Lending Club Loan Analysis

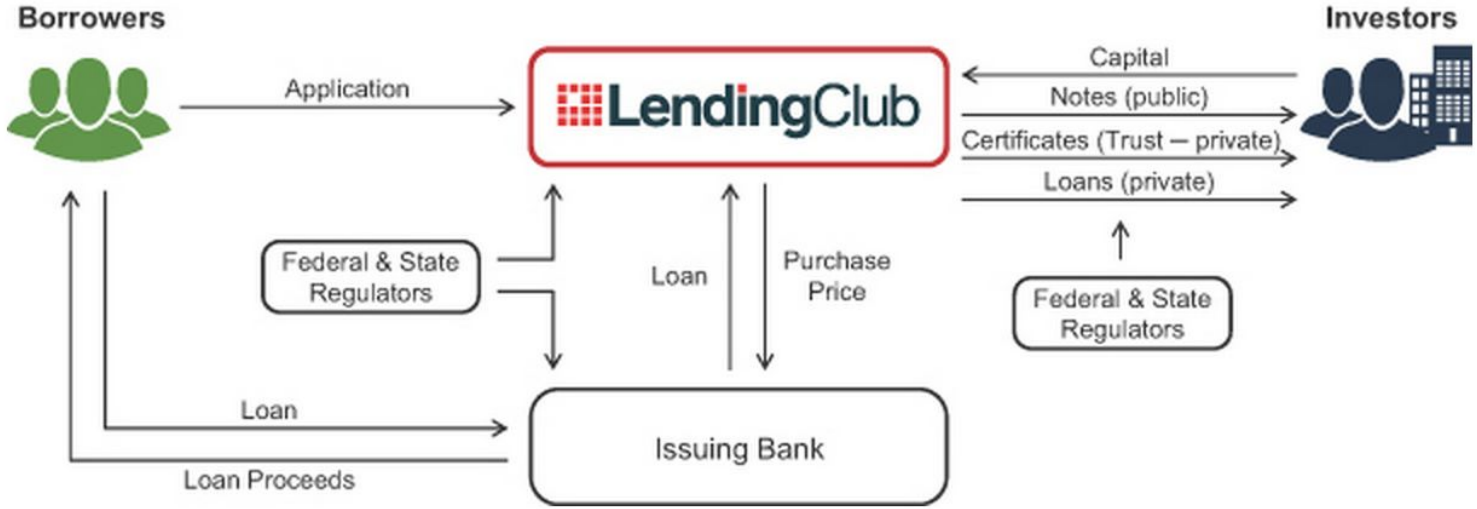
Ajay Kliyara Philip, Willard Williamson

About the Authors

- Will:
  - Software Engineer at Lockheed Martin in Syracuse NY
  - Used R and H2O
- Ajay:
  - Software Engineer at Xerox in Rochester NY
  - Used Python and SciKit Learn

Lending Club Web Site

Loan Issuance Mechanism



Project Goals

1. Make loan marketplace default predictions.
2. Make loan origination default predictions.
3. Explore the Lending Club developer API

First Goal

Loan Marketplace Prediction

- The loan marketplace is a website where notes are bought and sold after loan origination.
- We wrangled the data to take advantage of additional data beyond that which is available at loan origination time.
- Examples of additional data includes things like the borrower’s latest FICO score and total principal paid.

