Exploratory Analysis of Mortgage Dataset:

Visualizations and Classification Results

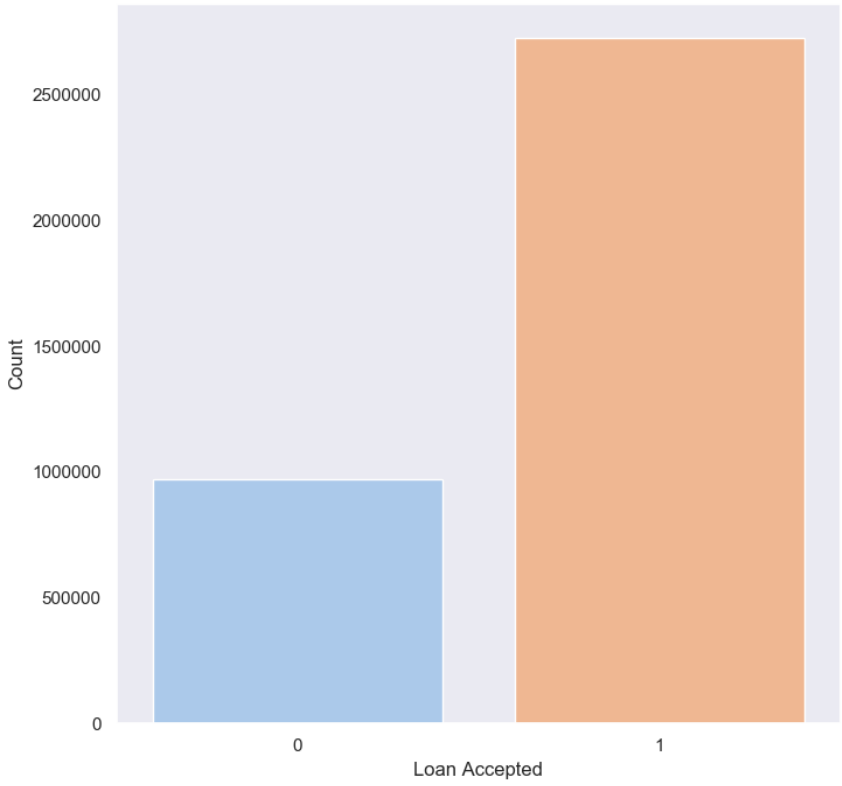
Data Description:

The data used in this report was gathered from public repository created after the Home Mortgage Disclosure Act. It contains 3.5 million data points detailing numerous categorical and numerical attributes from loan applications in the year 2017. This project focused only on applications for mortgage refinancing. Only the features that had a correlative value off > 10% we're used to train the machine learning models.

Visuals:

Ratio of Approved Loans and Rejected Loans

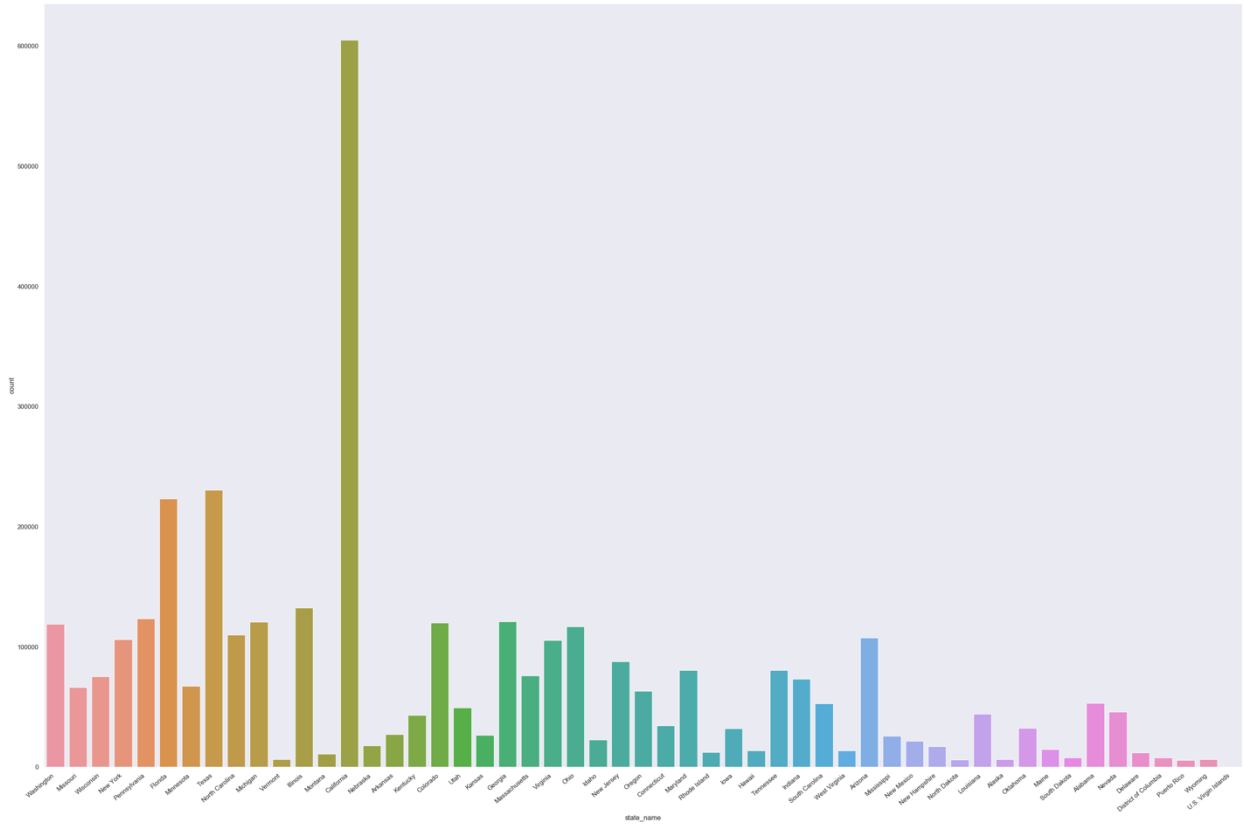
In these models we want to predict whether or not somebody will be approved for a refinancing loan. The first step is to plot the ratio of accepted vs rejected loans to get a better understanding of the distribution



