# 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?

1. Apply for loan → 2. Underwriting → 3. Loan Originated → 4. Loan Approved → 5. Loan Voided → 6. Loan Funded

- 1. Apply for loan: 577,682 Applied Loan, 577,426 with Loan ID
- 2. Underwriting: 357,588 (62%) Loans with underwriting
- 3. Loan Originated: 46,044 (8%) Loans with Originated
- 4. Loan Approved: 40,036 (87%) Loans are Approved
- 5. Loan Voided: 0.4% of Applied loans, 2,237 Loans are Voided
- 6. Loan Funded: 38,982 (97%) Loans are Funded

Introduce additional filtering layer using **Risky Loaners Model** prior to loan approval and funding to mitigate company losses

# 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

- Feature Engineering
- Data Transformation
- Features Selection
- Model Building
- Data Insights

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

** 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 |
|---|---|---|---|
| Incomplete Dataset | | -# of Paid Off Loan (21) | -# of unique inquires for the consumer seen by Clarity (5,034)<br>-# fraud indicators (5,049)<br>-Max # of unique SSNs with any bank account (5,049)<br>-Clear Fraud Score (5,116)<br>-Inquiry/on-file current address conflict, T/F (5,066)<br>-Name Address Match (5,058) |
| Complete Dataset | -Difference between Paid Off Amount and Original Scheduled Payment Amount<br>-Existing Debt Amount (RM)<br>-Median of Success Payment (RM)<br>-Median of Failed Payment (RM)<br>-# Previous Loan<br>-# Previous Bad Loan<br>-# Previous is Collection<br>-# Payment Count<br>-# of Success Payment<br>-# of Failed Payment | -Minutes from application to originated<br>-# of Paid Off Loan<br>-Annual Percentage Rate<br>-Loan Amount<br>-Original Scheduled Payment Amount<br>-Lead Cost<br>-Pay Frequency<br>-State<br>-Lead Type | |

## Imputation Method

### Current

Records with missing values are dropped.

### Recommended

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

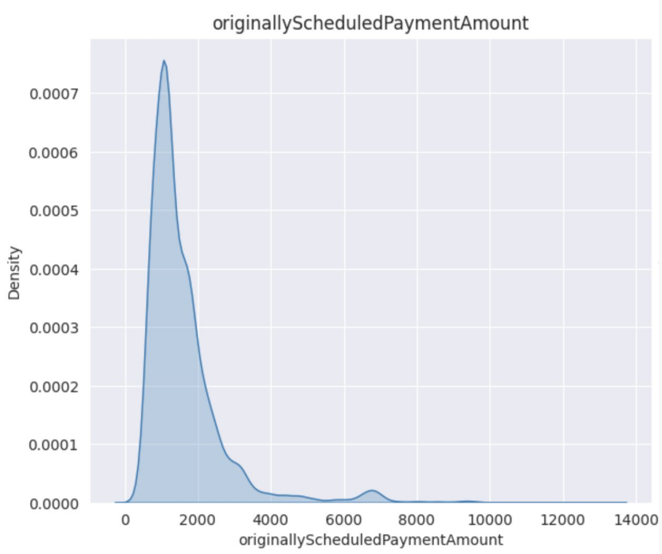
-More business context how Clarity Writing features (E.g. Clear Fraud Score)are calculated

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

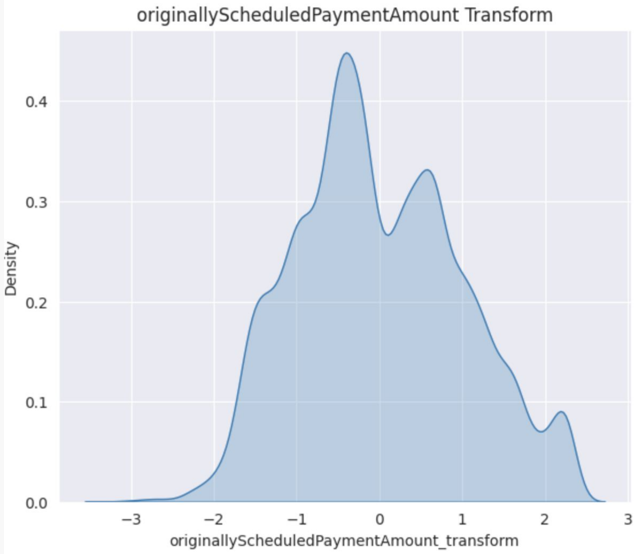
# Data Transformation for Model Building Experiment