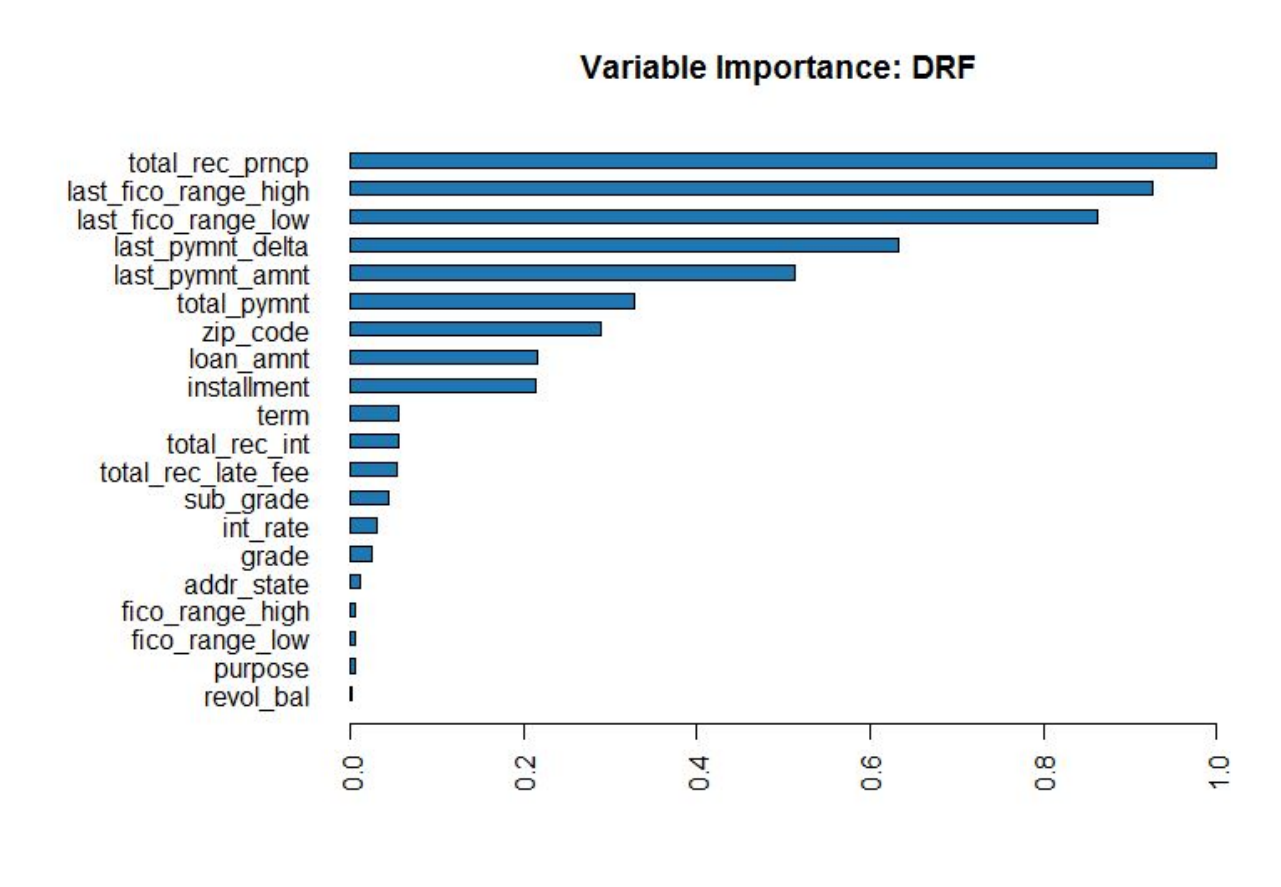
H2O Machine Learning Models

- Random Forest
- Naive Bayes
- Generalized Linear Model
- Neural Network
- Gradient Boosting Machine
- Stacking Ensemble