As shown above, roughly 64% of loans were approved in 2017 while 36% were denied

Ratio of Loan Applications by State

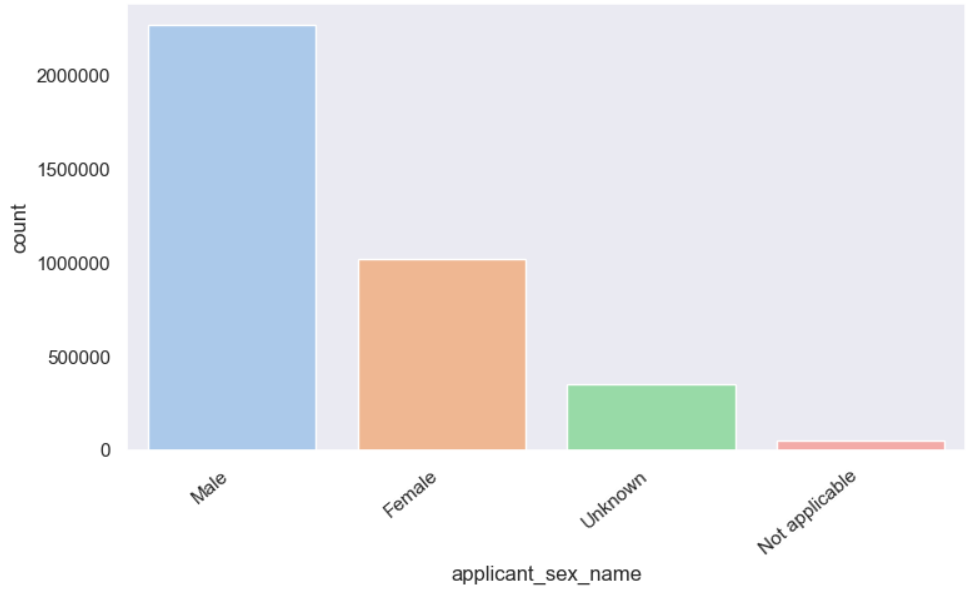
Here we can see the amount of mortgage applications by State. California outpaced the rest of country by a pretty wide margin while Vermont and Wyoming saw very little mortgage activity.



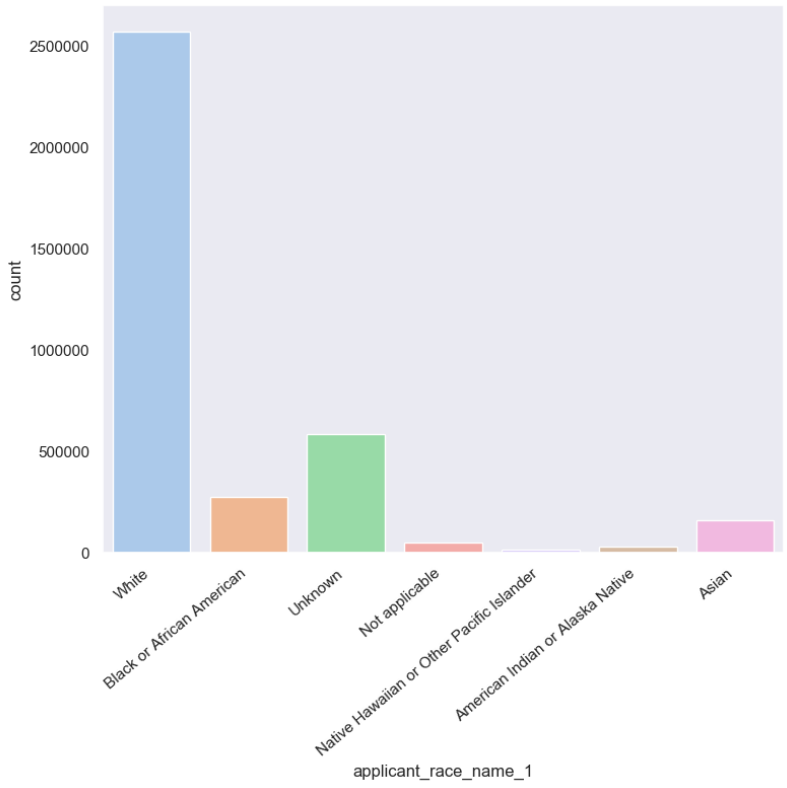
Demographics

Unfortunately, this dataset had very little personal information about each applicant. Variables such as **age**, **education**, and **credit history** may have had a significant impact in the models' predictive abilities.