**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

Not First Loan

| | 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 |

First Loan

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

### Transformed Dataset

Not First Loan

| | 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 |

First Loan

| | 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 :**

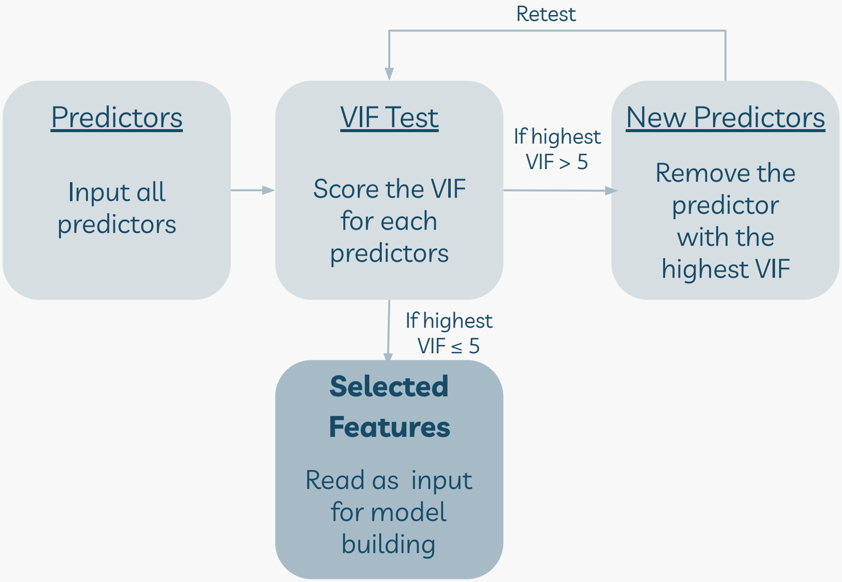
1. Previous loan count is positively correlated to Previous paid off loan count

**After Transform :**

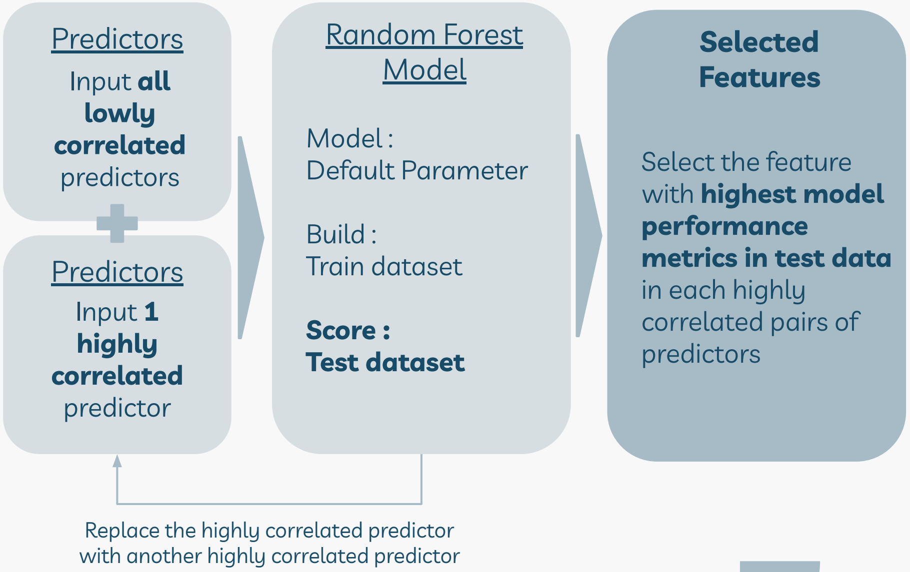
1. Existing loan debt amount (RM) is negatively correlated to Previous paid off loan count
2. Median of Failed Payment Amount (RM) is positively correlated to Number of Failed payment

# 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.

## 1 Baseline Model Selection

Comparing Model Performance across 5 different default models.

| Experiment | Model | Recall | AUC |
|---|---|---|---|
| Experiment (i) Non-normal | Logistic Regression | 78.0% | 63.0% |
| | SVM | 74.9% | 61.0% |
| | Random Forest | 74.6% | 63.2% |
| | Gradient Boosting | 78.0% | 63.6% |
| | XGBoost | 75.5% | 63.1% |
| Experiment (ii) Transformed | Logistic Regression | 77.7% | 63.1% |
| | SVM | 76.0% | 61.4% |
| | Random Forest | 75.5% | 63.4% |
| | Gradient Boosting | 79.1% | 64.2% |
| | XGBoost | 74.5% | 62.2% |

*RF has the highest AUC in out-time validation dataset in both experiments.*

## 2 Default RF Model

Comparing Model Performance after using feature elimination Approach (B) from step 1.

| Random Forest | Recall | AUC |
|---|---|---|
| **Experiment (i)** | | |
| In-time test : | 74.8% | 63.4% |
| Out-time validation : | 88.1% | 61.0% |
| **Experiment (ii)** | | |
| In-time test : | 75.8% | 63.3% |
| Out-time validation : | 87.2% | 60.4% |

## Final Model

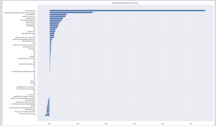
## 3 Model Optimization

Hyperparameter Tuning based on Grid search/Random Search CV for selected models from step 2.

| | Recall | AUC |
|---|---|---|
| **Random Forest** | | |
| In-time test : | 77.1% | 63.4% |
| Out-time validation : | 89.2% | 60.5% |
| **Gradient Boosting** | | |
| In-time test : | 77.7% | 63.8% |
| Out-time validation : | 88.6% | 58.7% |

## 4 Remove Redundant Features and Retrain

Remove features with permutation importances ≤ 0 and retrain with the best model from step 3.



