

# Loan Outcome Prediction

Anthony (Ernie) S. Leon-Guerrero


# The problem

Loan charge-offs for private lenders



The Federal Reserve's charge-off rate for consumer loans from all banks in Q3 2020 is 1.82%. This rate is much higher for private lenders due to a few factors:

- Faster approval times
- Informal application processes
- Work with more borrowers with bad credit
- Greater use of discretion



Could we create a prediction model to minimize the charge-off rate for private lenders?

# Who might benefit?

Individual private lenders

Private lending companies:

 **LendingClub**

**freedomplus** 

**PROSPER** 

 **LIGHTSTREAM**<sup>SM</sup>

  
**LENDINGPOINT**<sup>TM</sup>

**AVANT**

**Payoff**<sup>TM</sup>

  
**best egg**<sup>®</sup>

# Data

## Source

Approved loans from  
Lending Club

- <https://www.kaggle.com/wordsforthewise/lending-club>

## Date range

Jan 01 2007 to  
Dec 31 2018

## Size

- 1.3M observations
- 141 features

# Data cleaning challenges

## Challenge 1

### **Dual applicants**

The dataset contains ~26k observations for dual applicants with an entirely different set of features; we'll define these as outside of scope.

## Challenge 2

### **Highly correlated features**

Some features (such as grade and sub-grade) were nearly identical and were dropped to reduce complexity.

## Challenge 3

### **Many null values**

- Features with >90% were dropped
- Dropped 5% of observations which were all missing the same data
- Filled with max, 0, or median as needed



EDA

# Grade distinctions

- Borrower data decides grade
- Grade decides interest rate
- Interest rate correlates with charge-off rate

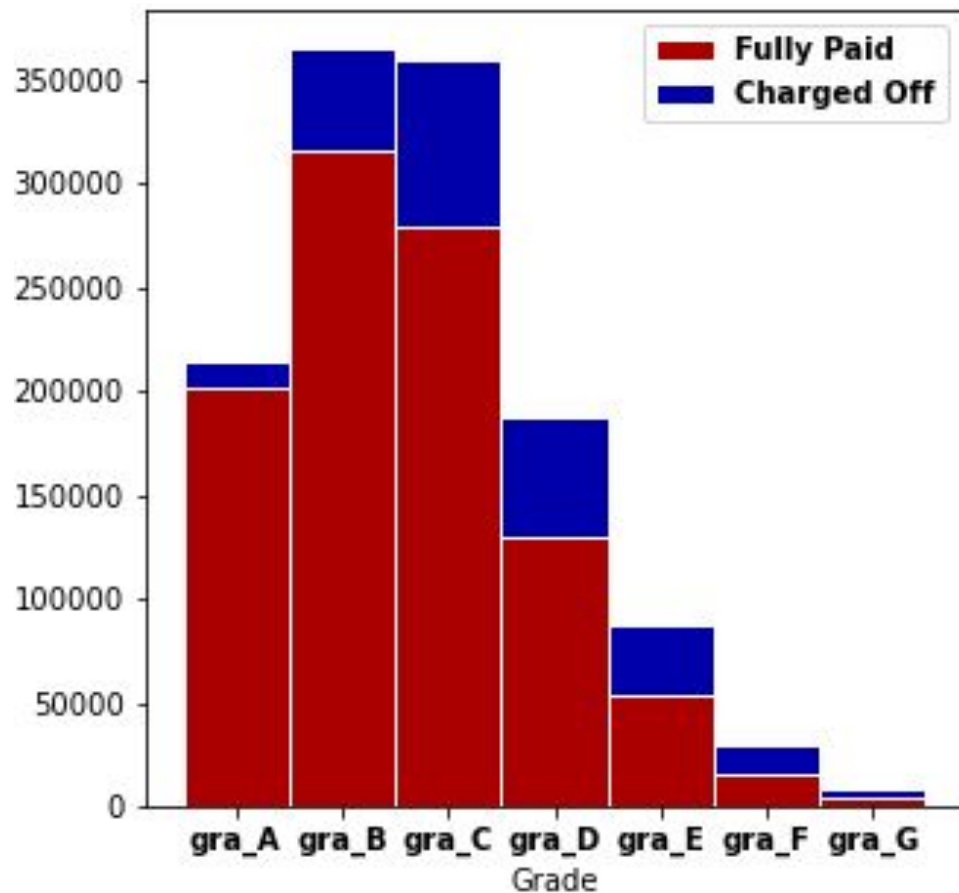




# Charge-off rate by grade

- 20.12% overall charge-off rate
- Grade A: 5.94% charge-off rate
- Grade G: 50.64% charge-off rate
- Each grade has a higher charge-off rate than the previous; ie Lending Club is good at rating relative risk

