

---

# **Risky Loaners Model**



---



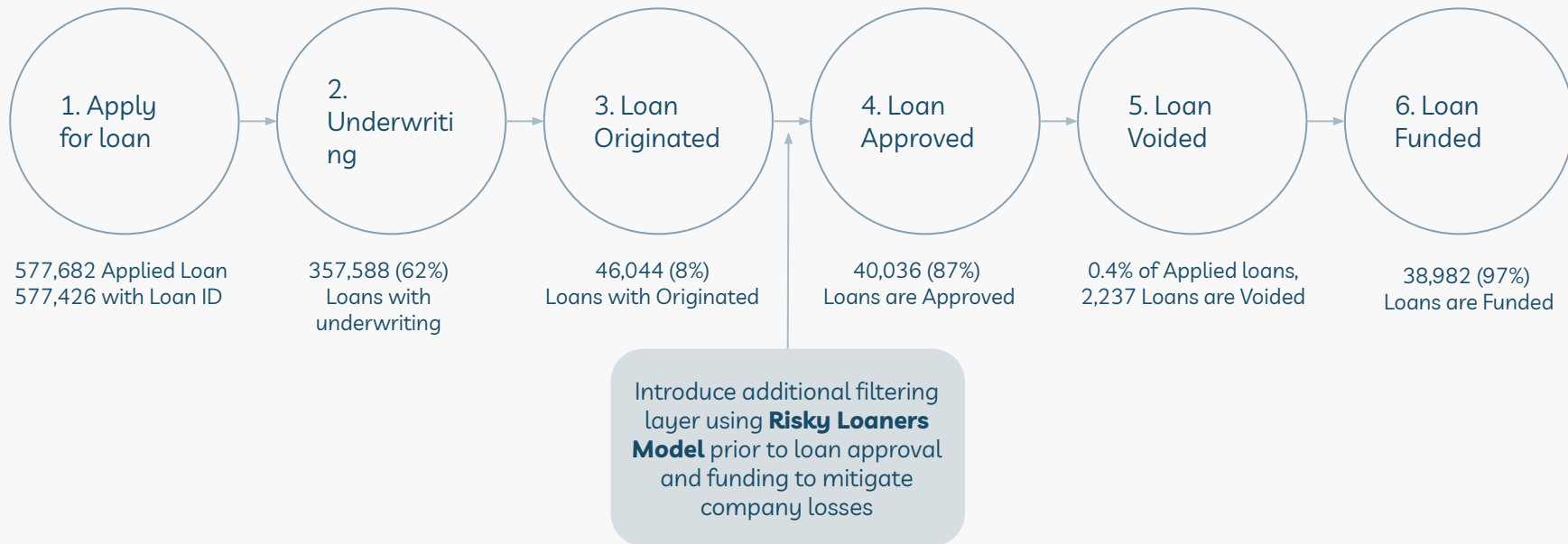
# Introduction



---

# Loan Application Procedure

How to identify potential bad loaners, whom having high loan risk from the loan application, before the loan is being approved and the loan is funded to the borrowers?



---

## Define Prediction Target : Good and Bad Loan Grouping

The distribution of Good Loan vs Bad Loan are **41%** (12,135) vs **59%** (17,383).

### Good Loan

Loans that are paid according to the agreed term and predefined schedule as set by the lender.

#### Loan Status:

Paid Off Loan  
Settlement Paid Off


### Bad Loan

Loans that carry a risk of default or are unlikely to be repaid according to the agreed term and predefined schedule. This leads to the possibility of cost incurred on third parties (i.e., Administrator, Legal fees) for late payments.