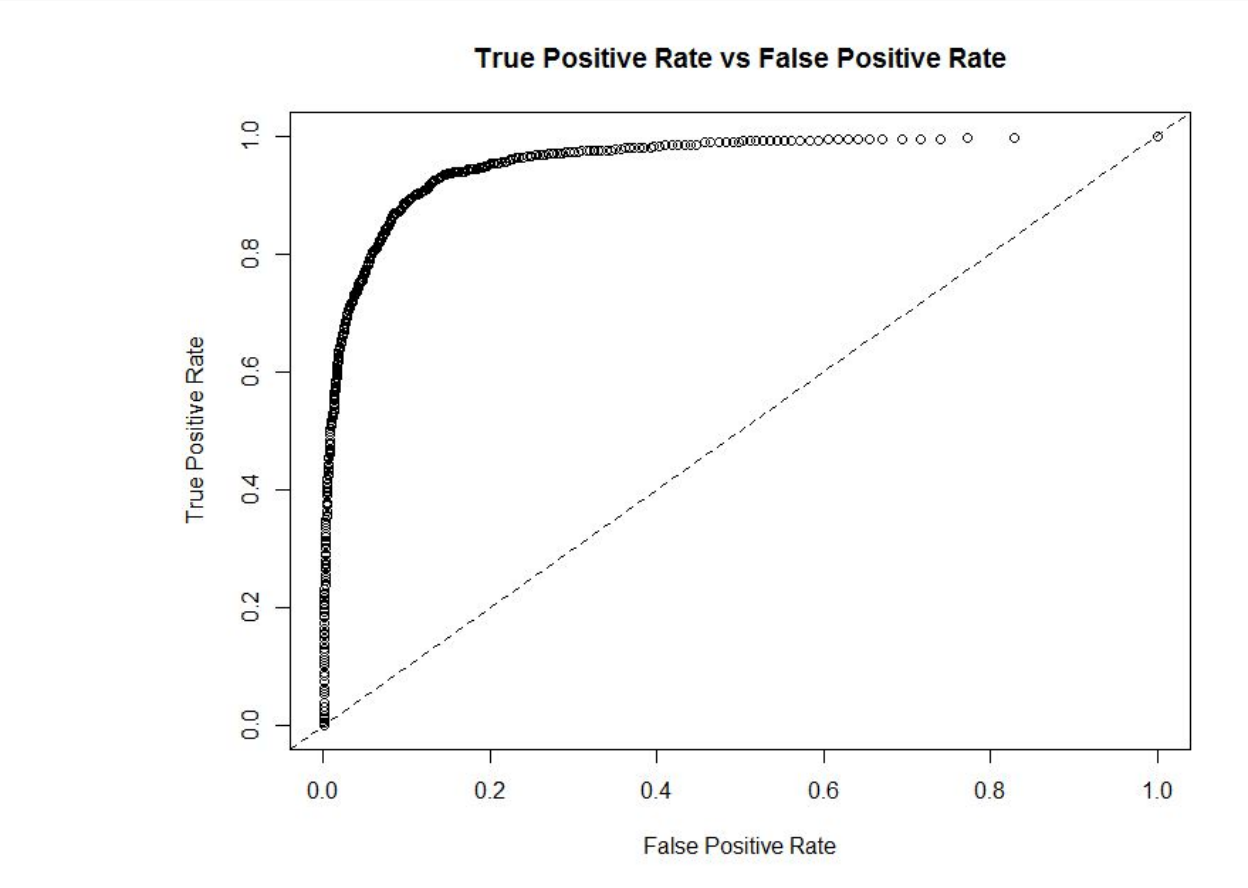
Data Wrangling

- Created an is_bad column from loan default, charge off, > 30 days late.
- Engineered a new feature called last_pymnt_delta = last payment date - loan origination date
- last_pymnt_delta introduced a time element to correlate time with features like last FICO
- last payment delta proved to be a very strong predictor

H2O Loan Marketplace Variable Importance Plot (Random Forest)



