

# Analysis of 2018 LendingClub Loan Data

---

Statistical Data Analysis & Machine Learning Models

Ngoc Phan | M.S. Business Analytics | [nphan20181@gmail.com](mailto:nphan20181@gmail.com) | [Github](#)

# Agenda

---



Project Background



Exploratory Data Analysis



Mission Statement



Machine Learning Models



Dataset



Recommendations



Data Wrangling

# Project Background

## LendingClub



Source: <https://www.pymnts.com/news/alternative-financial-services/2016/lending-club-timeline/>

# Mission Statement

---

- Assist LendingClub investors in the loan selection process
  - Exploring...
    - Fully paid vs. default loans
    - Return on investment (ROI) and loss of investment.
  - Estimating number of defaults for a selected sample
  - Developing machine learning models
  - Providing recommendation on loan selection

---

## 2018 LendingClub Loan Data

<b>Rows</b>	495,242
<b>Columns</b>	144
<b>Column Title</b>	Loan Amount, Interest Rate, Term, Annual Income, etc.

Dataset

---

# Data Wrangling

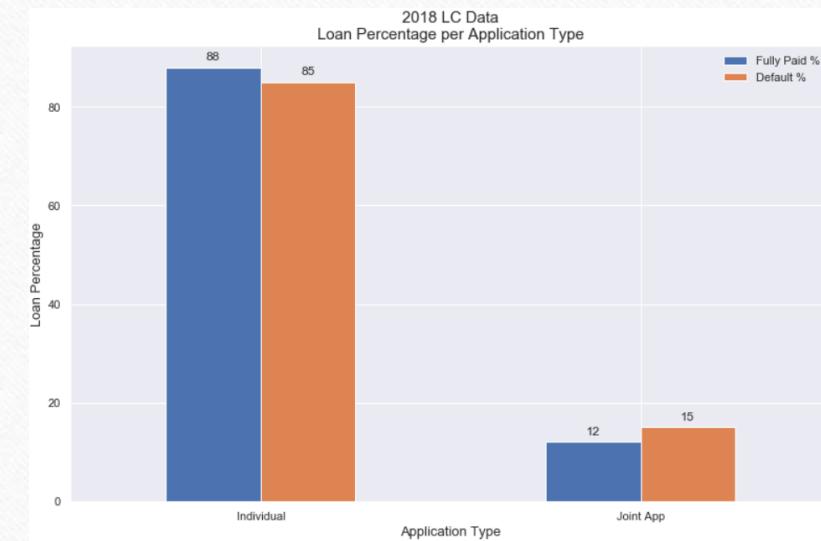
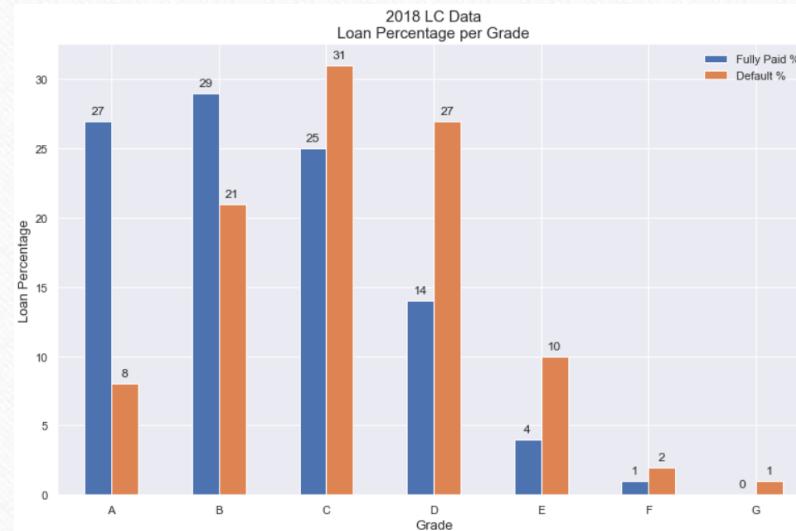
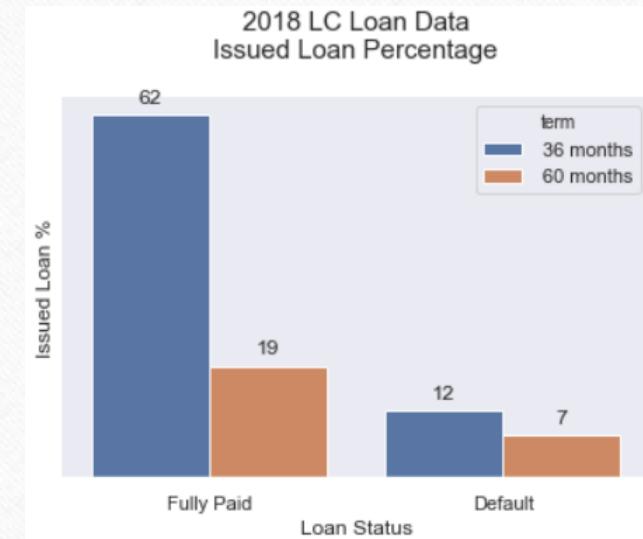
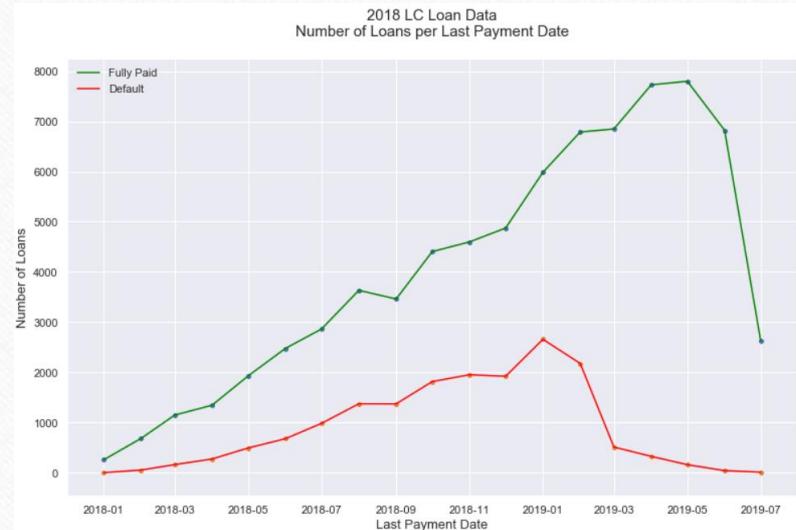
---

