# Lending Club

## Can we predict default vs. paid in full accounts based on customer information?

# What I will cover:

---

- Business Overview
- Dataset Introduction
- Question I am trying to answer
- Model Selection Process
- Ideal Model
- Shortcomings of this process
- Next steps

# Business Overview:

---

- LendingClub is a peer-to-peer lending company that offers loan trading on a secondary market. According to the company, it has made about $15.98 billion in loans through its platform as of 2015.
- An individual investor can create unsecured personal loans between $1,000 to $40,000 in their platform. LendingClub also makes traditional direct to consumer loans.
- LendingClub offers alternative to traditional borrowing (some borrowers may have been closed off to the traditional banking system due to credit history or lack thereof).

# Question

**Question: Can we predict if a borrower will default on a loan based on their financial history and variables provided?**

**Target: This would help investors decide if they should lend to clients based on previous financial history with LendingClub**

# Dataset Information

---

- The data looks at 2007 - 2011 loans issued by LendingClub.
- The dataset contained over 42,000 rows* and 143 features.
- A lot of features were empty which were deleted
- Columns without at least 35,000 values without NaN were deleted
- Confidential information about the borrower were not provided by the company hence, were empty (customer ID etc)

# Dataset Information

- Categorical values were converted to booleans
- Only relevant features were selected by reviewing the dataset and picking features that seemed most relevant to the analysis
- NaN values were filled with 0 (please note only rows with at least 35,000 column values with non-Nans were kept)
- There is a clear class imbalance (as with most of financial datasets regarding defaults - majority of borrowers pay back their loans)

# Benchmark

---

**Fully Paid:** 86%

**Charged Off:** 14%

# Predictor Variables in Dataset

The predictor variables that were kept were:

- Loan Amount
- Funded Amount
- Term (Length of the Loan)
- Interest Rate
- Grade of the Loan
- Employment Length of Borrower
- Home Ownership Status (Rent / Mortgage etc)
- Annual Income
- Last Payment Amount
- Open Accounts

