

# Loan Default Prediction

---

Samuel P. Sears

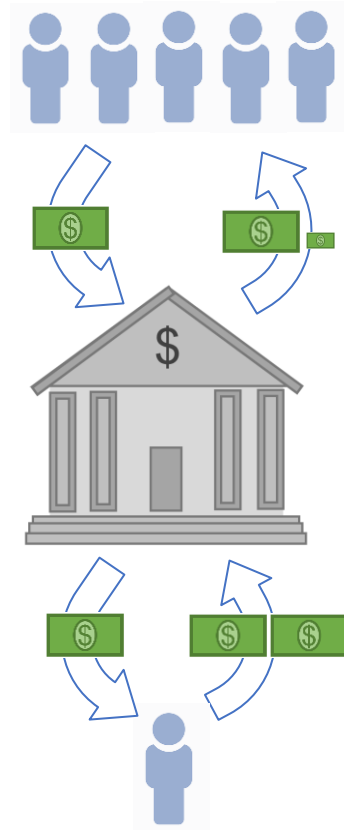
# Background

## Traditional Lending

Individuals deposit funds in banks and receive a **small** amount of interest.

Banks lend money to borrowers and receive a **larger** amount of interest.

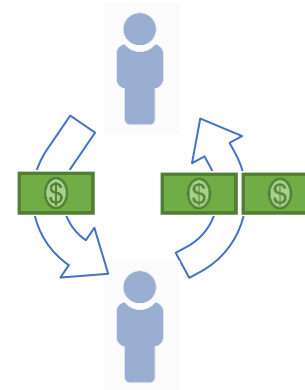
- Wells Fargo
- Bank of America
- US Bank



## Peer-to-Peer Lending

Individuals lend money to borrowers for a **larger** amount of interest.

- LendingClub
- Peerform
- Upstart
- Prosper



# Business Problem

## **Goal**

As an investor, I want to maximize my return on investment

## **Problems**

- If a loan is not paid back, I could experience a loss
- Safer loans result in lower returns

## **Need**

Identify riskier, high return loans that will be paid in full

# The Data

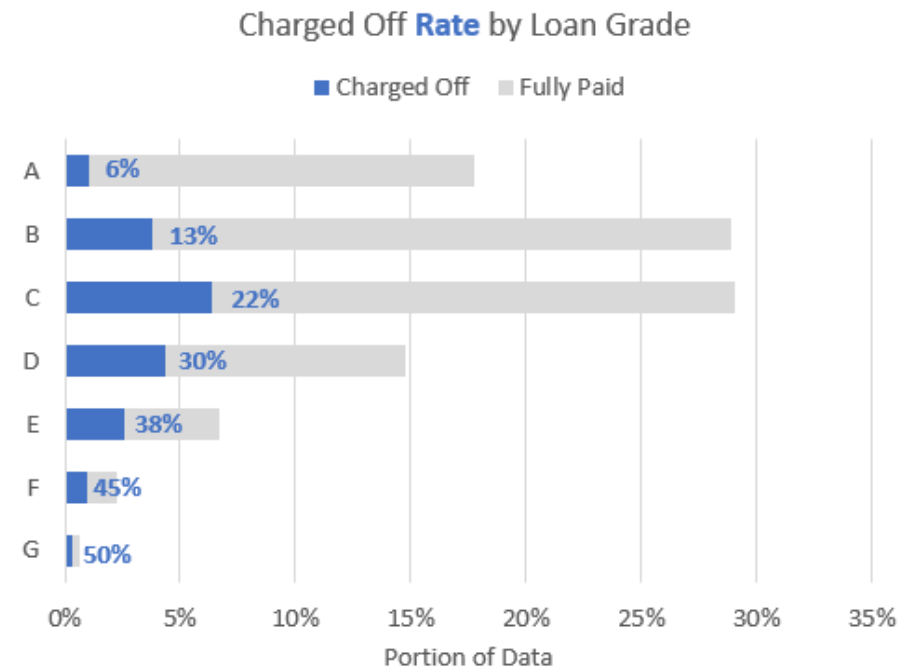
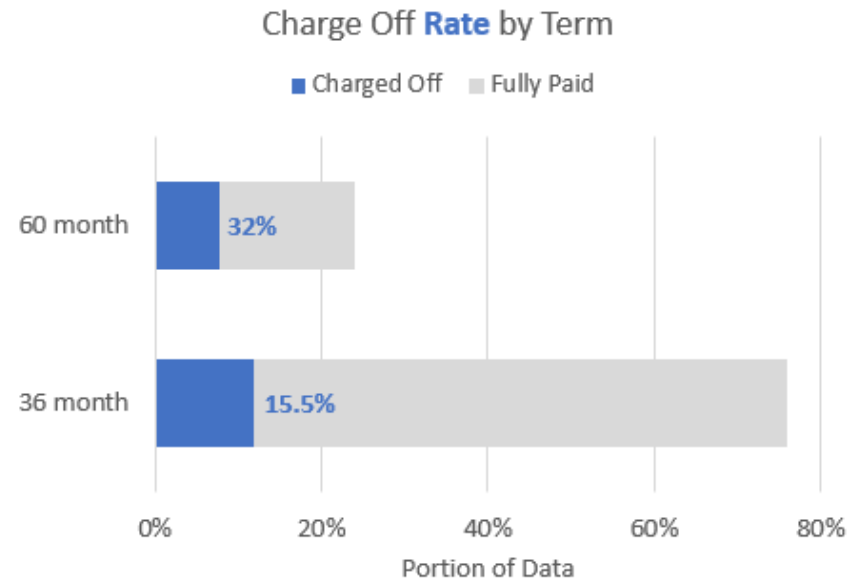
Kaggle – [All Lending Club loan data](#)