#### Loan Status:

Charged Off  
Paid Off  
Charged Off  
Internal Collection  
External Collection  
Settled Bankruptcy

---



---



# **Analytics Model Methodology**



---

---

## Solution Overview



# Engineered Features

**11** additional features are created to enrich data input and **7** categorical features are encoded.

	Payment	Loan	Clarity Writing	Recommended
Profile		-State*		-Income -Employment Status
Loan		-Lead Type* -Pay Frequency*		
Credit Score			-Inquiry on file current address conflict** -More than 3 inquiries in the last 30 days** -Overall Match Result** -Name Address Match**	
Credit History	-Difference between Paid Off Amount and Original -Scheduled Payment Amount -Existing Debt Amount (RM) -Median of Success Payment (RM) -Median of Failed Payment (RM) -# Previous Loan -# Previous Bad Loan -# Previous is Collection -# Payment Count -# of Success Payment -# of Failed Payment -Ratio of Failed over Success Payment			-Debt to Income Ratio -Credit Card Payment History  <i>These features are not included as they are not available</i>

\* One hot encoding is performed and Chi-Square Test is used to further select features that has relationship with the target

\*\* Label encoding is performed

# Strategies to Handle Missing Values

A total **5,118** missing records were dropped with **24,400** records remained.

## Data Collected

### Payment

### Loan

### Clarity Writing

### Current

Records with missing values are dropped.

### Recommended

K-means missing values imputation methodology, provided with the following information:

-More business context how Clarity Writing features (E.g. Clear Fraud Score) are calculated

-Collect other dataset, such as transaction information, credit card payment historical records, if available

Incomplete  
Dataset

Complete  
Dataset

-# of Paid Off Loan (21)

-# of unique inquires for the consumer seen by Clarity (5,034)  
-# fraud indicators (5,049)  
-Max # of unique SSNs with any bank account (5,049)  
-Clear Fraud Score (5,116)  
-Inquiry/on-file current address conflict, T/F (5,066)  
-Name Address Match (5,058)

-Difference between Paid Off Amount and Original Scheduled Payment Amount  
-Existing Debt Amount (RM)  
-Median of Success Payment (RM)  
-Median of Failed Payment (RM)  
-# Previous Loan  
-# Previous Bad Loan  
-# Previous is Collection  
-# Payment Count  
-# of Success Payment  
-# of Failed Payment

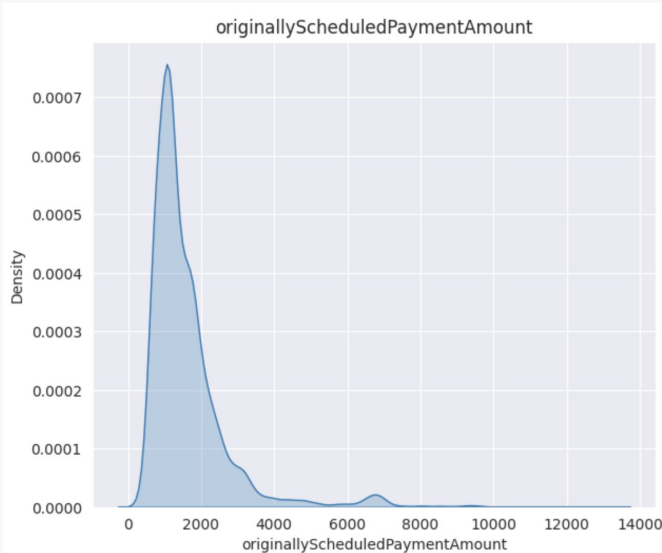
-Minutes from application to originated  
-# of Paid Off Loan  
-Annual Percentage Rate  
-Loan Amount  
-Original Scheduled Payment Amount  
-Lead Cost  
-Pay Frequency  
-State  
-Lead Type



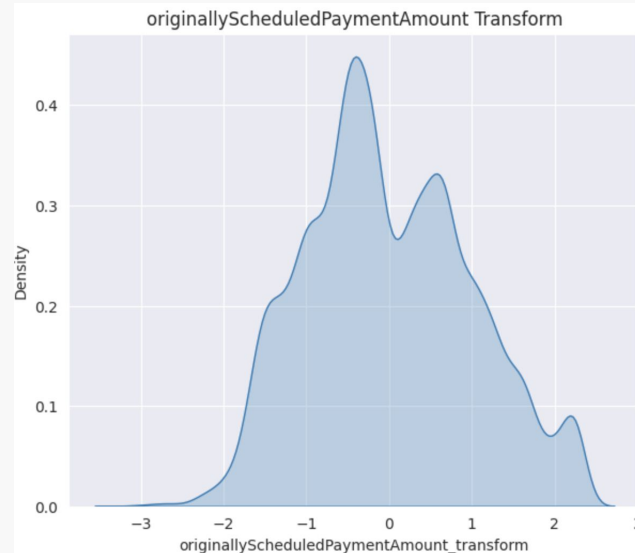
# Data Transformation for Model Building Experiment