Grade	Fully Paid	Percent	Charged Off	Percent
gra_A	202209	94.06	12773	5.94
gra_B	316125	86.62	48814	13.38
gra_C	278460	77.46	81024	22.54
gra_D	129898	69.32	57491	30.68
gra_E	53199	61.02	33987	38.98
gra_F	16074	54.21	13580	45.79
gra_G	4120	49.36	4227	50.64
All grades	1000085	79.88	251896	20.12



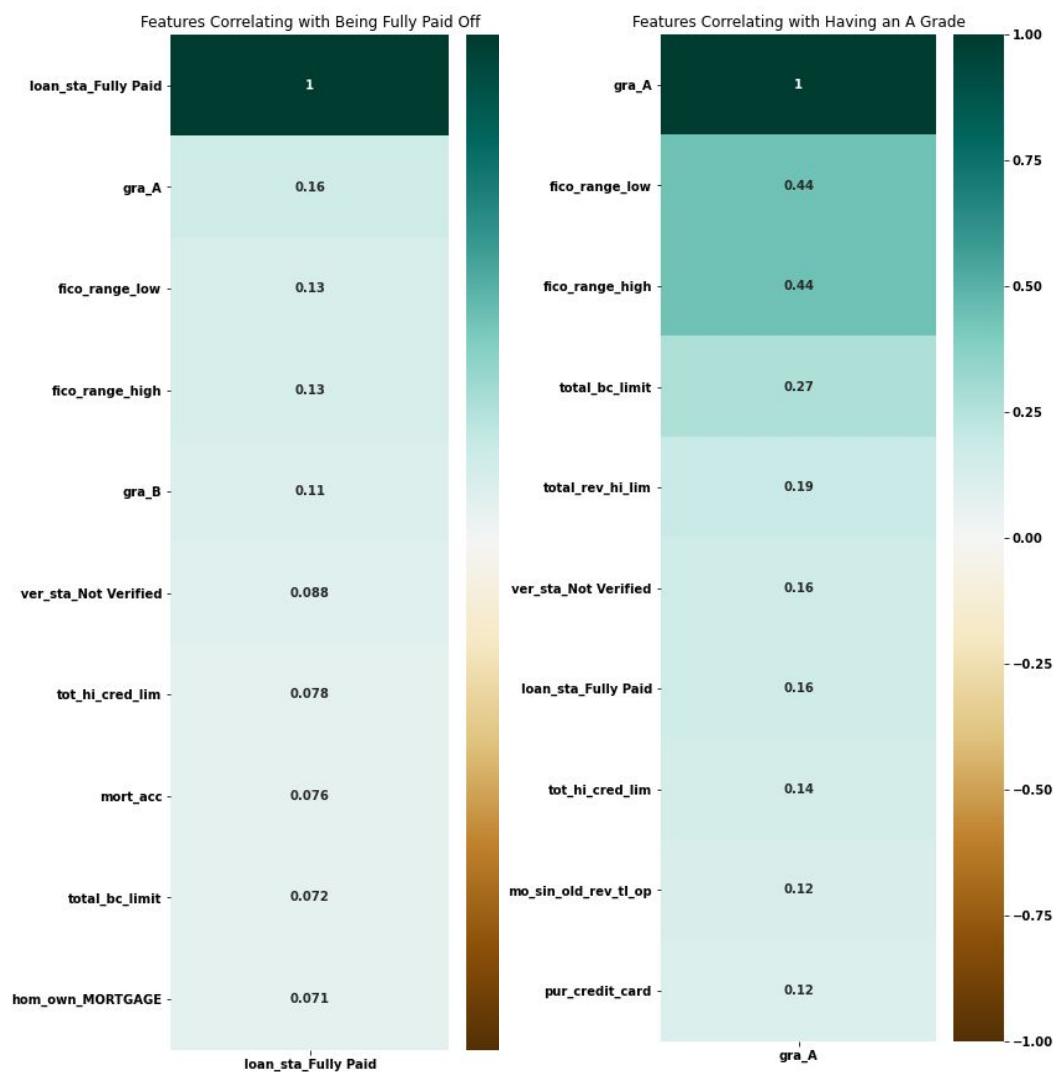
# Correlations between features

- There are some strong, but mostly expected correlations

First Feature	Second Feature	Correlation