Applicant Gender



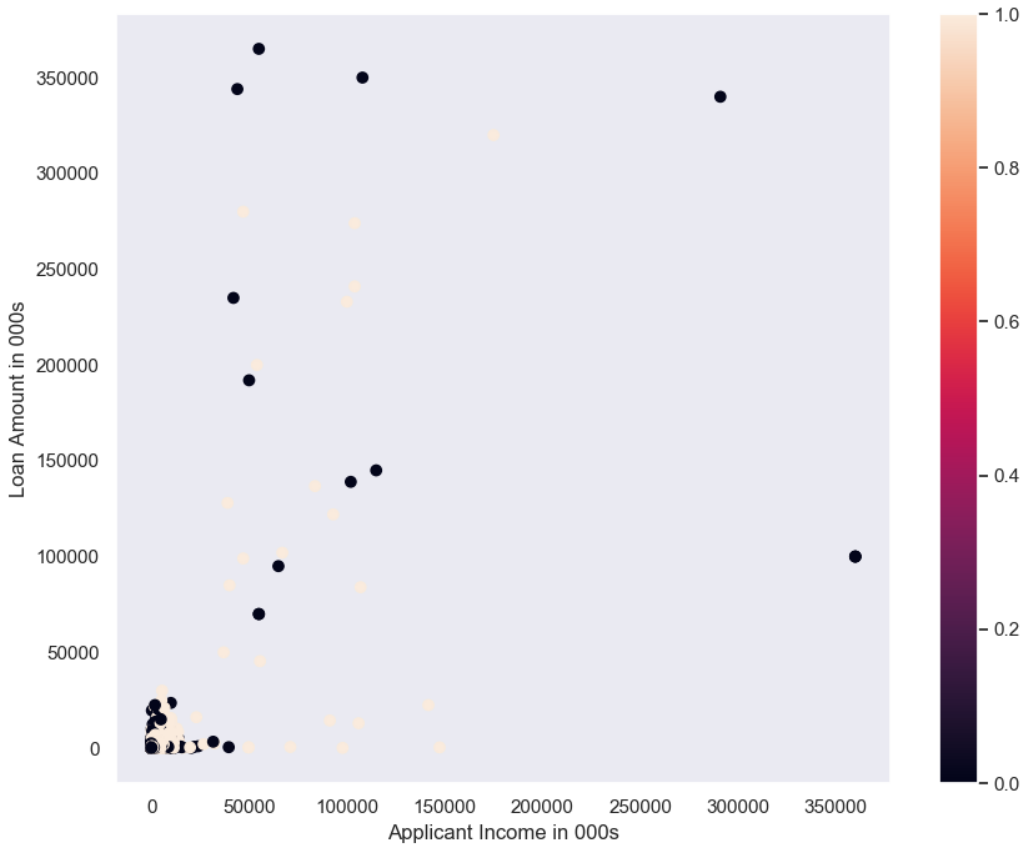
Applicant Ethnicity



Income to Loan Request Scatter Plot

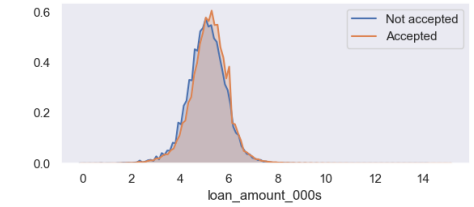
1 = Approved | 0 = Denied

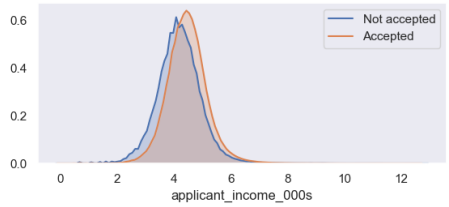
This is a scatter plot that maps the applicant's income with their loan request. Yellow dots represent an approved application. Purple dots represent a rejected application.

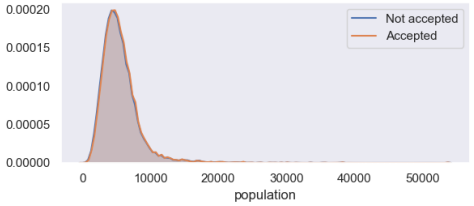


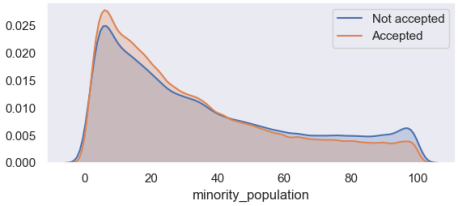
KERNEL DENSITY ESTIMATES

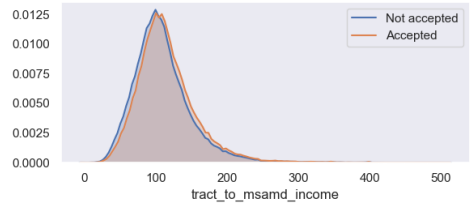
Kernel density estimation is a way to estimate the probability of a variable at a certain instance. The plots below show the density of loan acceptance based on all of the numerical features in the dataset. This gives a visual representation of what factors affect loan acceptance the most.