Variable	Loan Amount	Term	Installment	Grade	Annual Income	Purpose	Debt to Income Ratio	Open Accounts	FICO	...	Loan Status
Type	Numeric	Categorical	Numeric	Categorical	Numeric	Categorical	Numeric	Numeric	Numeric	...	Categorical
Values	500 ... 40,000	36 months 60 months	4.96 ... 1,719.83	A ... G	33 ... 10,999,200	14 different categories	-1 ... 999	1 ... 90	1 ... 163	...	Paid in full Charged off
Description	Dollar amount of loan	Term length of loan	Monthly repayment amount	Loan rating where A is least risky and G is most risky	Self-reported annual income of borrower	Category provided by borrower for loan request	Debt-to-income ratio of the borrower	The number of open credit lines in the borrower's credit file	Average of high and low FICO score ranges for the borrower	...	Status of loan

*Not Shown – emp\_length, home\_ownership, verification\_status, delinq\_2yrs, earliest\_cr\_line, inq\_last\_6mnths, pub\_rec, revol\_bal, revol\_util, total\_acct, application\_type*

1.15 million accepted loans from 2007 – 2018  
81% of loans paid in full | 19% charged off

# Insights from Data Exploration – No Surprises



As expected, loans with typical indicators of risk (low grade, low FICO, etc.) had higher charge off rates

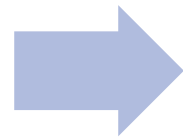
# Predictive Modeling Methods

## Preprocessing

Under sampled Fully Paid Loans

### 5 Preprocessing Combinations Tested

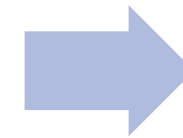
- Normalization
- Outlier Removal
- Transformations
- Changing Data Types
- Ignore Low Variance



## Model Selection

### 4 Models Tested

- CatBoost Classifier
- Gradient Boosting Machine
- Extreme Gradient Boosting Classifier
- Light Gradient Boosting Machine



## Final Model

### CatBoost Classifier

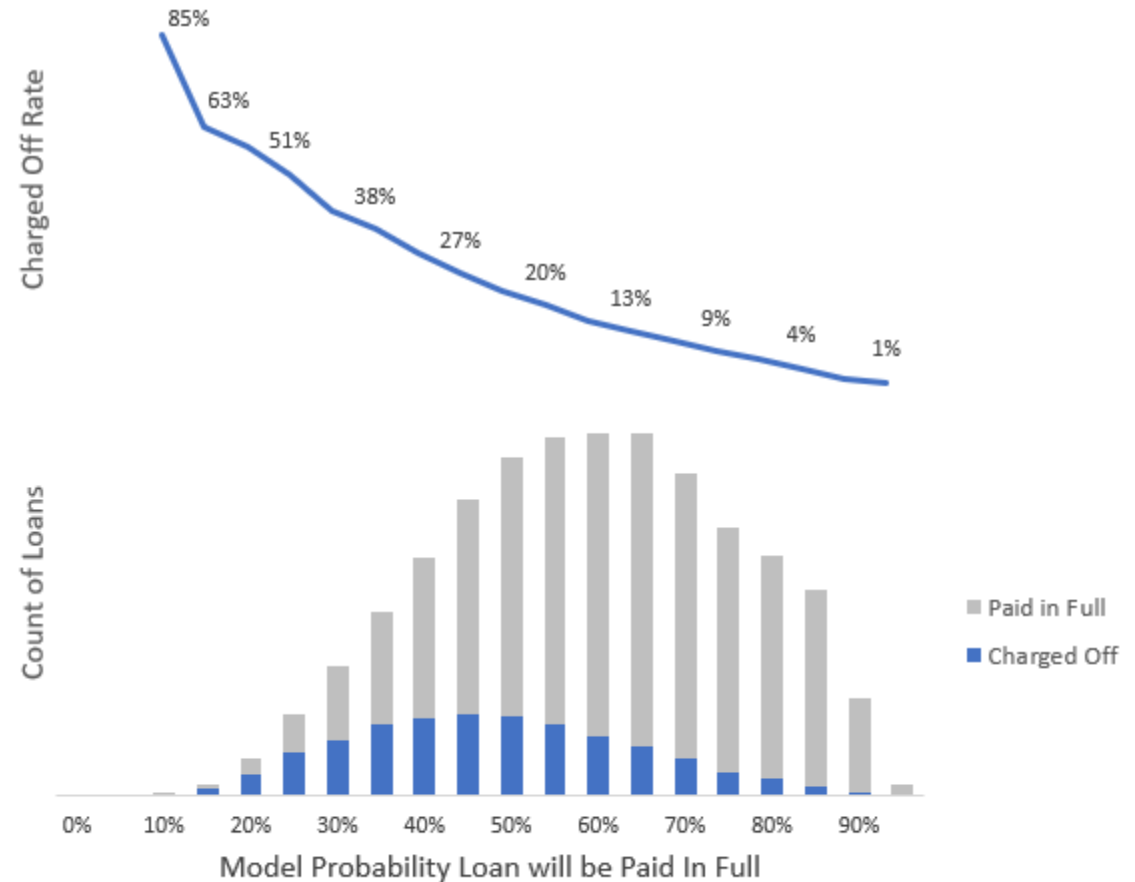
- Normalization
- Outlier Removal
- Loan Grade as Numeric