**15** statistically non-normal features are transformed to a **more normally distributed** features.

## Before



## After



### Explanation:

1. Perform Normality Test on Numerical Features
2. For non-Normal features, perform following data transformation using following steps
  - a. Standard scaling the feature to have mean of 0 and standard deviation of 1
  - b. Min max scaling the feature to have a min of 0 and max of 1
  - c. Perform Yeo-Johnson transformation on the feature

# Preliminary Data Exploration

6 pairs of **highly correlated features** are identified from Correlation Analysis.

## Non-normal Dataset

	level_0	level_1	correlation (Original)
34	thirtydaysago	ninetydaysago	0.802316
37	thirtydaysago	fifteendaysago	0.869999
167	sevendaysago	fifteendaysago	0.885578
346	loanAmount	originallyScheduledPaymentAmount	0.932922
446	prevLoan_cnt	prevPaidOff_cnt	0.992360

## Transformed Dataset

	level_0	level_1	correlation (Transformed)
34	thirtydaysago	ninetydaysago	0.802316
37	thirtydaysago	fifteendaysago	0.869999
162	sevendaysago	fifteendaysago	0.885578
333	loanAmount_transform	originallyScheduledPaymentAmount_transform	0.910899
382	existDebt_amt_transform	prevPaidOff_cnt_transform	-0.875509
406	medFailed_paymentAmt_transform	numFailed_payment_transform	0.999999

	level_0	level_1	correlation (Original, removed derived features)
26	thirtydaysago	fifteendaysago	0.866266
101	sevendaysago	fifteendaysago	0.870926
181	loanAmount	originallyScheduledPaymentAmount	0.948468

	level_0	level_1	correlation (Transformed, removed derived features)
26	thirtydaysago	fifteendaysago	0.866266
101	sevendaysago	fifteendaysago	0.870926
181	loanAmount_transform	originallyScheduledPaymentAmount_transform	0.915783

## Explanation:

1. Number of unique inquiries for the consumer seen by clarity in the last 30 days is positively correlated to 90 days and 15 days
2. Number of unique inquiries for the consumer seen by clarity in the last 15 days is positively correlated to 30 days and 7 days
3. Loan Amount is positively correlated to originally Scheduled Payment Amount, apply to before transformed and after transformed

### Before Transform :

1. Previous loan count is positively correlated to Previous paid off loan count

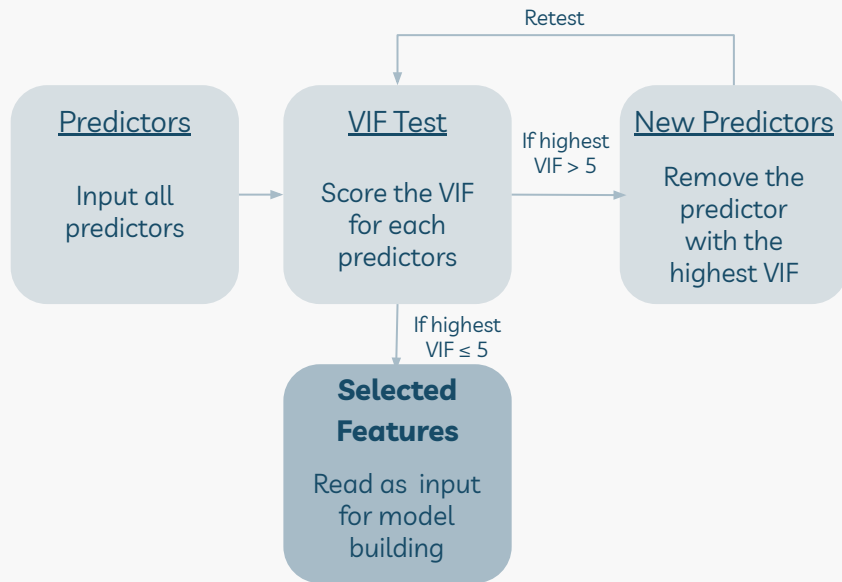
### After Transform :