H2O GLM Loan Marketplace ROC Curve: AUC = 0.9568
Top 5 Features

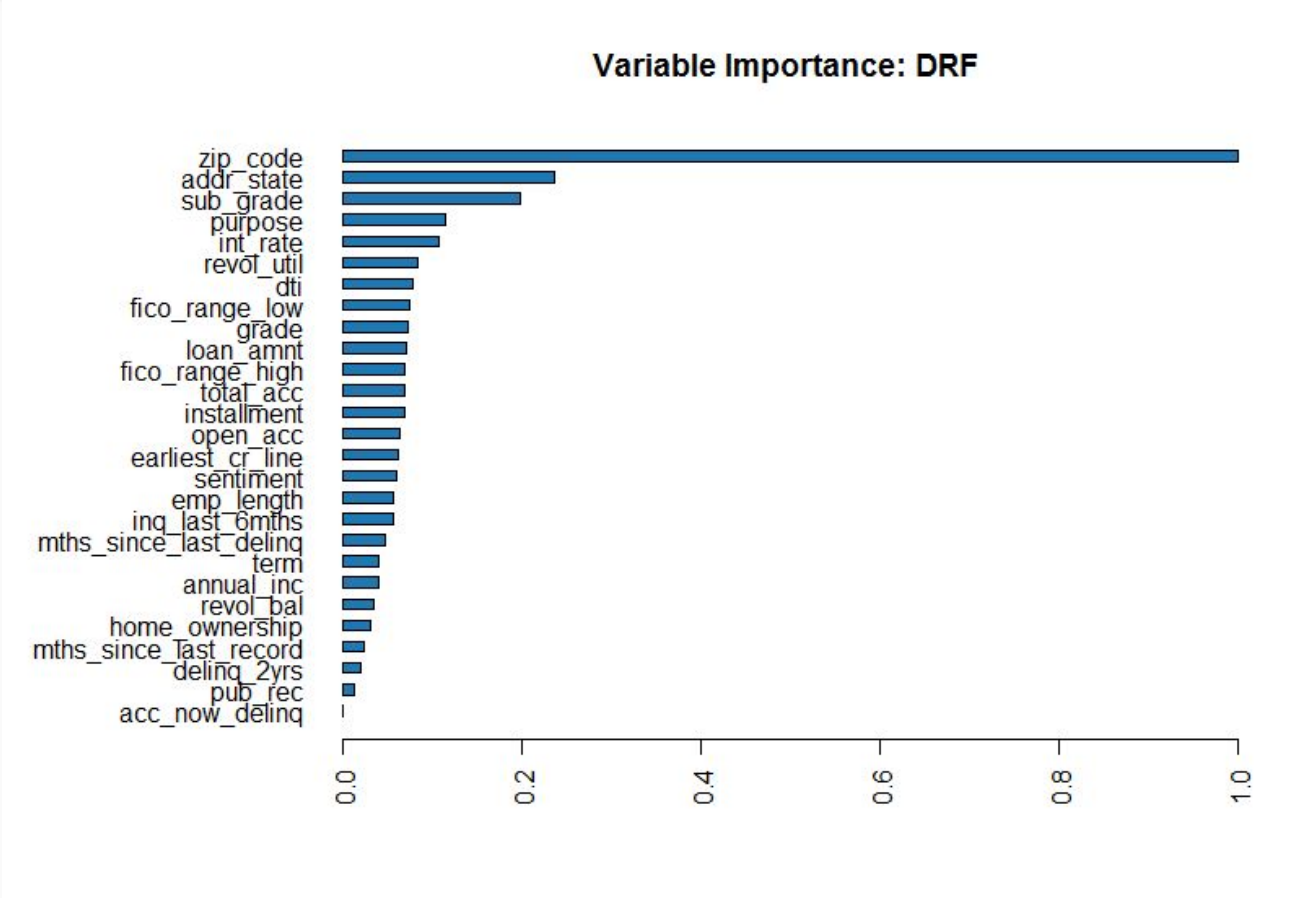


Second Goal

Loan Origination Prediction

- The goal was to predict loan default based on data ONLY available at loan origination time.
- The objective was to optimize from a lender's perspective

H2O Loan Origination Variable Importance Plot (Random Forest)