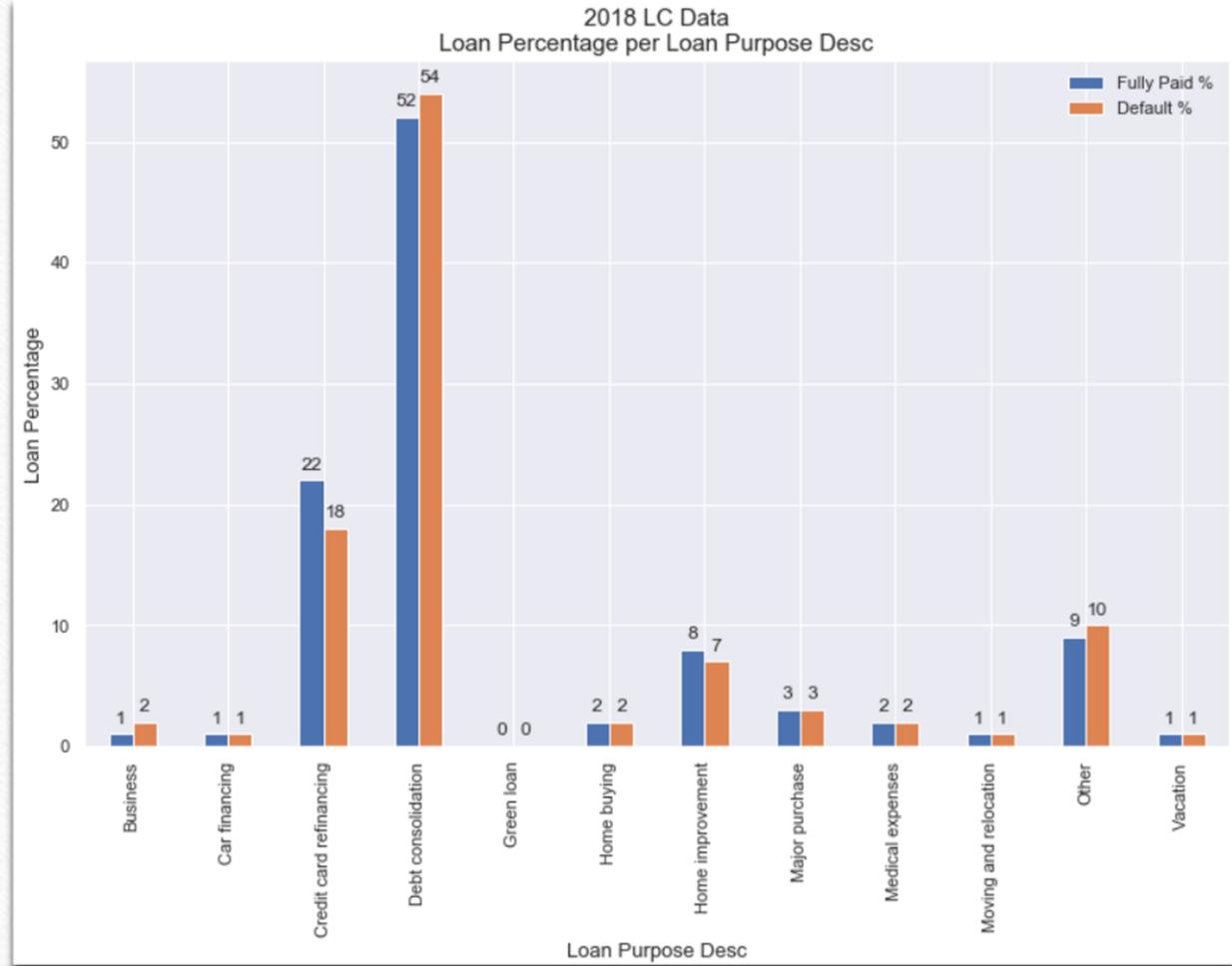
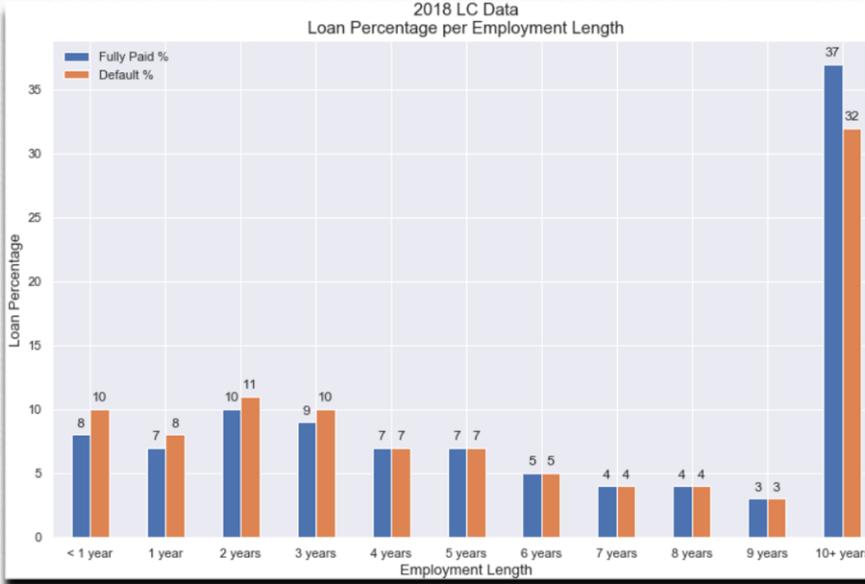
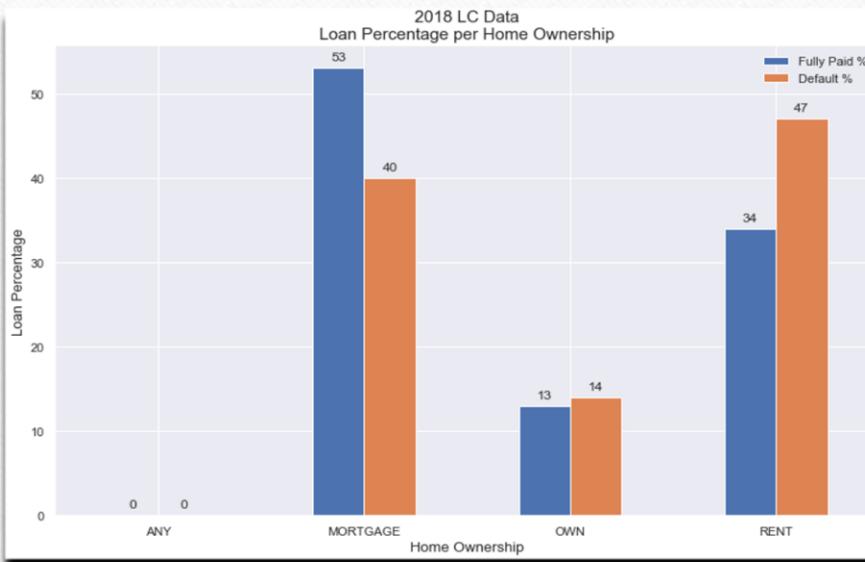
- Remove columns
  - Empty Columns
  - Columns with unimportant info
- Keep rows without status of fully paid / charged off / default
- Outliers
  - 64 columns and 36,355 rows
- Missing Values
  - Columns with  $> 25\%$  missing values
  - Columns with  $< 25\%$  missing values
- New Columns
  - Loan Status Flag: default, fully-paid
  - Return on Investment (ROI)
  - Months-in-Loan

