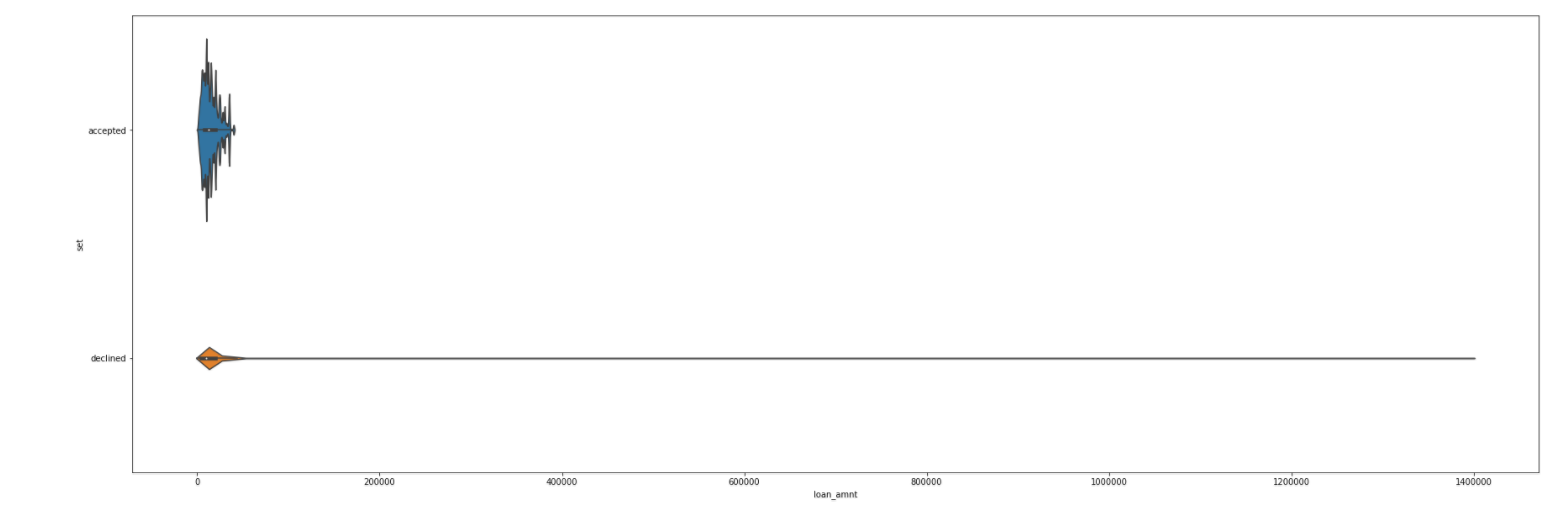


* The Debt to Income ratio distribution for the declined and the approved data is as shown below.



We notice that the declined data has higher DTI when compared to approved data.

* The distribution of loan amounts is as shown below:



We observe that there are several outliers for the declined data.

**Part 2A: Classification**

We combined the approved and declined datasets and used the following features to classify the model:

* Loan amount
* FICO score
* Debt to Income ratio
* Employment length
* Policy code

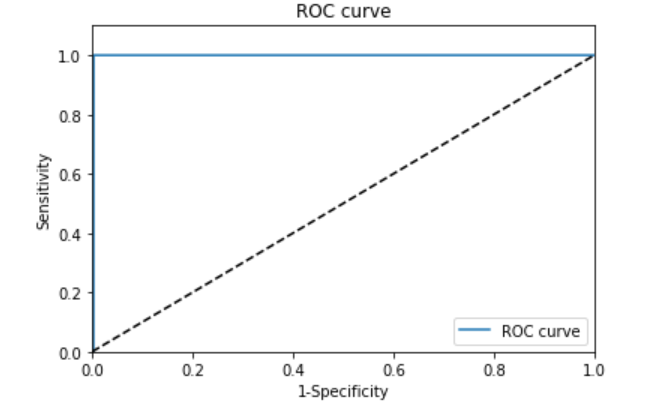
The confusion matrix and ROC curve for all the models shown below:

Confusion Matrix:

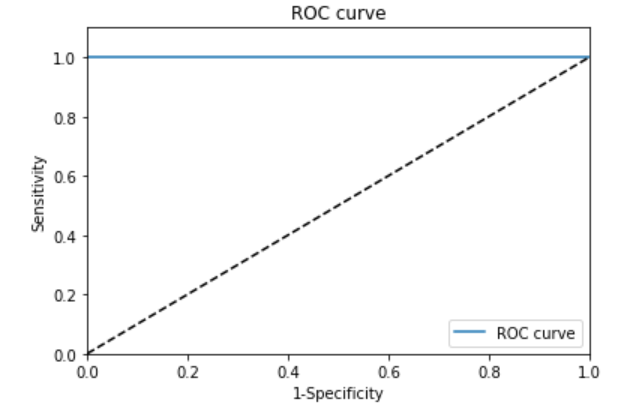
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **True Positives** | **False Positives** | **True Negatives** | **False Negatives** |
| Logistic Regression | 3216939 | 8949 | 302817 | 108 |
| Random Forest | 3226328 | 0 | 302485 | 0 |
| Neural Networks | 3224991 | 0 | 303822 | 0 |

ROC curve:

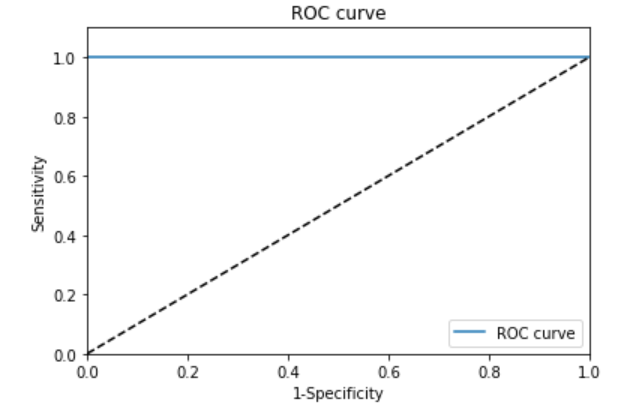
Logistic Regression:



Random Forest:



Neural Networks:



**Part 2B: Clustering**

**Part 2C: Prediction**

**Part 2D: Deployment**