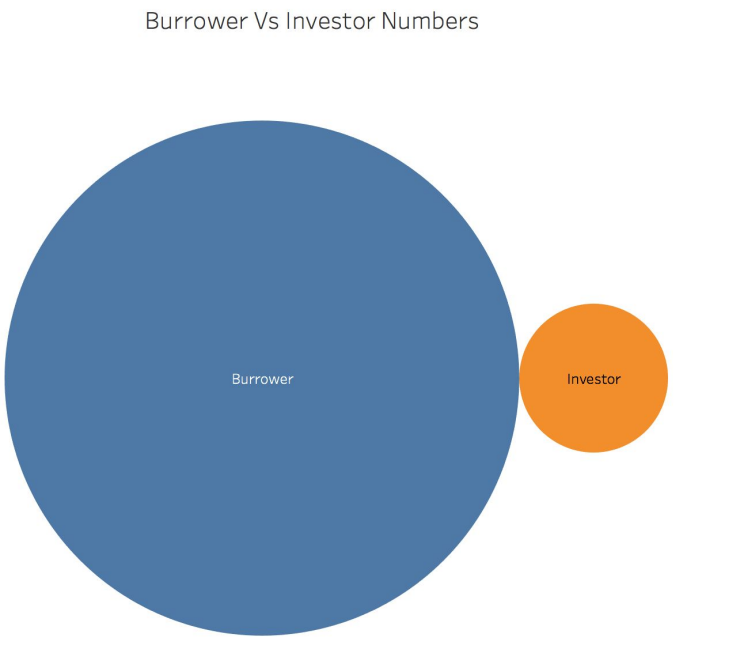


Best Model: H2O GLM Loan Origination ROC Curve: AUC = 0.7045
27 Features

H20 Loan Origination Model Performance Summary
Random Grid Search
10 Models Each, 5 Fold Cross Validation

Model	AUC
Ensemble (GLM)	0.7089
GLM	0.7045
Neural Network	0.7028
Random Forest	0.6878
GBM	0.6824
Naive Bayes	0.6657

Why optimize from a lender's perspective?

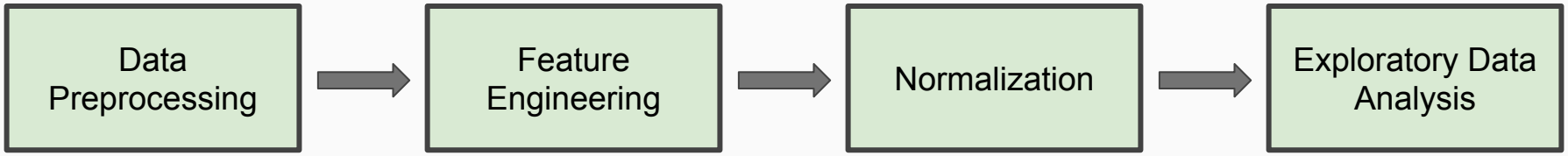


Per 2016 3rd Quarter Statistics released by Lending Club, its **borrower base is ~ 1.7 million** individuals and **investor base is ~142,000**.

Number of borrowers is **10 times** the number of investors.Every time a borrower's defaults, the investors in that loan loses their money.

This is a **250M $ company** and more such defaults, more investors are going to hesitate and the industry is going to take a hit.

The ‘Data Science’ Process



- **Binarize categorical** variables like purpose of loan, home ownership etc. to support sklearn models
- Recompute features as binaries like ‘# of defaults’ to ‘has defaulted?’

- **Sentiment Score -** Using AFINN() package determine overall sentiment score with loan request description.
- **TF-IDF score** of words in the loan request description
- **Months since last record**
- **Months since last delinquency**
- **Average FICO score -** derived from min and max score at loan initiation

Numeric non binary features were normalized to allow scaling of features to the same level for SciKit Learn.