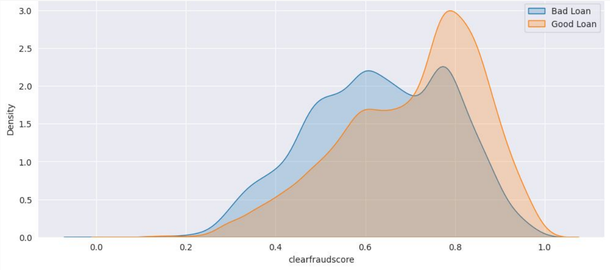
**Final Features**

| | Recall | AUC |
|---|---|---|
| **Random Forest** | | |
| In-time test : | 78.2% | 62.9% |
| Out-time validation : | 89.0% | 60.3% |

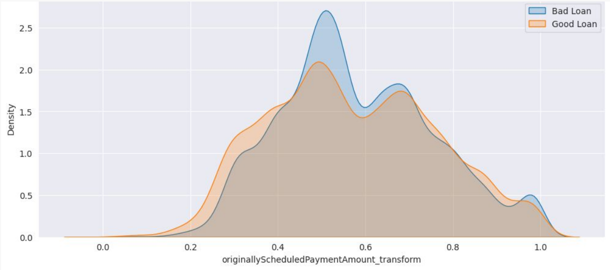
# Top 5 Features Distribution Plot based on Importances

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

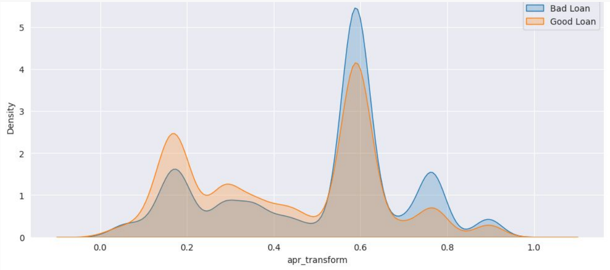
### 1 Clear Fraud Score



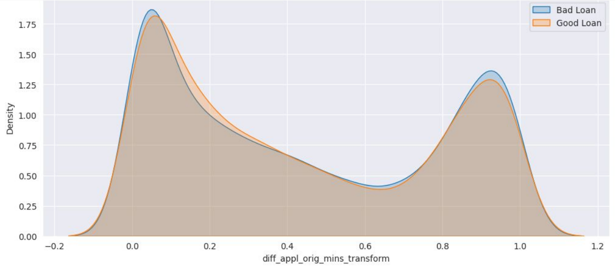

### 2 Originally Scheduled Payment Amount (Transformed)



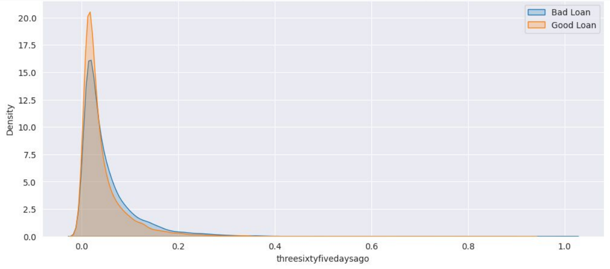

### 3 Annual Percentage Rate (Transformed)




### 4 Minutes from Application to Originated (Transformed)

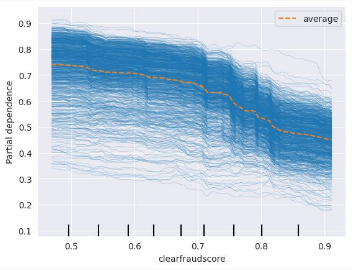



### 5 # of unique inquires for the consumer seen by Clarity in the last 365 days

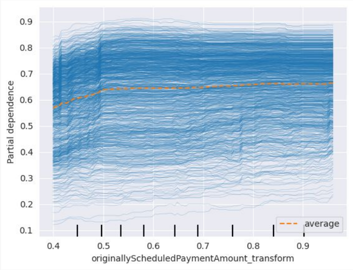


# Top 5 Features Partial Dependence Plot based on Importances

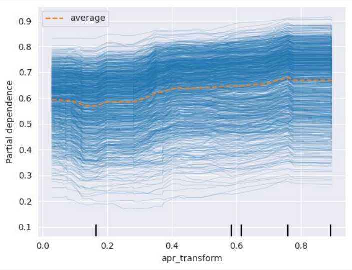
## 1 Clear Fraud Score



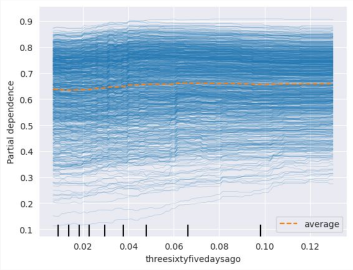

## 2 Originally Scheduled Payment Amount (Transformed)




## 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 Loan Risk 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

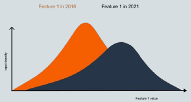
## Operationalizing Predictive Model

Prediction Schedule

Prediction Result

## Production Model Monitoring

Model Performance Metrics




Features Drift

# Thank You