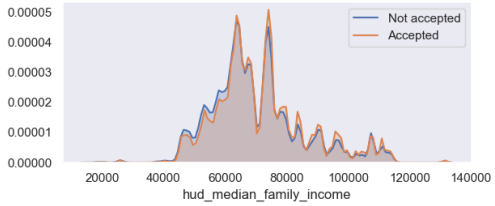


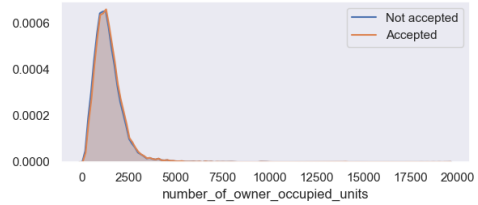


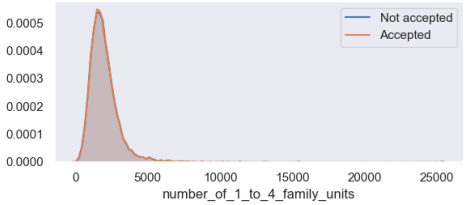


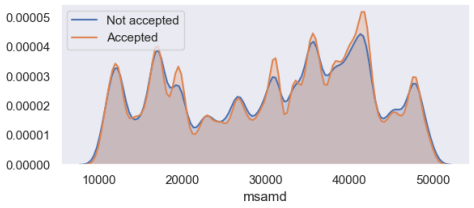


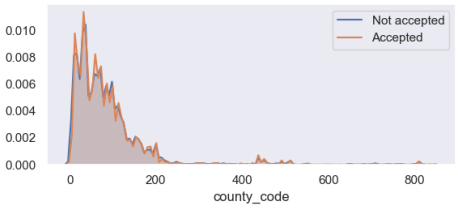




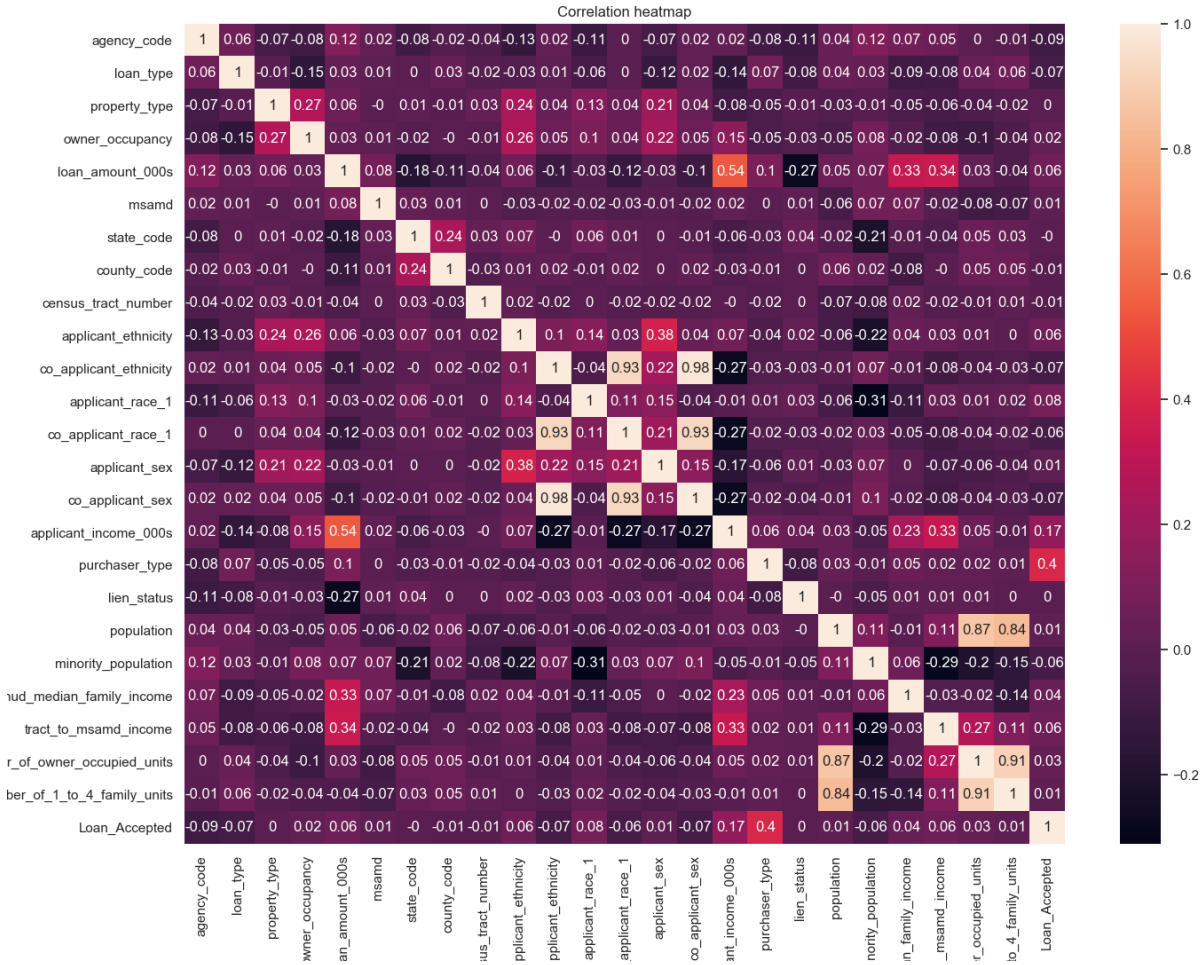








Correlation Matrix



Model Results

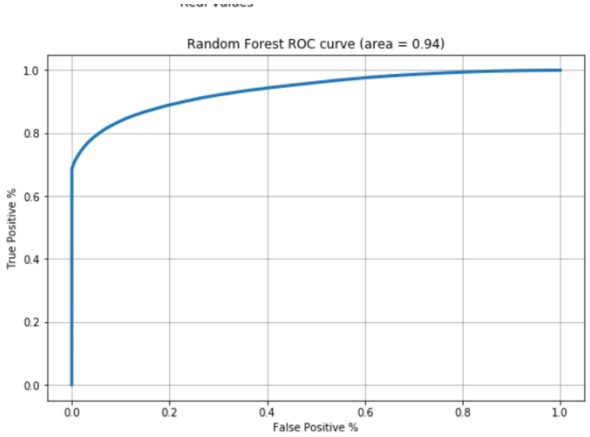
Results of the Random Forest Model

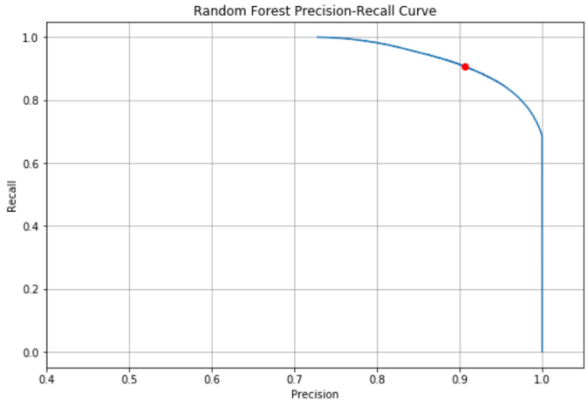
Precision: The model was correct 73% of the time in predicting whether somebody would be rejected for a loan. The model was correct 92% of the time in predicting whether somebody would be approved for a loan.

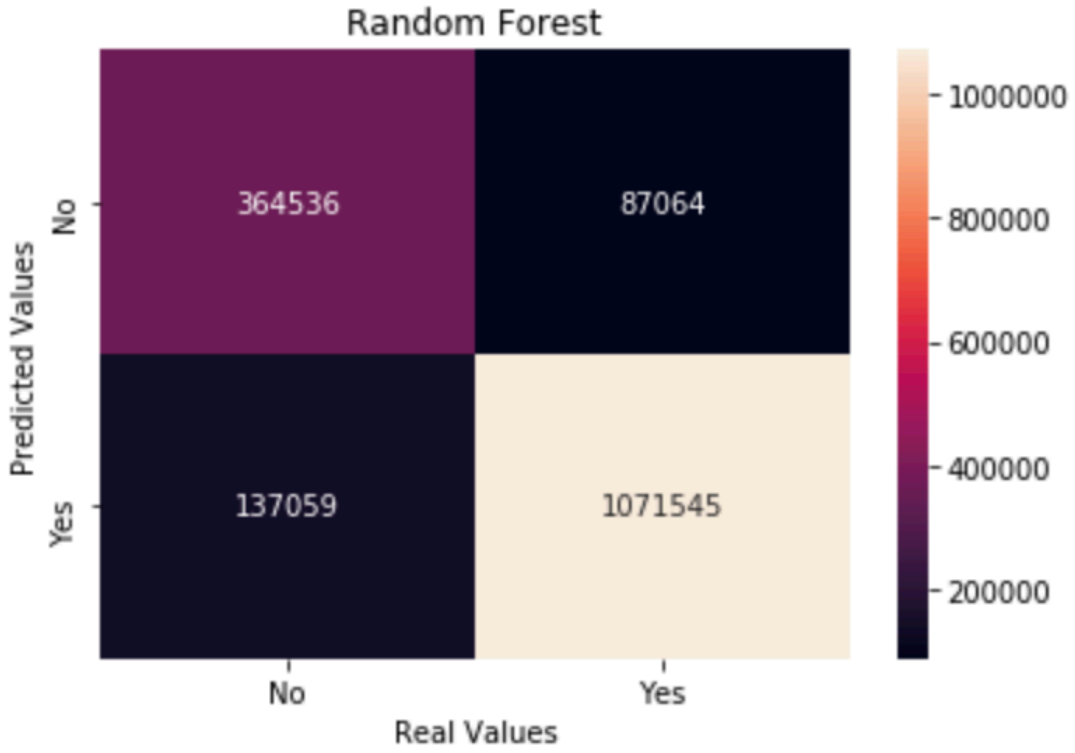
Recall: The model falsely labeled 19% of loan rejections as positive in the dataset. The model falsely labeled detect 11% of loan approvals as negative in the dataset.

True Positives: 364536
True Negatives: 1071545
False Negatives: 137059
False Positives: 87064

	precision	recall	f1-score	support
0	0.73	0.81	0.76	451600
1	0.92	0.89	0.91	1208604
accuracy			0.87	1660204
macro avg	0.83	0.85	0.84	1660204
weighted avg	0.87	0.87	0.87	1660204







The area underneath the **ROC** is equal to 94% of the total area, meaning this model is very accurate. This is a measure only of the model's ability to classify true positives and true negatives. It doesn't measure false negatives or false positives.

The **Precision-Recall** curve measures the trade-off of between Precision and Recall. As shown by the chart, a model with ~99% accuracy on labeling true positives and true negatives will mislabel false positives and false negatives ~40% of the time. A model that doesn't mislabel any false positives or false negatives will mislabel true positives and true negatives ~25% of the time. The ideal trade-off is indicated by the red dot and exists at ~90% Precision with ~85% Recall.