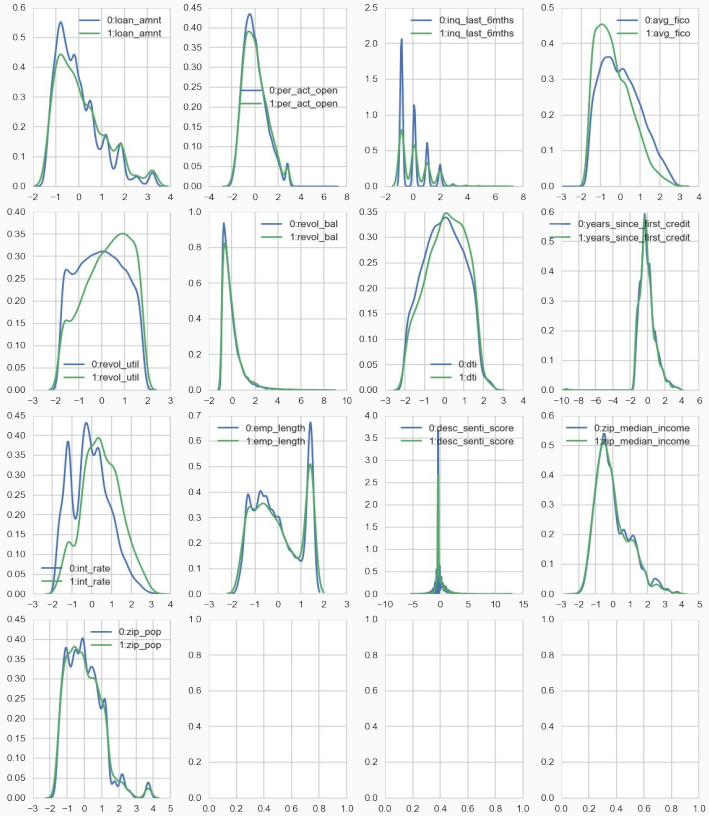
Data was visualized using KDE plot and histograms to identify if there are any undisputed predictors of good and bad loans.

KDE Plot was used to plot visualize continuous variables like loan amount, interest rate etc group by good and bad loans.

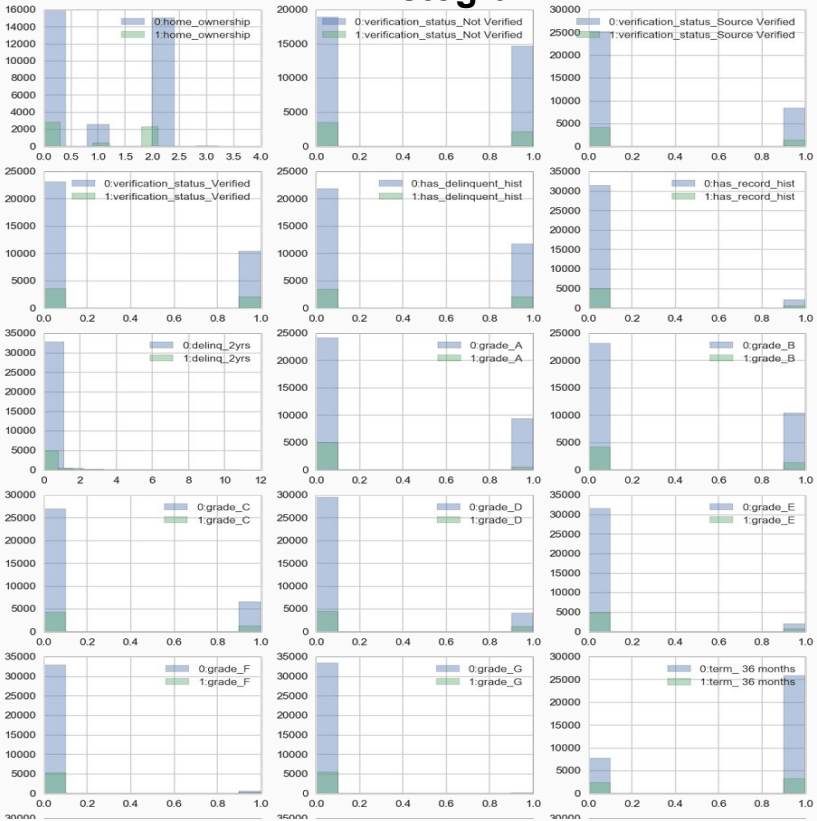
Histogram was used to analyze discrete variables home ownership, purpose of loan etc