1. Existing loan debt amount (RM) is negatively correlated to Previous paid off loan count
2. Median of Failed Payment Amount (RM) is positively correlated to Number of Failed payment

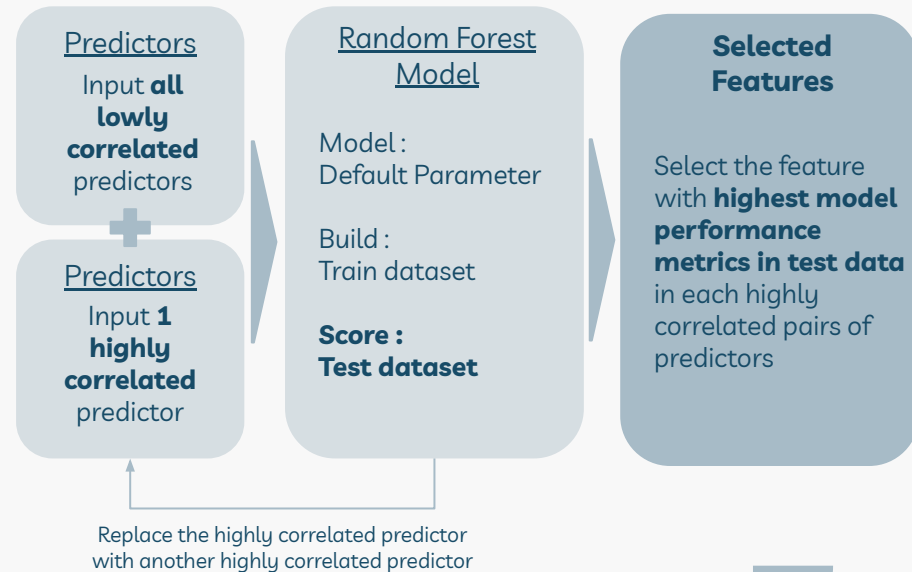
Not First  
LoanFirst  
Loan

# Highly correlated Predictors are removed to prevent multicollinearity for model building

## Approach A : VIF

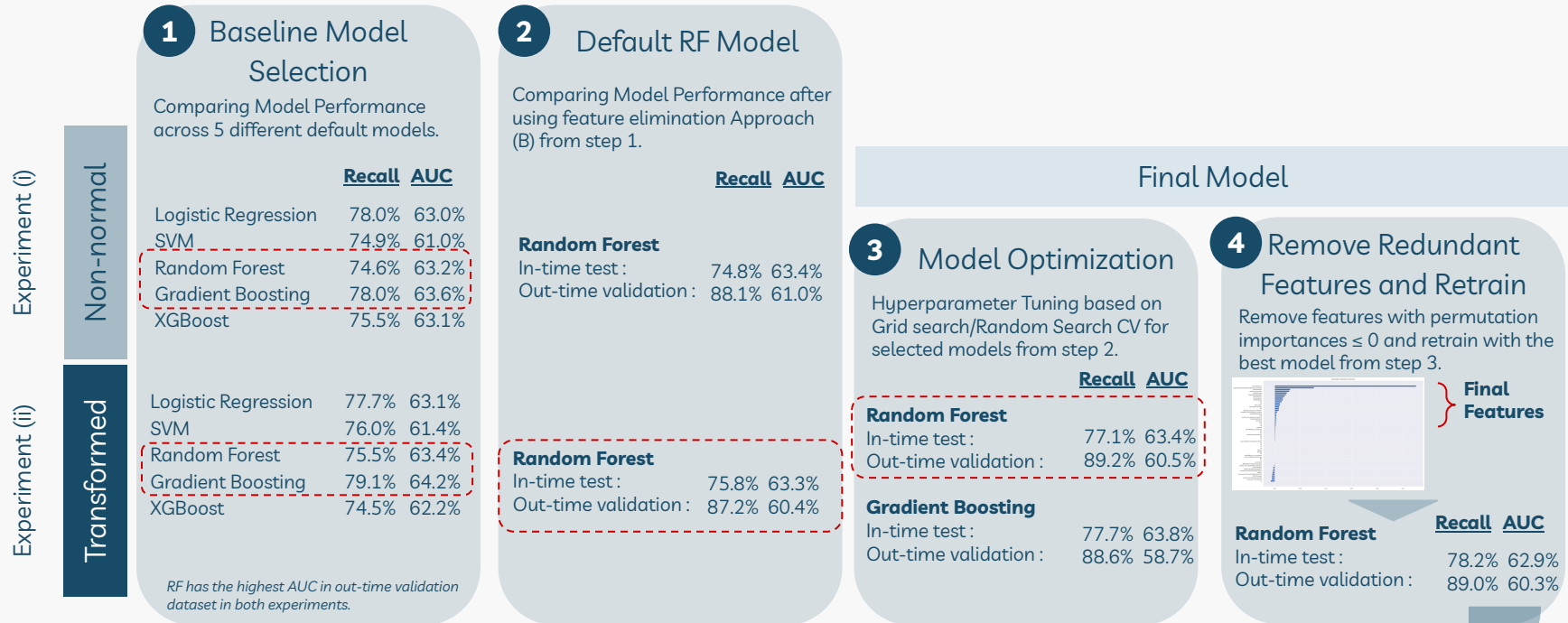


## Approach B : Performance Metrics



# Random Forest model has been selected