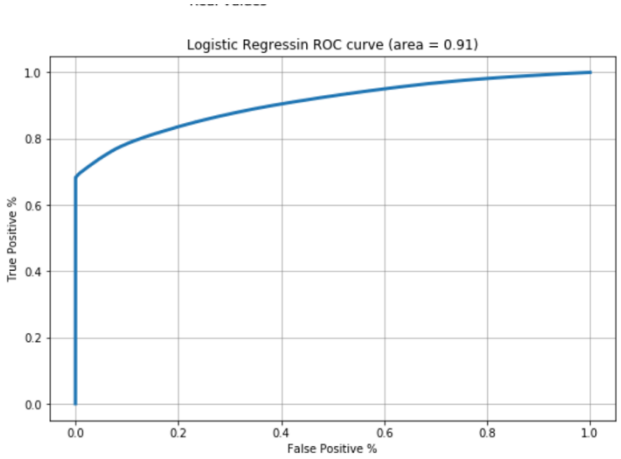
RESULTS OF THE LINEAR REGRESSION MODEL

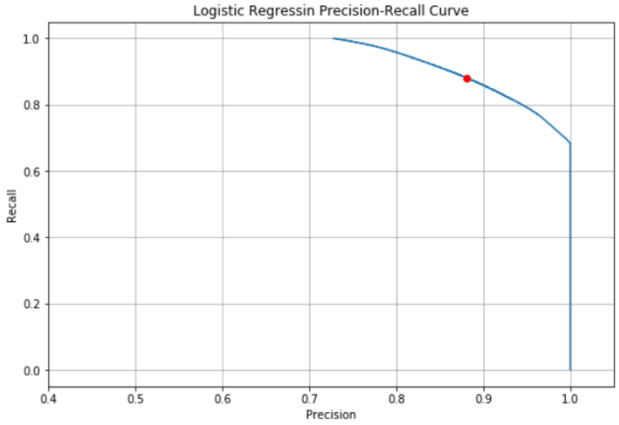
Precision: The model was correct 67% of the time in predicting whether somebody would be rejected for a loan. The model was correct 89% of the time in predicting whether somebody would be approved for a loan.

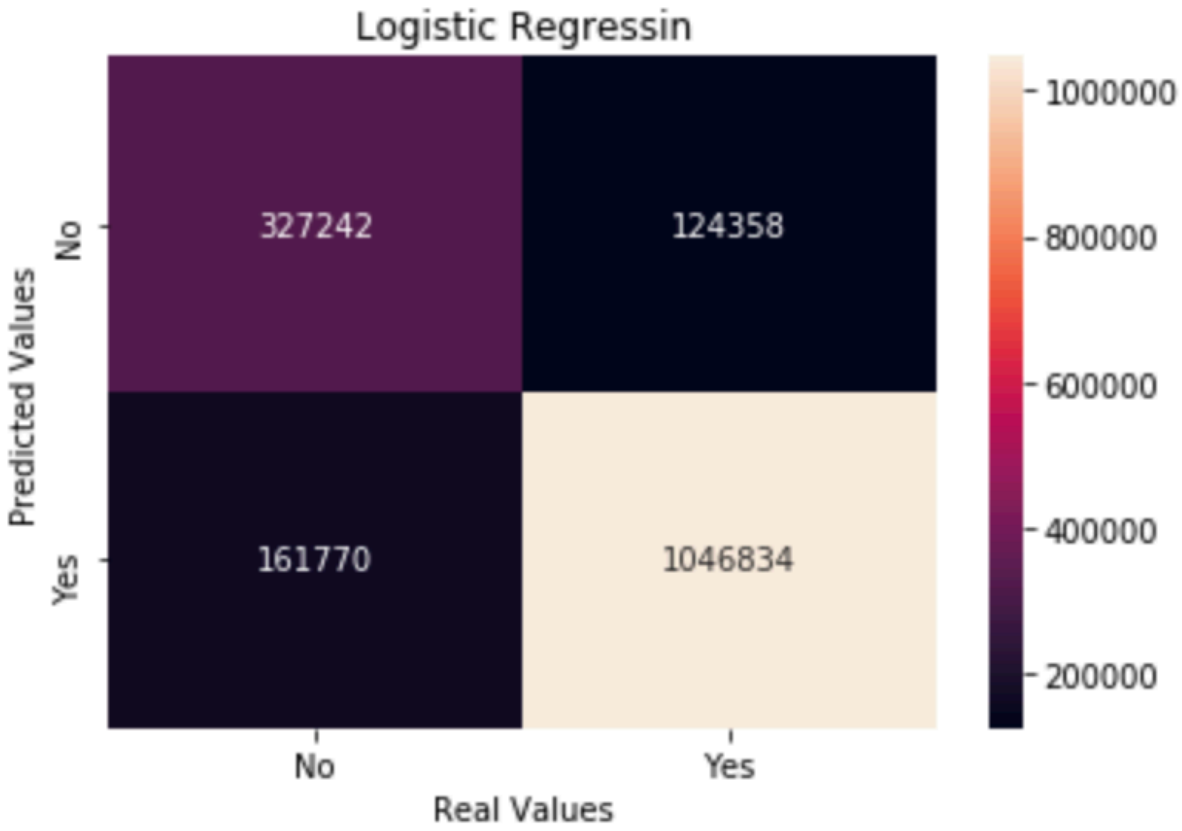
Recall: The model falsely labeled 28% of loan rejections as positive in the dataset. The model falsely labeled detect 13% of loan approvals as negative in the dataset.