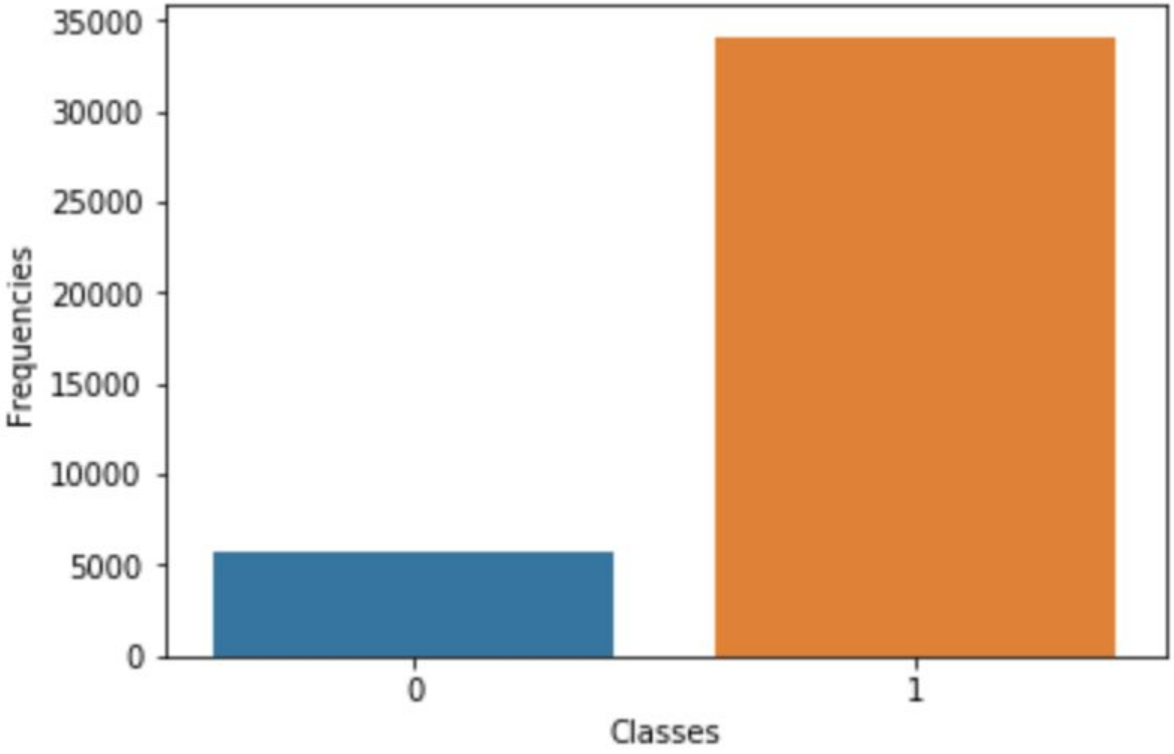
# Target Variables

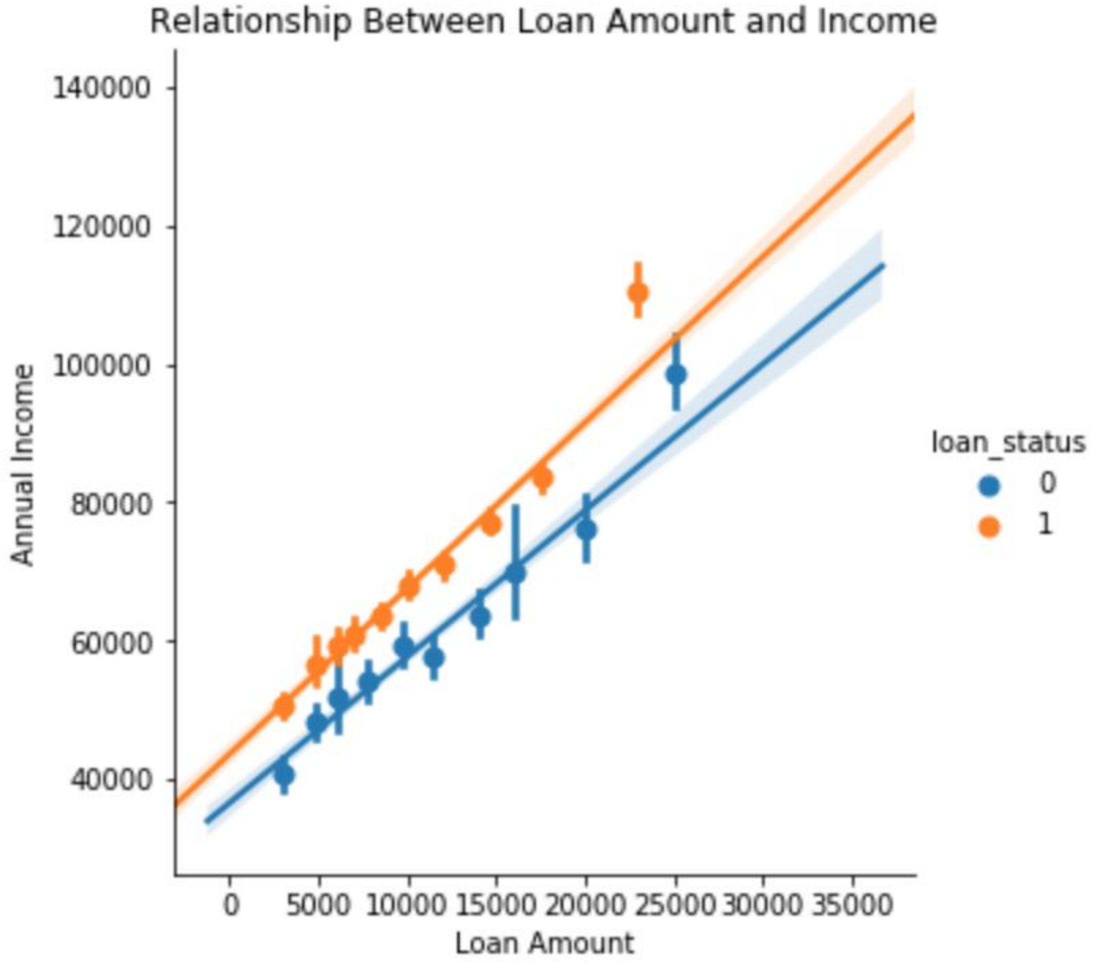
The target variable in my dataset was Loan Status. It has two different Values: Fully Paid v. Charged Off.