as Final Model

Recall metric increased by **2.4%** in in-time test and **1.8%** in out-time validation upon Model Optimization.

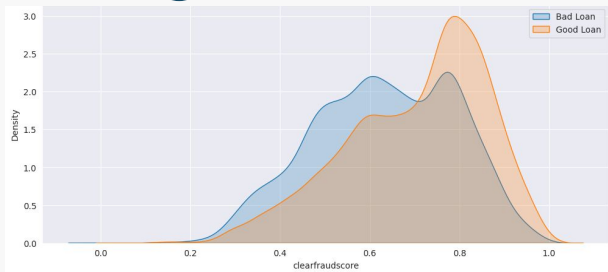


# Top 5 Features Distribution Plot based on Importances

High feature importance has more **distinctive distribution** between good and bad loans.

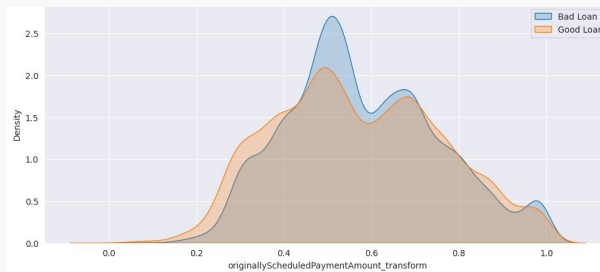
1

Clear Fraud Score



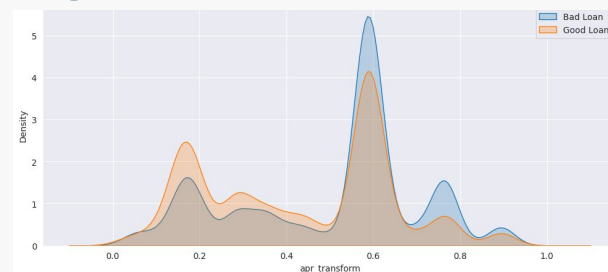
2

Originally Scheduled Payment Amount (Transformed)