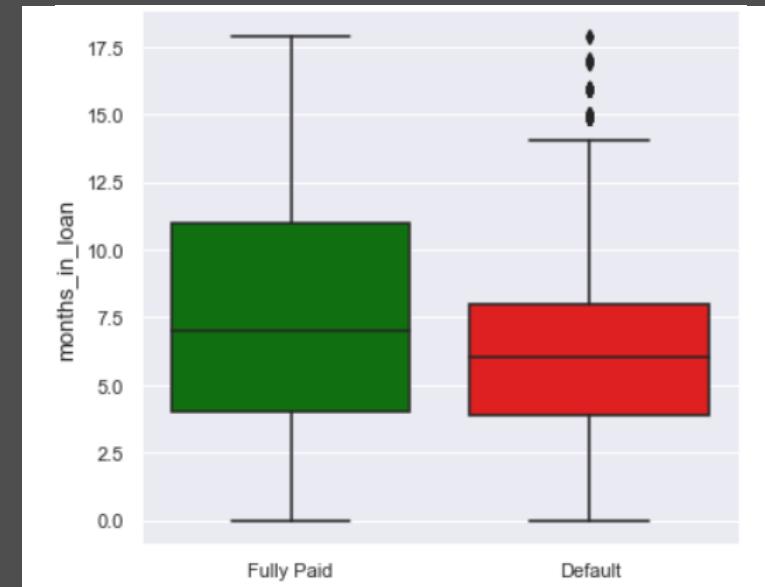
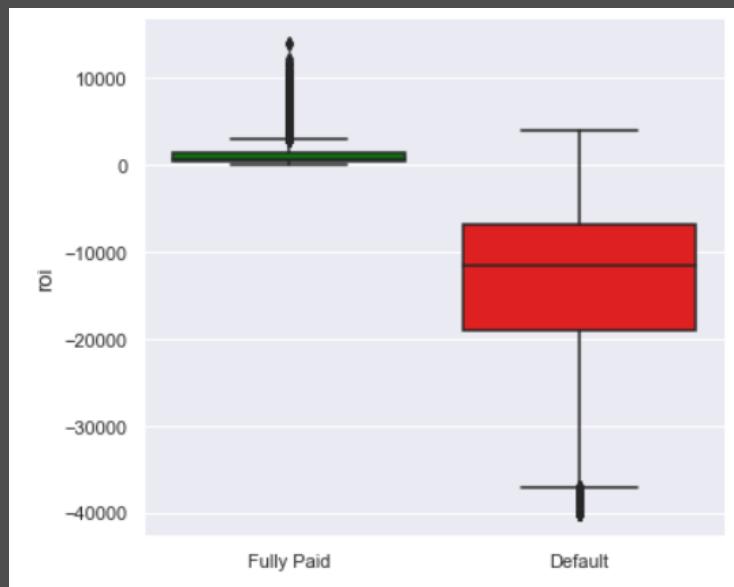
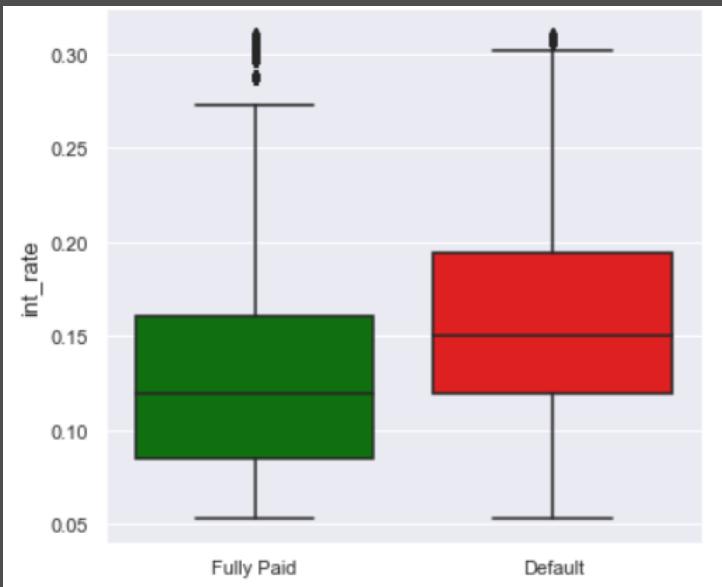
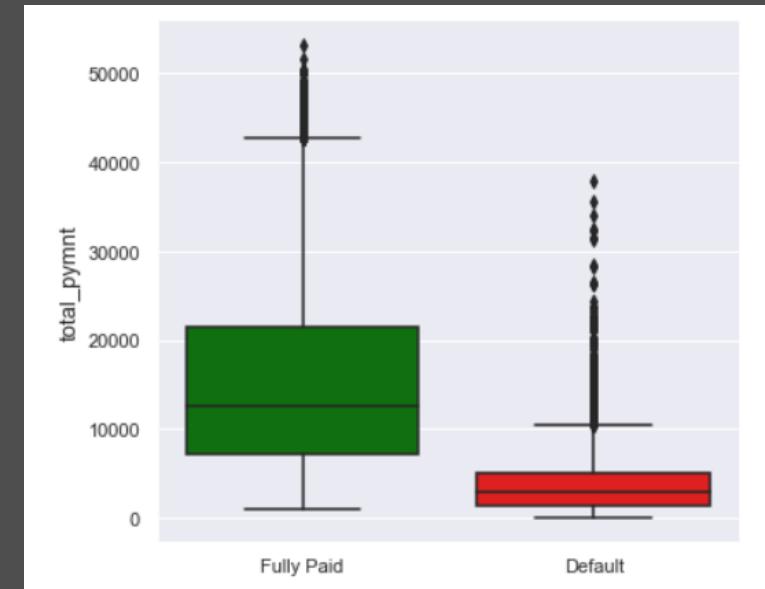
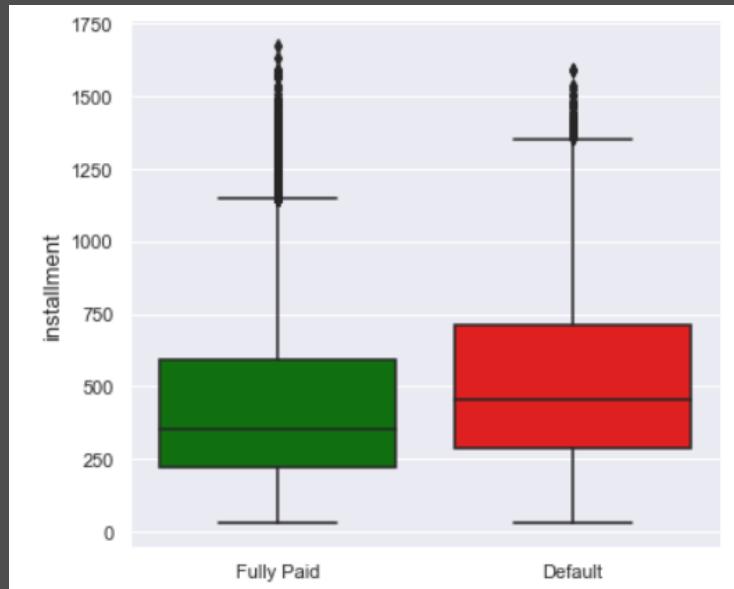
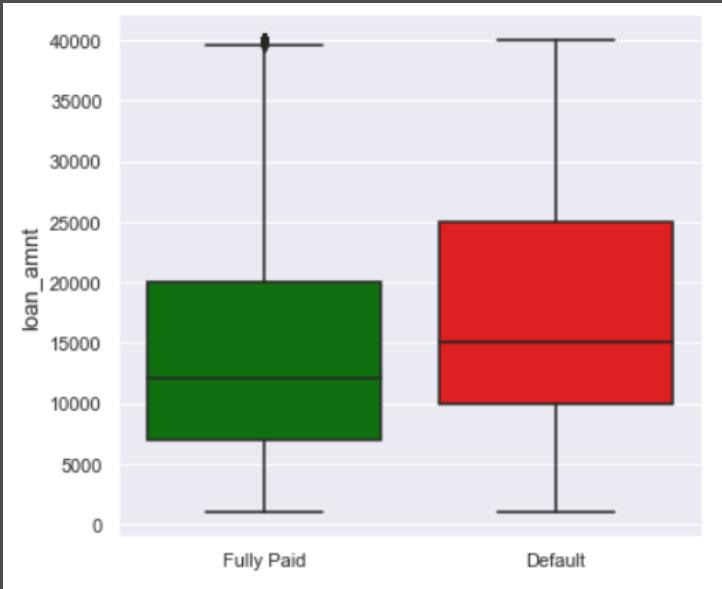
# Exploratory Data Analysis