True Positives: 364536
True Negatives: 1046834
False Negatives: 161770
False Positives: 124358

0	0.67	0.72	0.70	451600
1	0.89	0.87	0.88	1208604
accuracy			0.83	1660204
macro avg	0.78	0.80	0.79	1660204
weighted avg	0.83	0.83	0.83	1660204







The area underneath the **ROC** is equal to 91% of the total area, meaning this model is very accurate. This is a measure only of the model's ability to classify true positives and true negatives. It doesn't measure false negatives or false positives.

The **Precision-Recall** curve measures the trade-off of between Precision and Recall. As shown by the chart, a model with ~99% accuracy on labeling true positives and true negatives will mislabel false positives and false negatives ~40% of the time. A model that doesn't mislabel any false positives or false negatives will mislabel true positives and true negatives ~25% of the time. The ideal trade-off is indicated by the red dot and exists at ~88% Precision with ~85% Recall.

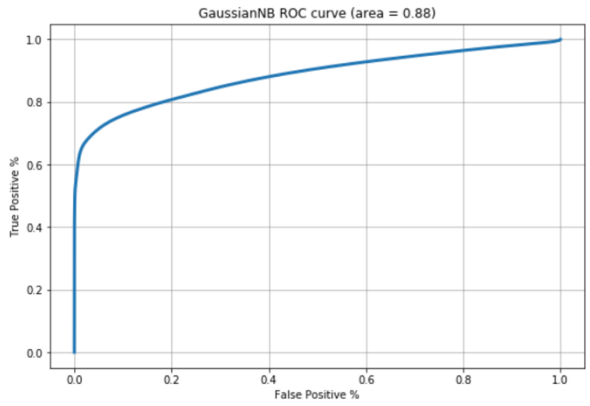
RESULTS OF THE GAUSSIAN NAIVE BAYES MODEL

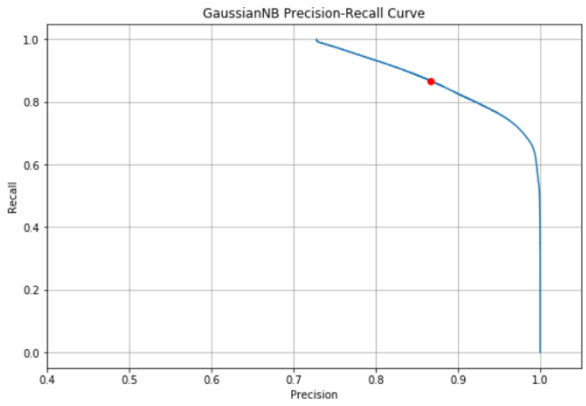
Precision: The model was correct 68% of the time in predicting whether somebody would be rejected for a loan. The model was correct 78% of the time in predicting whether somebody would be approved for a loan.

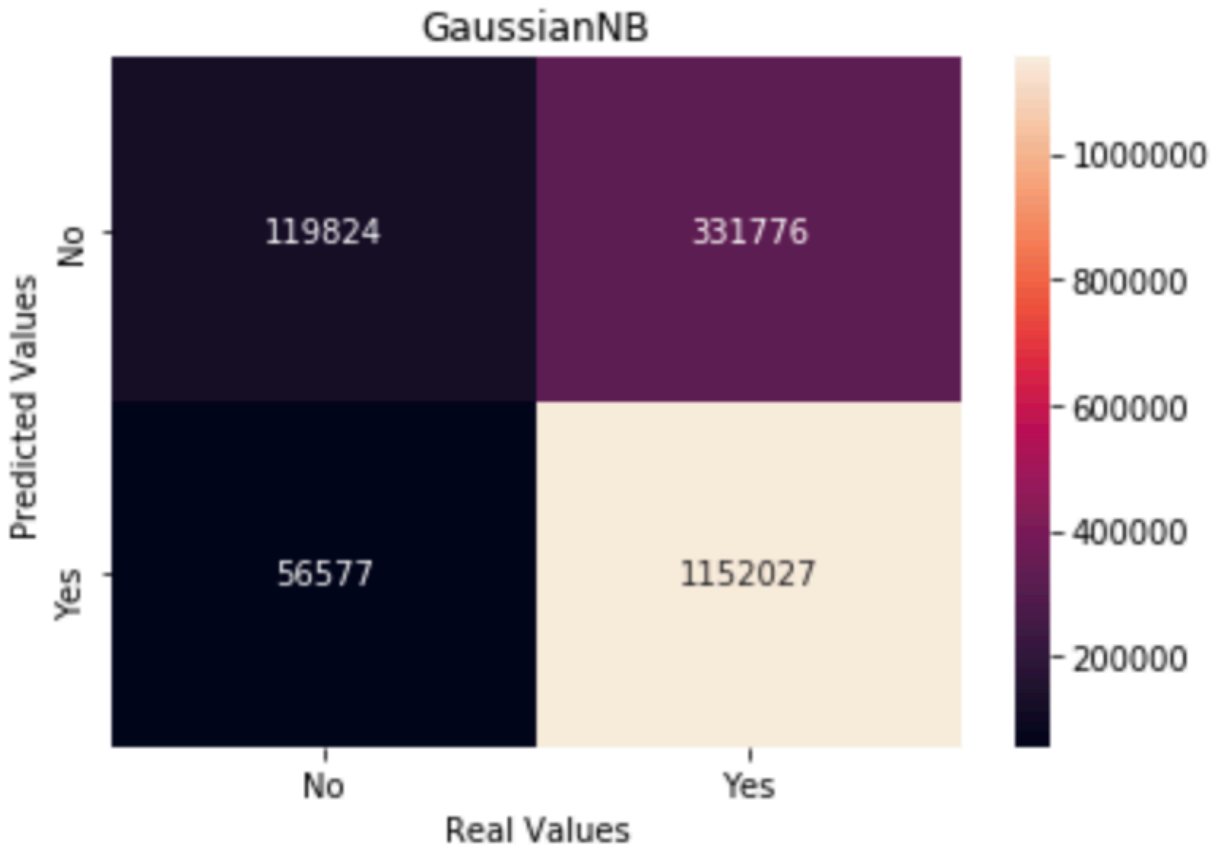
Recall: The model falsely labeled 73% of loan rejections as positive in the dataset. The model falsely labeled detect 5% of loan approvals as negative in the dataset.