### AUC

- Train: 0.7151
- Test: 0.655

# Model Results

Probability	Interval Total	Paid in Full	Charged Off	Paid In Full Rate
0.00-0.05	0	0	0	
0.05-0.10	0	0	0	
0.10-0.15	20	3	17	15%
0.15-0.20	263	98	165	37%
0.20-0.25	1030	434	596	42%
0.25-0.30	2296	1120	1176	49%
0.30-0.35	3660	2098	1562	57%
0.35-0.40	5182	3213	1969	62%
0.40-0.45	6689	4525	2164	68%
0.45-0.50	8374	6077	2297	73%
0.50-0.55	9554	7352	2202	77%
0.55-0.60	10118	8119	1999	80%
0.60-0.65	10263	8622	1641	84%
0.65-0.70	10268	8905	1363	87%
0.70-0.75	9113	8092	1021	89%
0.75-0.80	7574	6929	645	91%
0.80-0.85	6767	6309	458	93%
0.85-0.90	5820	5568	252	96%
0.90-0.95	2735	2679	56	98%
0.95-1.00	274	271	3	99%



# Strategies

## Selecting Loans by Grade

Grade	Investment	Return	Return Percent
A	\$245,454,075.00	\$12,622,301.91	5.10%
B+	\$622,146,300.00	\$31,666,138.74	5.10%
C+	\$1,035,893,950.00	\$41,948,090.44	4.00%
D+	\$1,258,940,825.00	\$42,853,184.45	3.40%
E+	\$1,376,237,325.00	\$40,603,722.92	3.00%
F+	\$1,416,926,275.00	\$38,697,269.68	2.70%
G+	\$1,429,701,975.00	\$37,753,355.10	2.60%

## Selecting Loans by FICO Score

FICO	Investment	Return	Return Percent
> 840	\$639,125.00	(\$19,325.81)	-3.00%
> 810	\$9,590,725.00	\$330,300.34	3.40%
> 780	\$42,523,125.00	\$1,363,990.21	3.20%
> 750	\$116,039,075.00	\$3,527,805.02	3.00%
> 720	\$313,465,825.00	\$10,546,842.22	3.40%
> 710	\$439,177,375.00	\$15,958,308.19	3.60%
> 700	\$594,961,275.00	\$21,532,477.57	3.60%

## Selecting Loans by Model Probability

Probability	Investment	Return	Return Percent
> 0.70	\$300,674,000.00	\$18,302,372.69	6.10%
> 0.75	\$220,936,150.00	\$13,460,299.83	6.10%
> 0.80	\$146,511,025.00	\$9,099,967.75	6.20%
> 0.85	\$75,170,350.00	\$4,729,705.43	6.30%
> 0.90	\$21,125,500.00	\$1,231,376.73	5.80%
> 0.95	\$582,000.00	\$31,937.90	5.50%
> 0.96	\$93,475.00	\$5,614.41	6.00%
> 0.97	\$12,050.00	\$184.33	1.50%



# Proposed Solution

## Return Potential:

$$\frac{((\text{Monthly Installments} \times \text{Term Length}) - \text{Loan Amount})}{\text{Loan Amount}}$$

## Grid Search:

All Model Probabilities and Return Potentials tested on 5 different test sets

## Best Results:

Return Potential > 45%

Model Probability > 0.55

*Investing in all loans with return potential greater than 45% and model probability greater than 55%*

Test Set	Investment	Return	Return Percent
A	\$2,094,700.00	\$341,931.27	16%
B	\$1,651,150.00	\$287,614.09	17%
C	\$2,048,050.00	\$324,628.08	16%
D	\$1,763,950.00	\$199,015.67	11%
E	\$1,808,325.00	\$319,421.78	18%

# Further Analysis

## **Validation**

- Dataset specific to platform used
- Return results varied across test data set

## **Different Approach**

Predict the return amount as opposed to whether or not loan will be paid in full



# Project Repository

<https://github.com/ssears219/Loan-Default-Prediction>