---

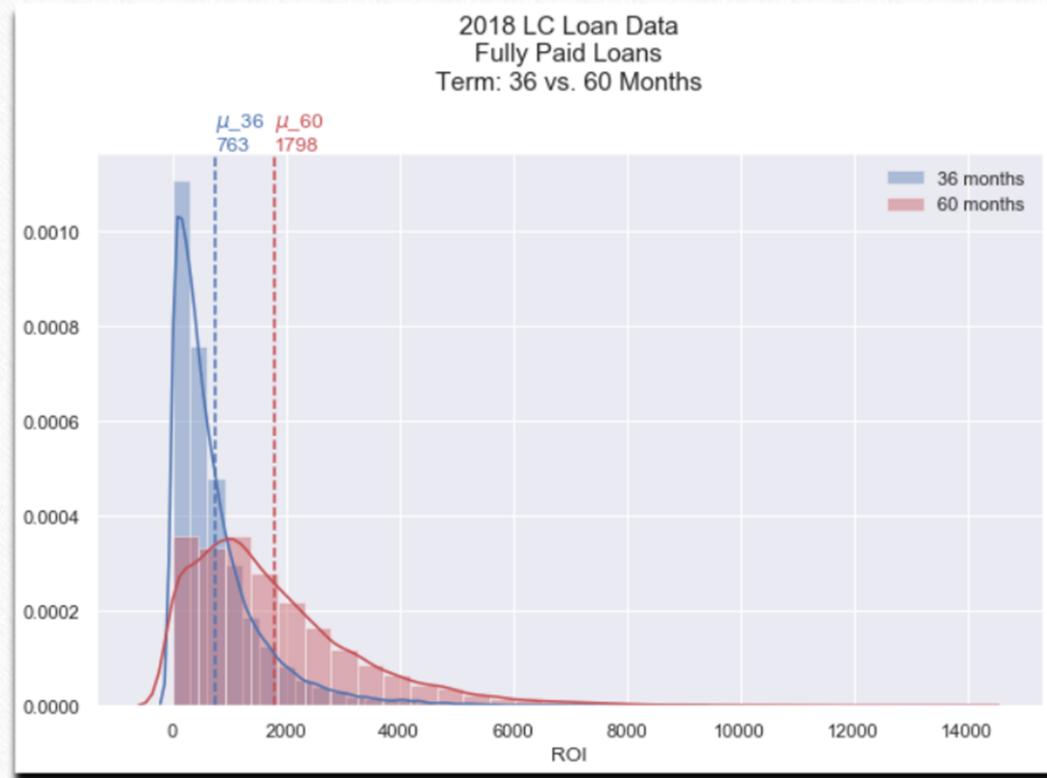
Descriptive Statistics & Inferential Statistics