- **Fully Paid:** Loan has been paid fully
- **Charged Off:** Loan is not expected to be paid off

Since our goal is to predict if a borrower will default on a loan, I am looking at these two factors only. I changed Fully Paid to 1 and Charged Off to 0.

# Class Imbalance





Relationship Between Loan Amount and Income
Annual Income
140000
120000
100000
80000
60000
40000
0
5000
10000
15000
20000
25000
30000
35000
Loan Amount
loan_status
0
1

# Models Applied & Reasoning

- Naive Bayes
  - Simplistic model
  - Easy to start and understand at the beginning
- Random Forest with Grid Search
  - Reduces bias based on single important feature (in this case: loan payment status with feature importance of 30%)
- Logistic Regression with Grid Search
  - Easy to understand and the algorithm can be regularized to avoid overfitting
- SVM with Grid Search
  - Easy to see distinction between data (good visualization)
  - Once a hyperplane is found, data is easy to understand (outside of hyperplane is redundant)
- KNN Neighbors
  - Addresses the classification problem and low computational power

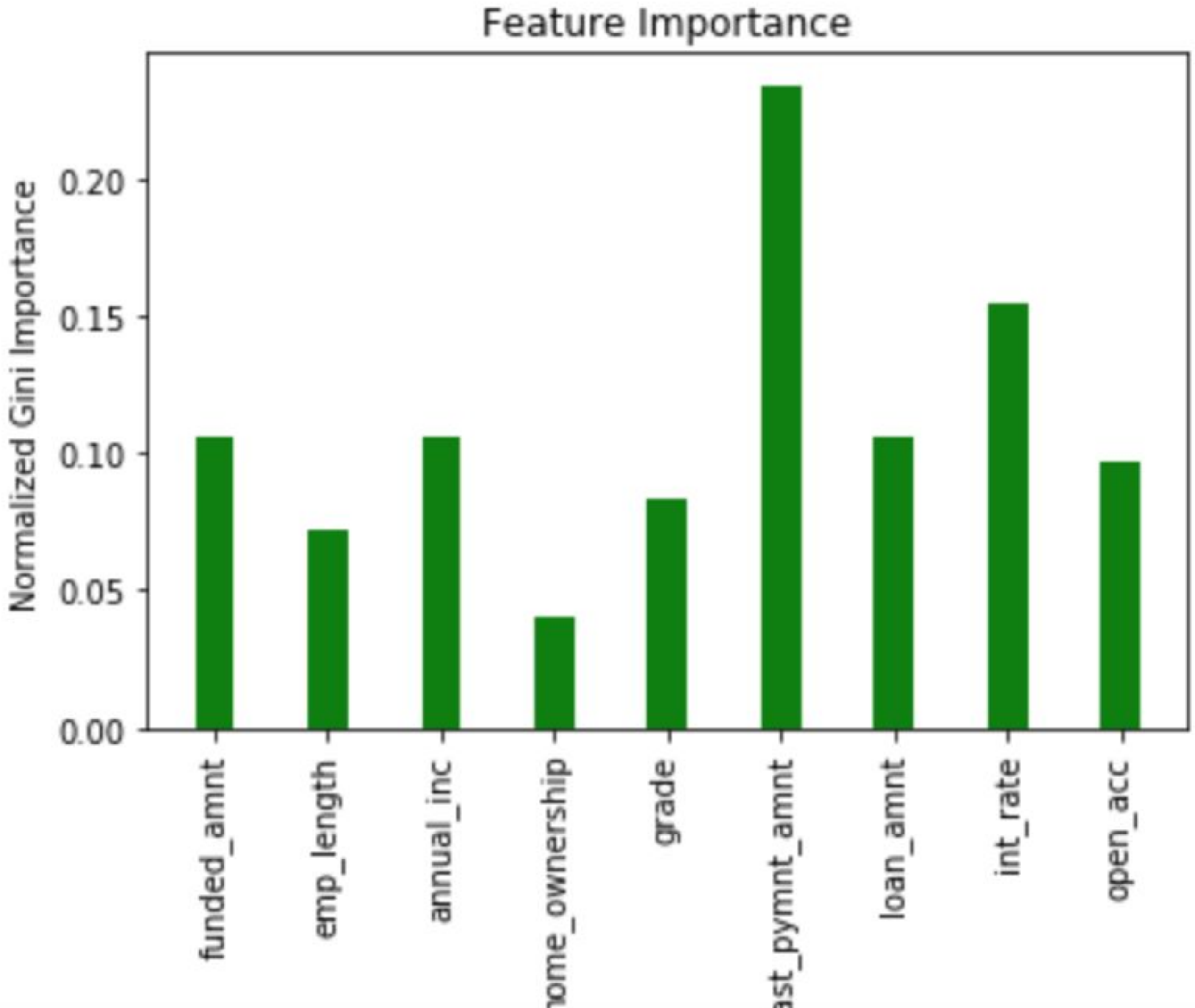
# Accuracy Scores (Before & After Grid Search)