Exploratory Analysis

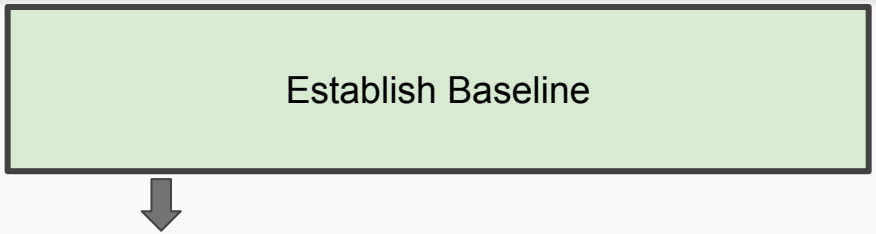
KDE Plot



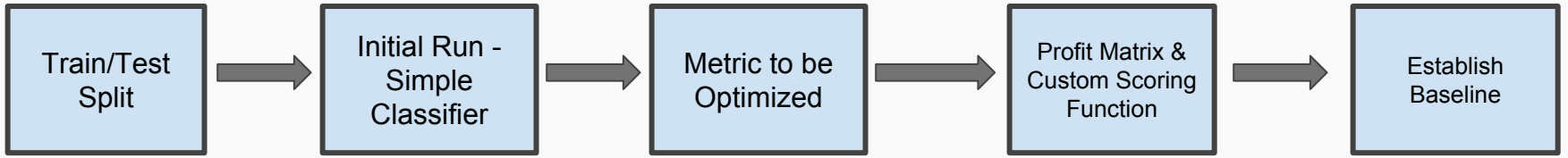
Histogram



The ‘Data Science’ Process (Contd)




```
[ 1333.,     -0.]
[-6108.,     -0.]
```



Train/Test Split and **5 fold cross validation** was carried out for all classifiers to follow.

Ran a simple classifier - Linear SVM, got an **accuracy of 85%** on the blind test dataset. But our dataset is skewed, 85% of our data are good loans, so an accuracy of 85% is same as predicting all loans as good.

Accuracy is not our primary metric, we want to make the model bias towards the investor, want to **maximize investor profit**.

Optimizing Metric - Profit Per Loan for the investor

Objective was to penalize bad loans classified as good
TN (0,0) 1333$ obtained by averaging accrued interest across all good loans

FP(1,0) - $ 6108 average of principal lost on bad loans and interest lost as well
FN, TP = 0 = no monetary gain
Multiply Profit matrix with CM

Predict all as good loans and also all loans as bad.
Determine the profit for each classification.