True Positives: 119824
True Negatives: 1152027
False Negatives: 56577
False Positives: 331776

	precision	recall	f1-score	support
0	0.68	0.27	0.38	451600
1	0.78	0.95	0.86	1208604
accuracy			0.77	1660204
macro avg	0.73	0.61	0.62	1660204
weighted avg	0.75	0.77	0.73	1660204





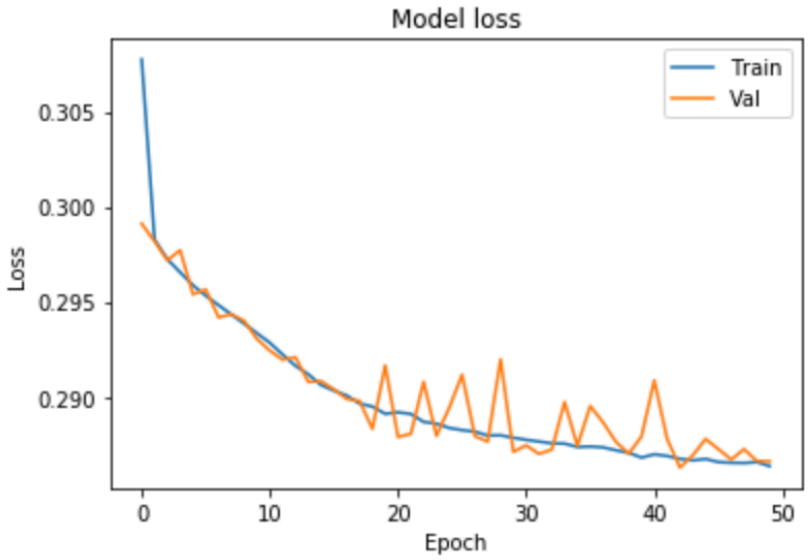


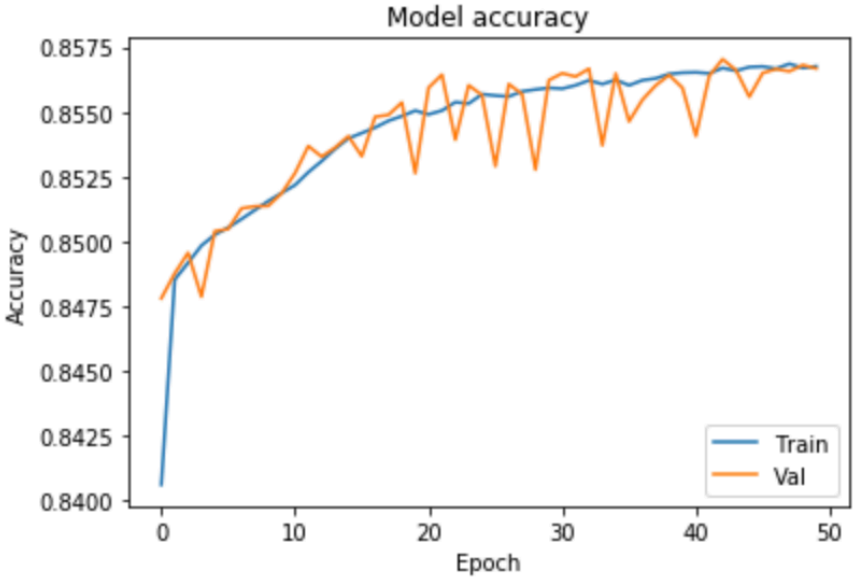
The area underneath the **ROC** is equal to 88% of the total area, meaning this model is very accurate. This is a measure only of the model's ability to classify true positives and true negatives. It doesn't measure false negatives or false positives.

The **Precision-Recall** curve measures the trade-off of between Precision and Recall. As shown by the chart, a model with ~99% accuracy on labeling true positives and true negatives will mislabel false positives and false negatives ~40% of the time. A model that doesn't mislabel any false positives or false negatives will mislabel true positives and true negatives ~25% of the time. The ideal trade-off is indicated by the red dot and exists at ~88% Precision with ~85% Recall.

Neural Network

Result: 85.42%





Above we can see that the Validation data is more erratic than the Training data and does not completely match the line. This is an indication that the model is slightly over-fitting.