upper boundary of borrower's FICO score	lower boundary of borrower's FICO score	1.000
number of open credit lines	number of satisfactory accounts	.999
amount of loan	amount of monthly payments	.954
months since oldest revolving account opened	years between opening of first line of credit and loan issuance	.918
number of revolving accounts	number of bankcard accounts	.838

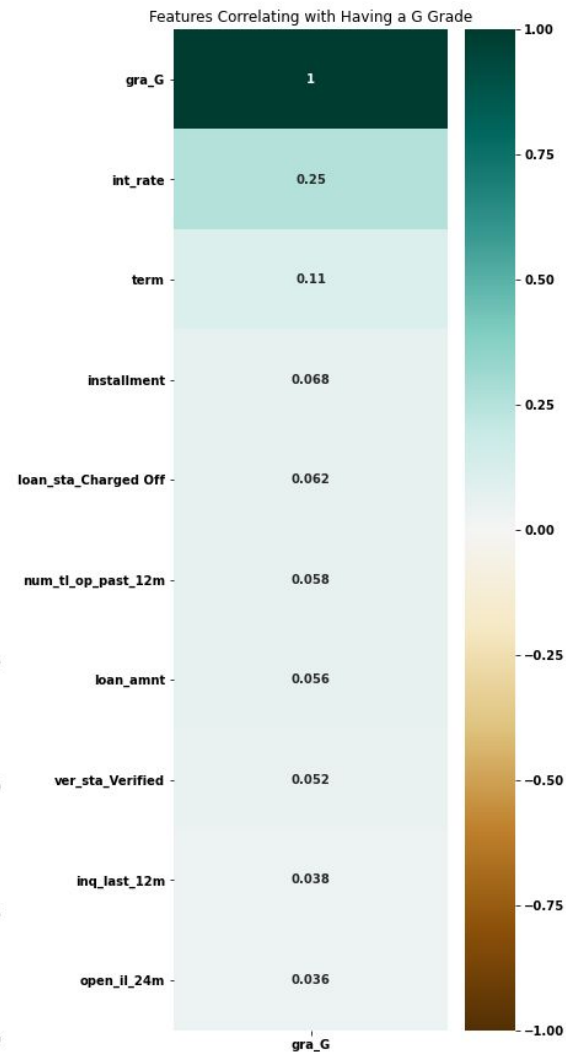
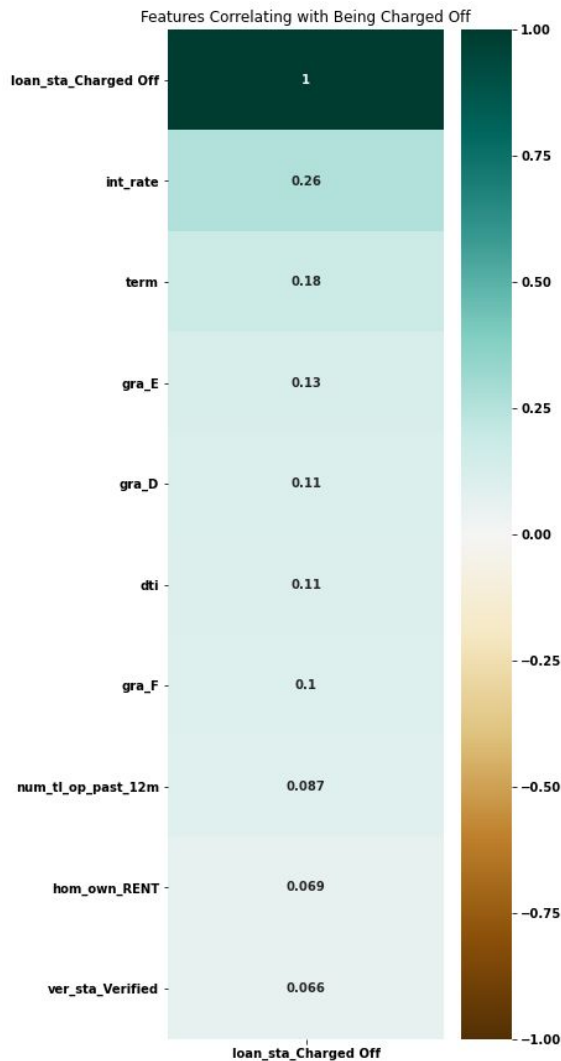
# Correlations for fully paid loans