| Models | Train | Test | GridSearch |
| --- | --- | --- | --- |
| Naive Bayes | 0.8574 | 0.8614 | - |
| Random Forest | 0.8607 | 0.8566 | 0.8752 |
| Logistic Regression | 0.8606 | 0.8571 | 0.8601 |
| KNN | 1 | 0.8069 | 0.8608 |
| SVM | Too large to run | Too large to run | Too large to run |

# Best Performer

The best model appears to be Random Forest in this case because, it has a higher Test Score than the Train Score. After applying the Grid Search, the accuracy score is 87% which is the highest accuracy score.

This also takes into account of one single feature importance (last_pymnt_amnt) which has an importance of over 30%. Random Forest reduces bias based on one single importance.



Feature Importance
Normalized Gini Importance
0.00
0.05
0.10
0.15
0.20
funded_amnt
emp_length
annual_inc
home_ownership
grade
ast_pymnt_amnt
loan_amnt
int_rate
open_acc

# Random Forest Classification Report

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.00 | 0.00 | 0.00 | 1108 |
| 1 | 0.86 | 1.00 | 0.93 | 6850 |
| micro avg | 0.86 | 0.86 | 0.86 | 7958 |
| macro avg | 0.43 | 0.50 | 0.46 | 7958 |
| weighted avg | 0.74 | 0.86 | 0.80 | 7958 |

# Class Imbalance with weights (Random Forest):

---

When I ran it with class weights as balanced:

| Train Score | Test Score |
| --- | --- |
| 0.9949 | 0.8589 |

# Classification Report:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.73 | 0.01 | 0.01 | 1108 |
| 1 | 0.86 | 1.00 | 0.93 | 6850 |
| micro avg | 0.86 | 0.86 | 0.86 | 7958 |
| macro avg | 0.79 | 0.50 | 0.47 | 7958 |
| weighted avg | 0.84 | 0.86 | 0.80 | 7958 |

# Class Imbalance with weights (Log Reg):

When I ran it with class weights as balanced:

| Train Score | Test Score |
| --- | --- |
| 0.6928 | 0.6883 |

# Classification Report

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.31 | 0.85 | 0.45 | 902 |
| 1 | 0.97 | 0.68 | 0.80 | 5464 |
| micro avg | 0.70 | 0.70 | 0.70 | 6366 |
| macro avg | 0.64 | 0.77 | 0.62 | 6366 |
| weighted avg | 0.87 | 0.70 | 0.75 | 6366 |

# Limitations

- Class Imbalance is an issue
  - Most of the data says the loan status is “Fully Paid” so, model is likely to predict “Fully Paid” instead of “Default”
  - Class weights is not perfect

- SVM model did not work as it require large computational power / time

# Application

- Can be used by investors looking to lend again to repeat individuals / businesses in the LendingClub platform
  - 
- Can be used by LendingClub to evaluate risk based on financial history of a business / individual
  - 
- Similar model can be used by companies in the alternative lending space

# Next steps:

- Collecting more samples would be beneficial as it would allow us to address the class imbalance in a more applicable way

- Other models might be more efficient that were not run due to computational problems