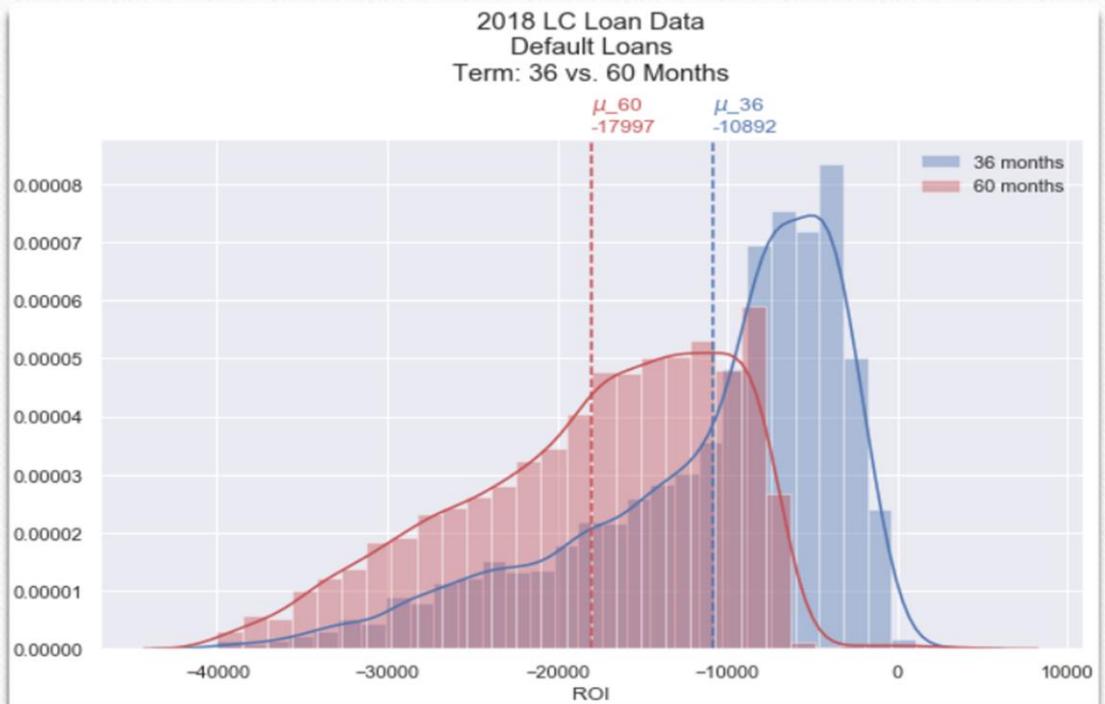


# Return on Investment (ROI)



75% Confident Interval for ROI

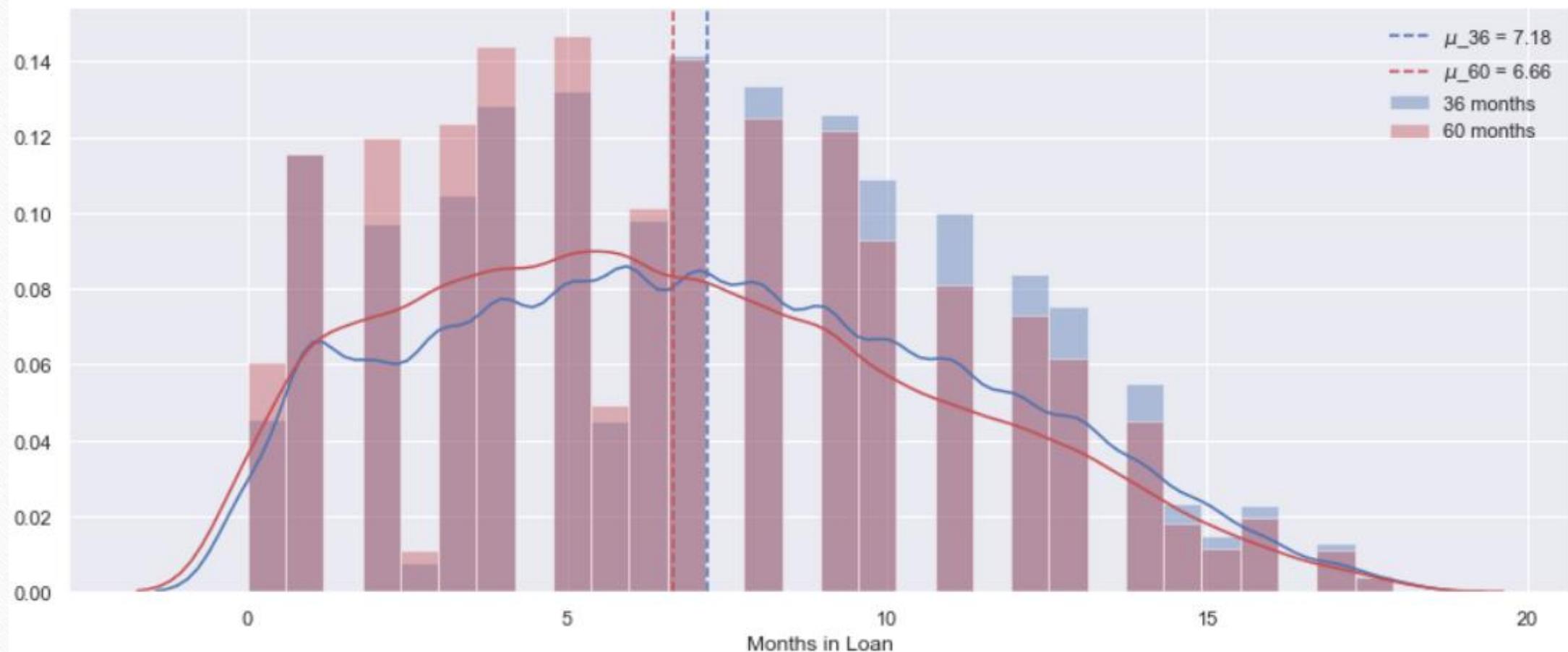
- 36 months: between 0 and 2,439 dollars
- 60 months: between 0 and 4,817 dollars



75% Confident Interval for Lost of Investment

- 36 months: between 0 and 26,661 dollars
- 60 months: between 2,270 and 33,724 dollars

2018 LC Loan Data  
Months in Loan  
Term: 36 vs. 60 Months



# Estimating Months-in-Loan for Selected Sample

---

## Binomial Distribution

- Probability of Default: 18.74%

Number of Loans	5	10	15	20	25	30	35	40	45	50
Expected No. of Defaults	1	2	3	4	5	6	7	7	8	9