- Having a high grade and being fully paid off have similar correlations
- No strong predictors for being fully paid off
- FICO score is primary decider for grade



# Correlations for charged-off loans

- Having a low grade and being charged-off have similar correlations
- Interest rate and length of term have highest correlations for both



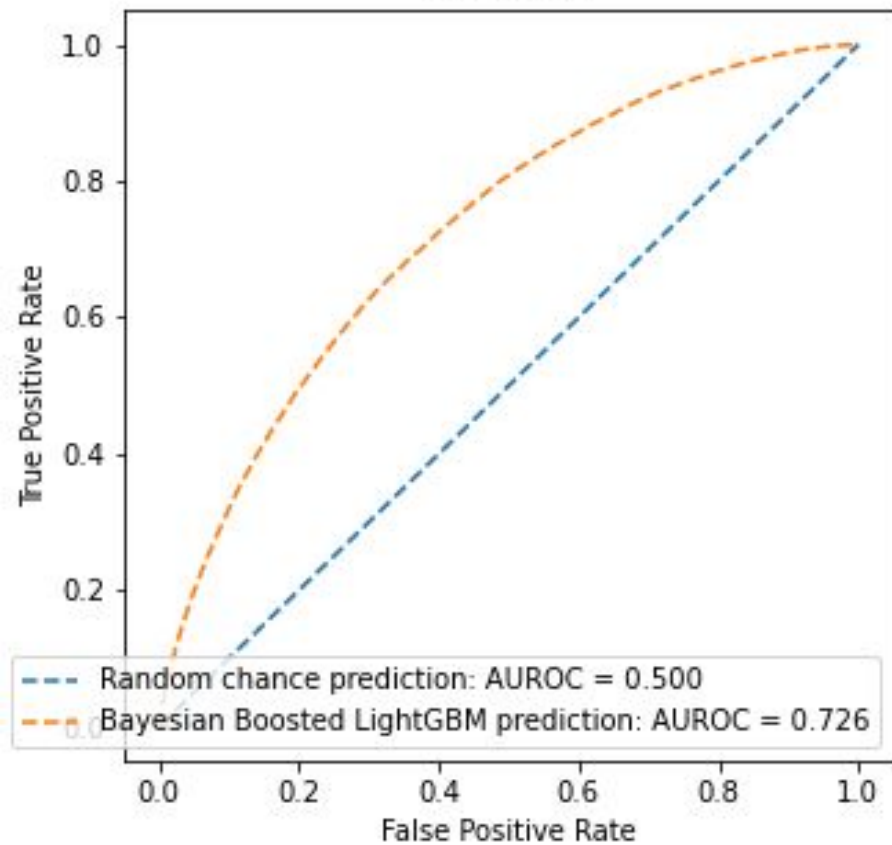


# Machine Learning & Modeling

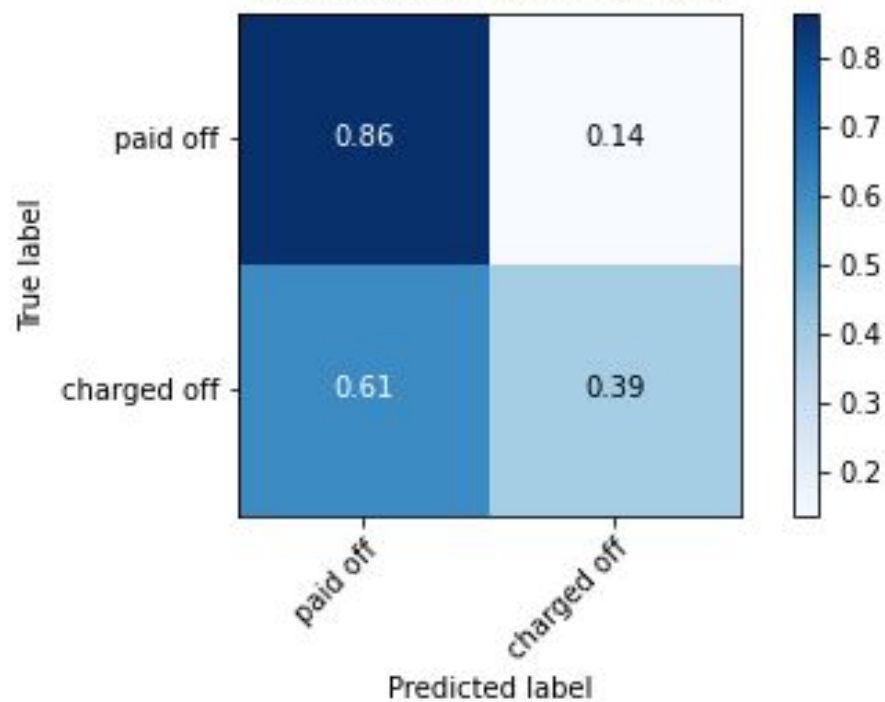
# ML methods

- Bayesian optimization with LightGBM (Gradient Boosting Machine) algorithms tried:
  - Traditional gradient boosting decision tree
  - Random forest
  - Dropouts meet multiple additive regression trees
  - Gradient-based one-side sampling (GOSS)
- GOSS performed best
  - Iterations = 6
  - CV = 5

ROC Plot



Normalized confusion matrix



# Measuring improvement

- Assuming 50% unpaid principal for charged-off loans
- Threshold = .65 maximizes profit via interest
  - 15.00% charge-off rate versus 20.12% of original dataset
  - Inclusion of un-approved loans could increase total profit
- Further lowering threshold increases TPR (predicting a charge-off), increases profit per loan
  - TNR (predicting a fully paid off loan) decreases, decreasing total profit