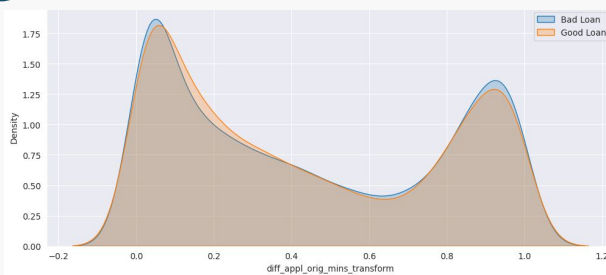
3

Annual Percentage Rate (Transformed)



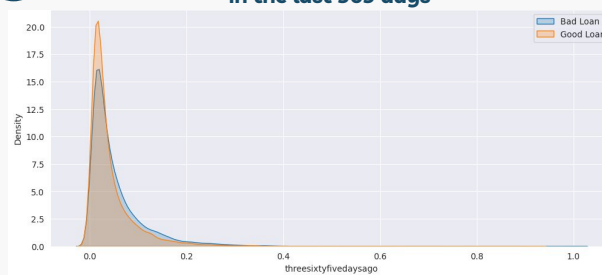
4

Minutes from Application to Originated (Transformed)



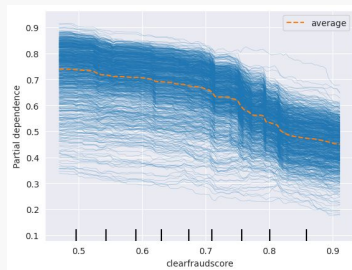
5

# of unique inquires for the consumer seen by Clarity in the last 365 days

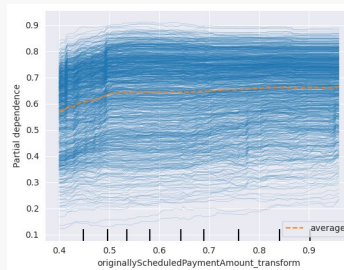


# Top 5 Features Partial Dependence Plot based on Importances

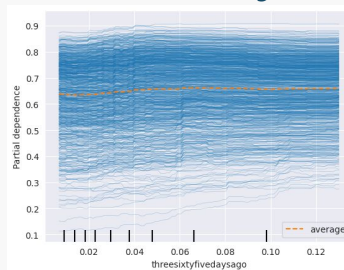
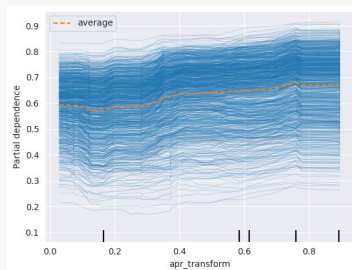
## 1 Clear Fraud Score



## 2 Originally Scheduled Payment Amount (Transformed)



## 3 Annual Percentage Rate (Transformed) 5 # of unique inquires for the consumer seen by Clarity in the last 365 days



### Explanation:

From the below Partial Dependence Plot, it is observed that

1. The **higher** the Clear Fraud Score, the **lower** the loan risk
2. The **higher** the Originally Scheduled Payment Amount Transform, the **higher** the loan risk
3. The **higher** the Annual Percentage Rate (%), the **higher** the loan risk
4. The **higher** the Number of unique inquiries for the consumer seen by Clarity in the last 365 days, the **higher** the loan risk

