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

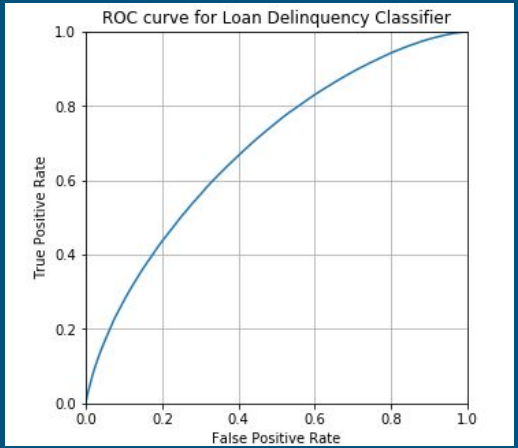
Results

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.