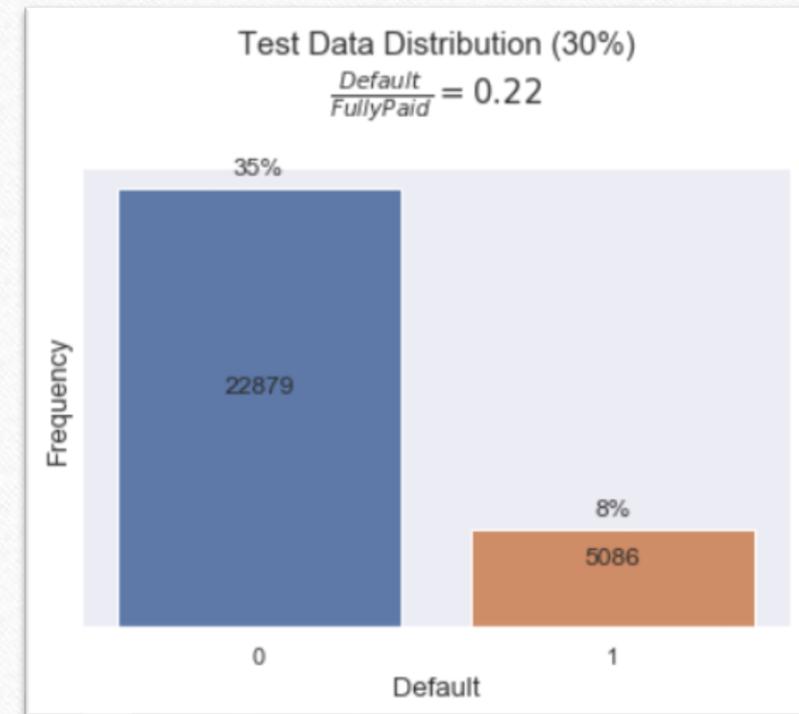
# Machine Learning Models

---

Random Forest Classifier: Classify Loan

Random Forest Regressor: Estimate Months-in-Loan

# Train Data vs. Test Data



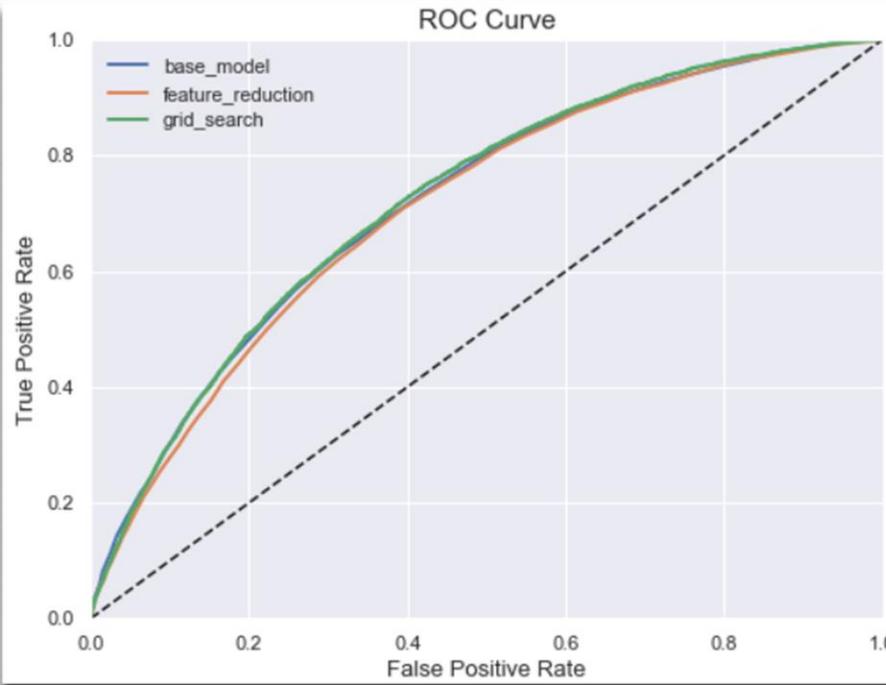
# Model Development

Base Model	Model 1	Model 2
<ul style="list-style-type: none"><li>• Use all features</li><li>• Use the following parameter: <pre>param = {'bootstrap': [True], 'n_estimators': [100]}</pre></li></ul>	<ul style="list-style-type: none"><li>• Use same parameter as base model.</li><li>• Use features with important score greater than 0.2.</li><li>• Exclude some features with strong correlation.</li></ul>	<ul style="list-style-type: none"><li>• Use same features are model 1.</li><li>• Use the best parameters from grid search.</li></ul>

# Random Forest Classifier: Classify Loan

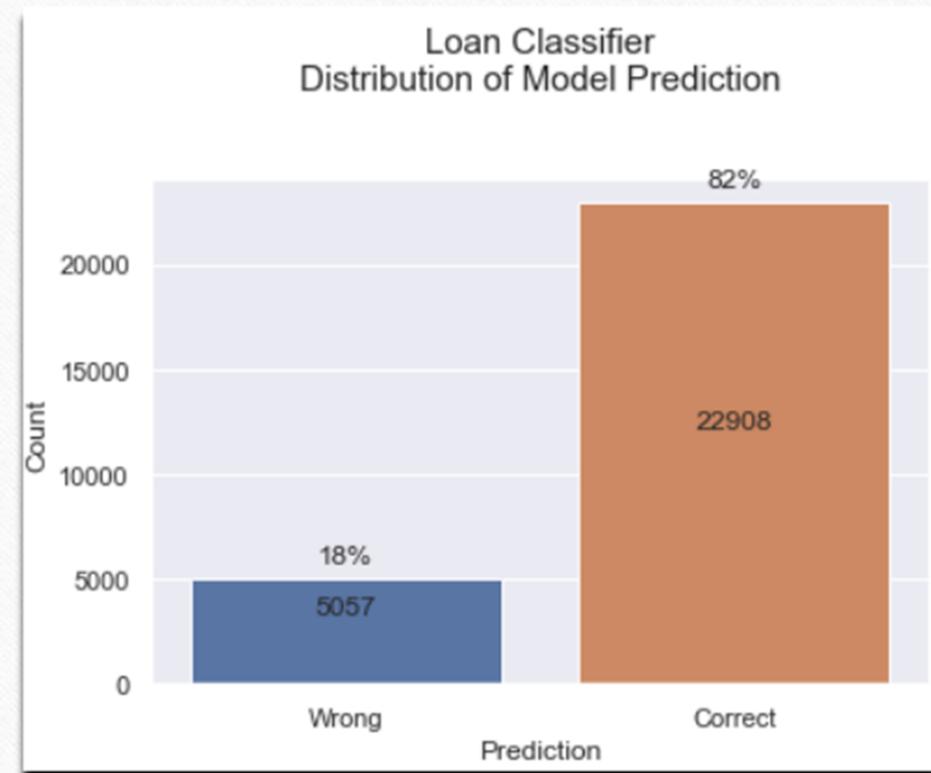
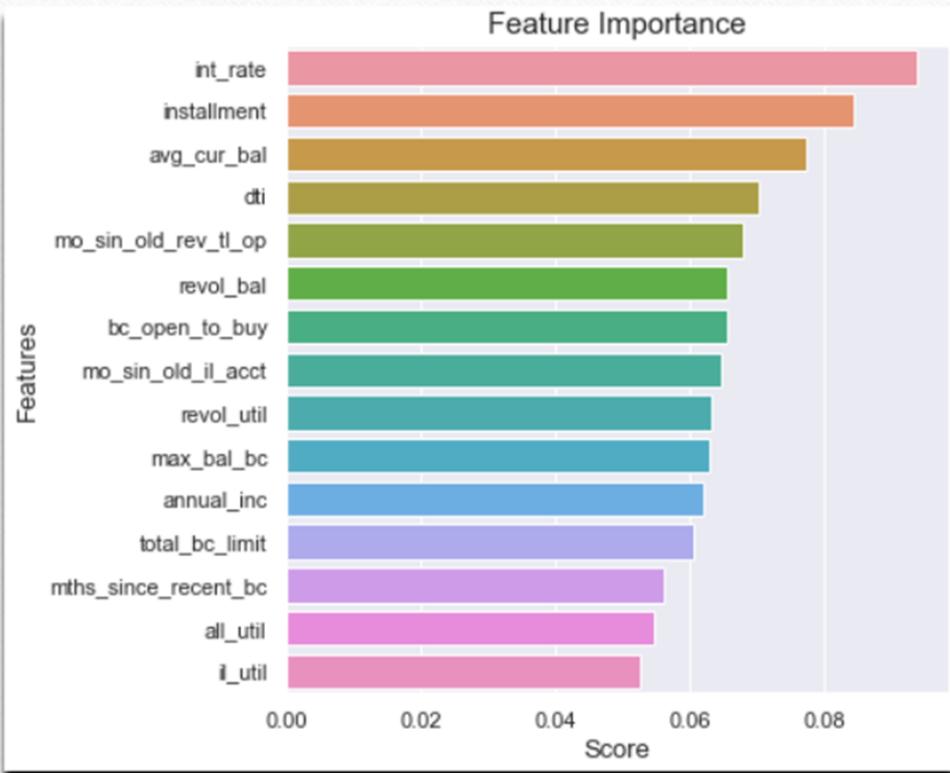
Model Evaluation & Selection | Model Prediction