---



# **Business Impact**



# \$246k of Net Profit is Generated with Risky Loaners Model in 2017 Q1

Without Model

Actual	Predicted	Loan Status	Loan Count	Loan Amount (USD)	Amount Collected as Scheduled Originally (USD)	Recovery Rate	Costs Rate	Revenue (USD)	Cost (USD)	Total (USD)
Good Loan	Good Loan	Paid Off Loan	326	\$281,007	\$739,503	100%	0%	\$458,496		\$458,496
Good Loan	Good Loan	Settlement Paid Off	1	\$400	\$1,336	100%	0%	\$936		\$936
Good Loan	Bad Loan	Paid Off Loan	693	\$487,503	\$1,480,469	100%	0%	\$992,966		\$992,966
Good Loan	Bad Loan	Settlement Paid Off	13	\$9,600	\$29,860	100%	0%	\$20,260		\$20,260
Bad Loan	Good Loan	Charged Off Paid Off	6	\$10,350	\$24,205	100%	20%		\$2,070	-\$2,070
Bad Loan	Good Loan	External Collection	8	\$5,075	\$13,244	60%	30%		\$3,553	-\$3,553
Bad Loan	Good Loan	Internal Collection	340	\$370,501	\$853,812	60%	15%		\$203,776	-\$203,776
Bad Loan	Good Loan	Settled Bankruptcy	8	\$7,500	\$15,954	60%	20%		\$4,500	-\$4,500
Bad Loan	Bad Loan	Charged Off	1	\$1,800	\$5,138	0%	20%		\$2,160	-\$2,160
Bad Loan	Bad Loan	Charged Off Paid Off	4	\$5,800	\$18,078	100%	20%		\$1,160	-\$1,160
Bad Loan	Bad Loan	External Collection	225	\$121,450	\$383,334	60%	30%		\$85,015	-\$85,015
Bad Loan	Bad Loan	Internal Collection	2678	\$1,793,106	\$5,241,509	60%	15%		\$986,208	-\$986,208
Bad Loan	Bad Loan	Settled Bankruptcy	22	\$13,000	\$34,647	60%	20%		\$7,800	-\$7,800

**Net Profit** \$176,417

With Model

Actual	Predicted	Loan Status	Loan Count	Loan Amount (USD)	Amount Collected as Scheduled Originally (USD)	Recovery Rate	Costs Rate	Revenue (USD)	Cost (USD)	Total (USD)
Good Loan	Good Loan	Paid Off Loan	326	\$281,007	\$739,503	100%	0%	\$458,496		\$458,496
Good Loan	Good Loan	Settlement Paid Off	1	\$400	\$1,336	100%	0%	\$936		\$936
Good Loan	Bad Loan	Paid Off Loan	693	\$487,503	\$1,480,469	100%	0%	\$0		\$0
Good Loan	Bad Loan	Settlement Paid Off	13	\$9,600	\$29,860	100%	0%	\$0		\$0
Bad Loan	Good Loan	Charged Off Paid Off	6	\$10,350	\$24,205	100%	20%		\$2,070	-\$2,070
Bad Loan	Good Loan	External Collection	8	\$5,075	\$13,244	60%	30%		\$3,553	-\$3,553
Bad Loan	Good Loan	Internal Collection	340	\$370,501	\$853,812	60%	15%		\$203,776	-\$203,776
Bad Loan	Good Loan	Settled Bankruptcy	8	\$7,500	\$15,954	60%	20%		\$4,500	-\$4,500
Bad Loan	Bad Loan	Charged Off	1	\$1,800	\$5,138	0%	20%		\$0	\$0
Bad Loan	Bad Loan	Charged Off Paid Off	4	\$5,800	\$18,078	100%	20%		\$0	\$0
Bad Loan	Bad Loan	External Collection	225	\$121,450	\$383,334	60%	30%		\$0	\$0
Bad Loan	Bad Loan	Internal Collection	2678	\$1,793,106	\$5,241,509	60%	15%		\$0	\$0
Bad Loan	Bad Loan	Settled Bankruptcy	22	\$13,000	\$34,647	60%	20%		\$0	\$0

**Net Profit** \$245,534

## Assumptions:

Given that Recovery Rate for bad loans are set as 60%.

Costs for bad loans includes legal fees, external agency costs, administrative costs, write-offs and provisions, as well as unrecovered loan amounts.

Charged Off Paid Off are assumed to be 100% recovered, with no additional revenue generated.



---



# Next Step



---

---

# Implementation Plan

## Model Deployment



Final Model Registry



CI/CD Pipeline



Code Repository

## Operationalizing Predictive Model



Prediction Schedule

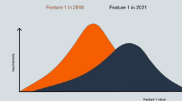


Prediction Result

## Production Model Monitoring



Model Performance Metrics



Features Drift

---



**Thank You**



---