Model Threshold	Profit per Loan	Improvement Over Original	Total Profit	Improvement Over Original	Loans Misclassified as Charge-off	Total Un-earned Profits	Loans Correctly Classified as Charge-Offs	Total Losses Prevented
Original Dataset	\$1,497	0.00%	\$1,873,755,479	0.00%	0	\$0	0	\$0
.7	\$1,926	28.72%	\$2,097,808,571	11.96%	90,008	\$345,141,121	73,050	\$568,543,632
.65	\$2,074	38.55%	\$2,102,113,032	12.19%	140,012	\$536,886,188	98,239	\$764,593,160
.6	\$2,254	50.64%	\$2,087,673,433	11.42%	200,017	\$766,980,268	125,948	\$980,247,641



# Future improvements

## Dual applicants

There was much less data for dual applicants, and it used different features; modeling could improve outcome classification here as well.

## Denied applicants

The only data available was for approved loans. Looking at denied applications could reveal a portion of low-risk loans.

## Number of iterations

Performing more than 6 iterations of Bayesian optimization could reveal better results.

# Thank you!

Special thanks to Springboard  
mentor Tony Paek

Anthony (Ernie) S. Leon-Guerrero

[AnthSLG@GMail.com](mailto:AnthSLG@GMail.com)

[www.linkedin.com/in/anthslg/](https://www.linkedin.com/in/anthslg/)

[github.com/ErnieLG/Projects/](https://github.com/ErnieLG/Projects/)

[github.com/ErnieLG/Projects/tree/  
main/Loan\\_Outcome\\_Classifier](https://github.com/ErnieLG/Projects/tree/main/Loan_Outcome_Classifier)

---