# Loan Classifier: Model Evaluation

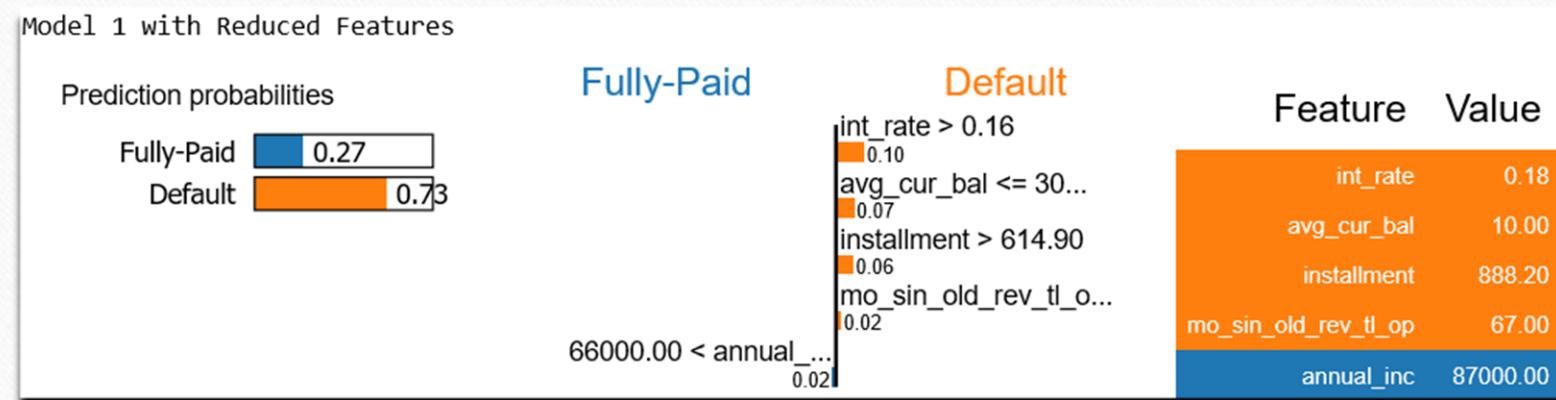
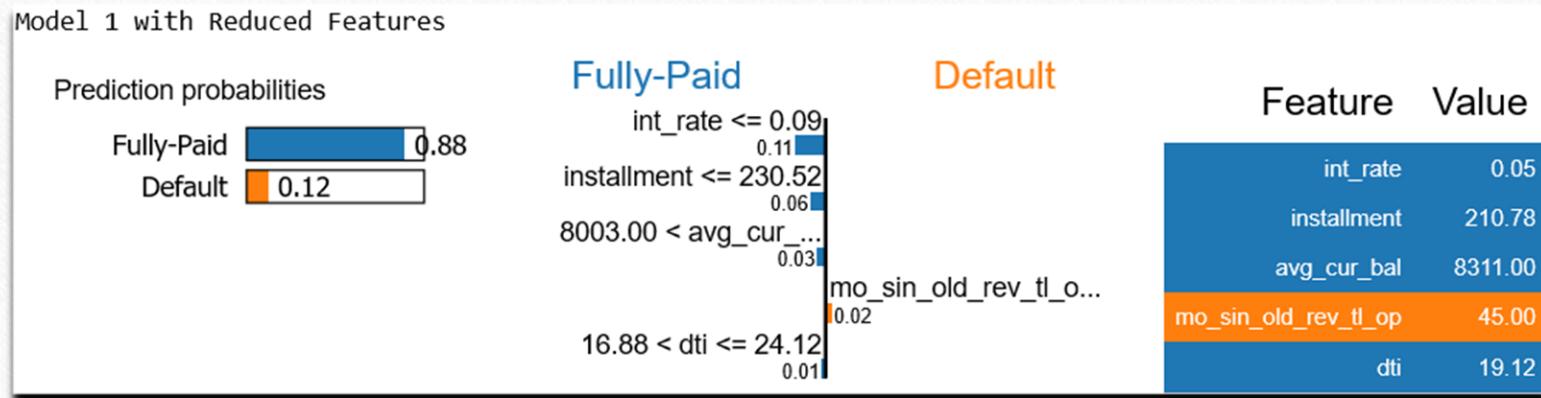


model_name	true_neg	true_pos	false_neg	false_pos	accuracy	precision	recall	f1_score	training_time
base_model	22692	241	4845	187	0.820061	0.563084	0.047385	0.087414	68
feature_reduction	22684	224	4862	195	0.819167	0.534606	0.044042	0.081381	41
grid_search	22777	145	4941	102	0.819667	0.587045	0.028510	0.054378	1179