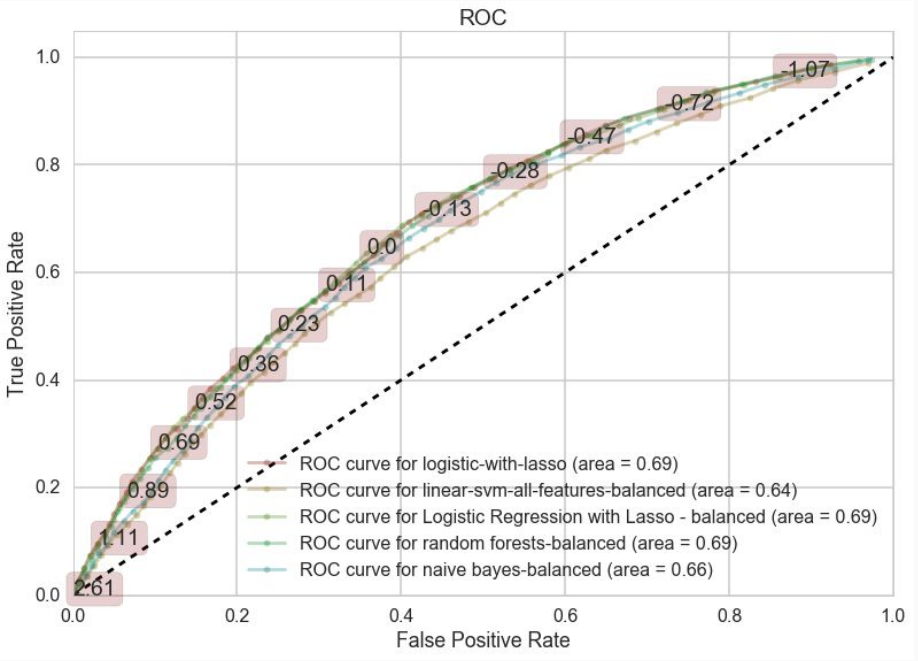
Baseline classifier was ‘predicting all loans as good with a profit of 260.8 per loan

	classifier	Profit Score	AUC	Sensitivity
0	Base Classifier:Predicting all loans = good	260.804342	None	0

The ‘Data Science’ Process (Contd)

Classification with 5 fold cross validation

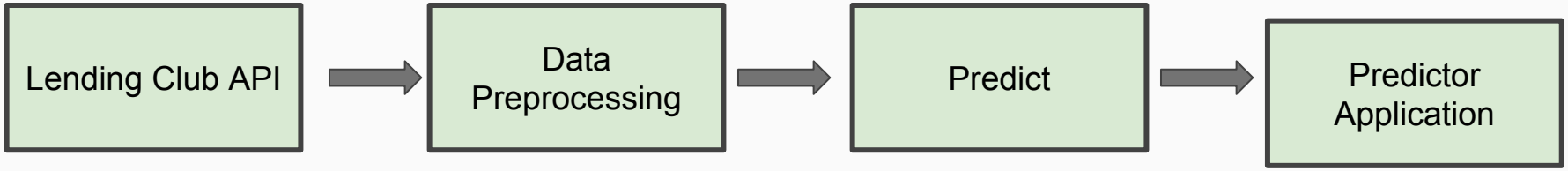
	classifier	Profit Score	AUC	Sensitivity
0	Base Classifier:Predicting all loans = good	260.804342	NaN	0.00
1	Logistic Regression with Lasso	268.093800	0.69	0.01
2	Logistic with Lasso modified Threshold = 0.16	407.410483	NaN	0.55
3	Linear SVM	260.804342	0.59	0.00
4	Linear SVM - Balanced	347.380544	0.64	0.63
5	Logistic Regression with Lasso - balanced	412.554406	0.69	0.64
6	Random Forests	400.557544	0.69	0.65
7	Naive Bayes	385.217878	0.66	0.50



Project Goal 3
Lending Club API

- The Lending Club web site makes an extensive API available to developers.
- We wanted to explore the API and make loan default predictions by obtaining data through the API and running it through our best model.
- Make a loan default prediction and present it to the Investor

Lending Club API(Contd)



Lending Club makes available a REST API to access its data.

It has a listing api, which allows pulling data around loans which are available to invest.

API - https://api.lendingclub.com/api/investor/v1/loans/listing

Data extracted using the API is preprocessed through all the steps our training data went through such as normalization and feature engineering.

The processed loan data is passed through the model to make a prediction.

The data flows into the predictor application, which takes as input the loan id the investor is trying to make a decision on and goes ahead and makes a prediction. It also makes recommendation.

Predictor Application Demo

Future Work

- Build a web based application to allow actual investor to interact with predictor application
- The marketplace loan data provides monthly FICO score data since loan origination but the loan downloaded loan data only provides the final FICO score. Use the lending club API to collect monthly FICO score data over time and work it into the model.