# Loan Classifier: Model 1 (Feature Reduction)



# Default Prediction with Loan Classifier



# Random Forest Regressor: Estimate Months-in-Loan

---

Model Evaluation & Selection | Model Prediction

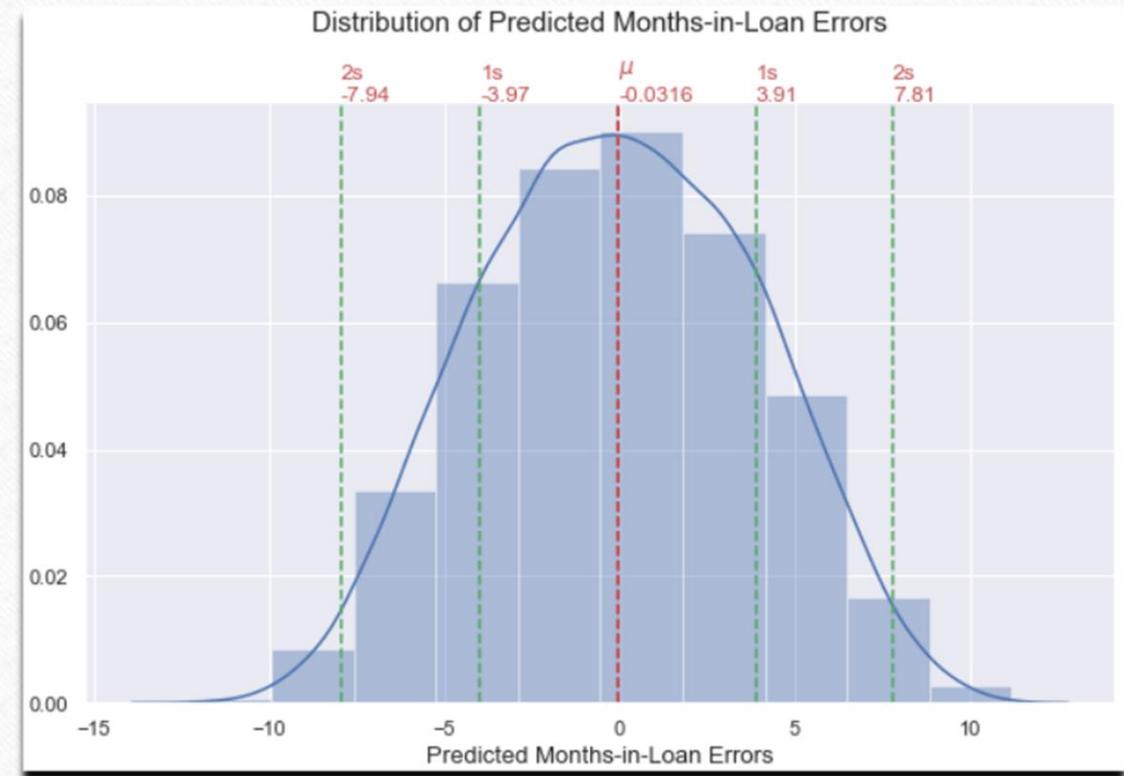
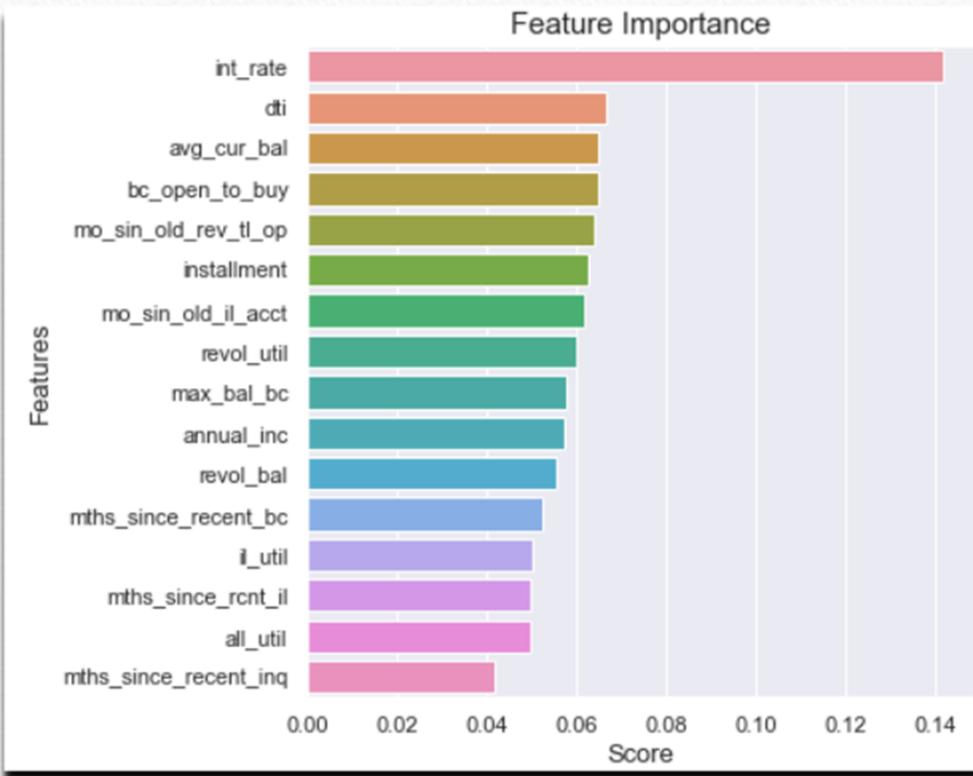
# Months-in-Loan Estimator Model Evaluation

---

<b>model_name</b>	<b>mse</b>	<b>rmse</b>	<b>mae</b>	<b>R-squared</b>	<b>training_time</b>
base_model	15.385022	3.922374	3.237386	0.871849	240.061900
feature_reduction	15.510222	3.938302	3.246415	0.871484	87.549685
grid_search	16.500856	4.062125	3.373749	0.109093	338.585933

# Months-in-Loan Estimator

## Model 1 (Feature Reduction)



# Model Prediction

loan_amnt	installment	actual_label	predict_label	prob_fully_paid	prob_default	actual_mil	predict_mil	predict_error	predict_mil_68cf	predict_mil_95cf
7000	210.78	fully-paid	fully-paid	0.88	0.12	7	9.9	2.9 (5.93, 13.81)	(1.96, 17.71)	
1600	48.74	fully-paid	fully-paid	0.98	0.02	15.9	9.3	-6.6 (5.33, 13.21)	(1.36, 17.11)	
20000	693.51	fully-paid	fully-paid	0.69	0.31	9	4.9	-4.1 (0.93, 8.81)	(0, 12.71)	
7000	240.85	fully-paid	fully-paid	0.77	0.23	3	5.3	2.3 (1.33, 9.21)	(0, 13.11)	
20000	684.33	fully-paid	fully-paid	0.84	0.16	14.9	7	-7.9 (3.03, 10.91)	(0, 14.81)	
10000	275.34	default	default	0.47	0.53	9.1	7.6	-1.5 (3.63, 11.51)	(0, 15.41)	
8000	251.4	default	fully-paid	0.97	0.03	1	4.2	3.2 (0.23, 8.11)	(0, 12.01)	
10000	381.13	default	default	0.48	0.52	7	6.8	-0.2 (2.83, 10.71)	(0, 14.61)	
6000	224.18	fully-paid	fully-paid	0.86	0.14	10	7	-3 (3.03, 10.91)	(0, 14.81)	
15000	481.06	fully-paid	fully-paid	1	0	11	7.6	-3.4 (3.63, 11.51)	(0, 15.41)	
8950	321.19	default	fully-paid	0.63	0.37	2	7.7	5.7 (3.73, 11.61)	(0, 15.51)	
4800	154.71	fully-paid	fully-paid	0.96	0.04	0	7.7	7.7 (3.73, 11.61)	(0, 15.51)	
4800	166.52	fully-paid	fully-paid	0.75	0.25	11	9	-2 (5.03, 12.91)	(1.06, 16.81)	
30000	606	fully-paid	fully-paid	0.8	0.2	5	7.1	2.1 (3.13, 11.01)	(0, 14.91)	
35000	745.03	default	fully-paid	0.79	0.21	3	5.2	2.2 (1.23, 9.11)	(0, 13.01)	
4000	152.46	default	fully-paid	0.83	0.17	3	8.1	5.1 (4.13, 12.01)	(0.16, 15.91)	

actual\_mil: actual months-in-loan

predict\_mil: predicted months-in-loan

predict\_mil\_68\_cf: 68% confident interval for predicted months-in-loan

predict\_mil\_95\_cf: 95% confident interval for predicted months-in-loan

# Recommendation to Investors

---

- Invest in a fractional part of a loan
- Invest in a loan that has
  - 36 months loan term
  - Low interest rate
  - Low installment
  - Low average current balance of all accounts

**Thank You!**