Credit Scoring & Client Default Prediction

kaggle.com

MSDS-PT 2025B Term 2 LT 4

Co | Quiddaoen | Semual | Tan

Overview

Home Credit Business Model
Business Objectives

Fast Loans, Easy Installments!

Eyeing the latest cellphone, laptop, TV, refrigerator, or air conditioner? Simply buy now and pay later from over 100 trusted brands in Home Credit partner stores near you. Repay loan amounts over the counter or via your bank account. Enjoy our fast loan application and easy installments today.

Get Product Installments Today

Cash Loans? Get Fast Approvals Online!

Our online cash loans are ready for your needs. Existing customers who have previously availed of our installment shopping deals are eligible for our cash loan offers. Enjoy competitive interest rates and flexible payments! Pay in 12 months or more, depending on the loan.

Use our quick cash loans for your small business, home renovation, school tuition, or other needs. Check for offers, apply conveniently, and get fast approvals.

Check Cash Loan Offers

* New customers must first get our installment shopping deals to be eligible for cash loans.
Learn More

Chat with us

HOME CREDIT

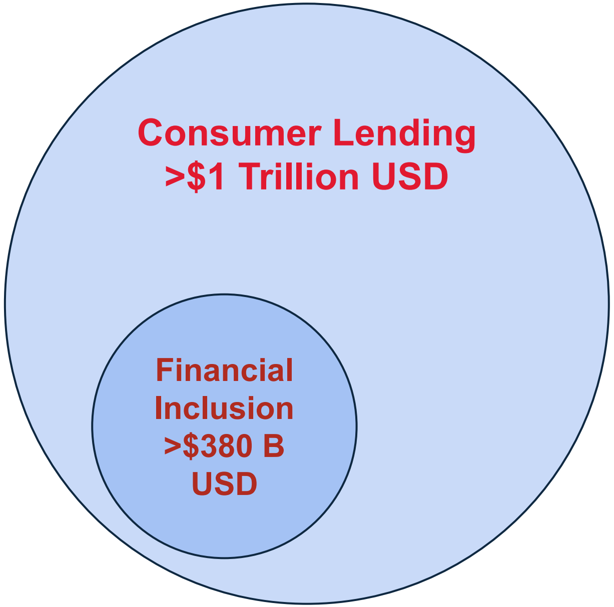
About the Company

An international consumer finance provider in multiple European & Asian Markets

With a focus on responsible lending primarily to individuals with little to no credit

https://www.homecredit.net/about-us.aspx/

Market Environment



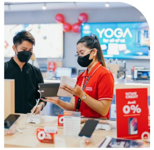
Consumer Lending Market Size, Share & Forecast, 2032 (businessresearchinsights.com)

Financial Inclusion, Accenture and CARE International UK Study

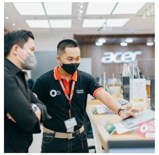
Market Environment

HOME CREDIT

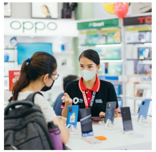
Broadening Financial Inclusion
1.152 Million
New Customers Added in 2022



Women Empowerment
572k
New Women Customers in 2022



Youth Empowerment
298k
Customers are under 30 years old



Digitalization & Innovation
76%
of loans processed digitally for POS

[Home Credit loans hit P296 billion | The Freeman (philstar.com)]
[https://www.homecredit.ph/about-home-credit]

Business Model

Versus Traditional Lending and Credit Scoring

Financial Inclusion

is achieved through

Easier Application Process

Less Rigid Metrics for Approval

Interest Rates Include Risk Premiums

Problem Statement

- **Credit History**

  Without traditional data, someone with little to no credit history is likely to be denied.

- **Dynamic Financial Market**

  Loan providers aren't able to spot potential problems any sooner…
  Stability in the future is critical, as a sudden drop in performance means that loans will be issued to worse clients on average.

- **Our Mission:**

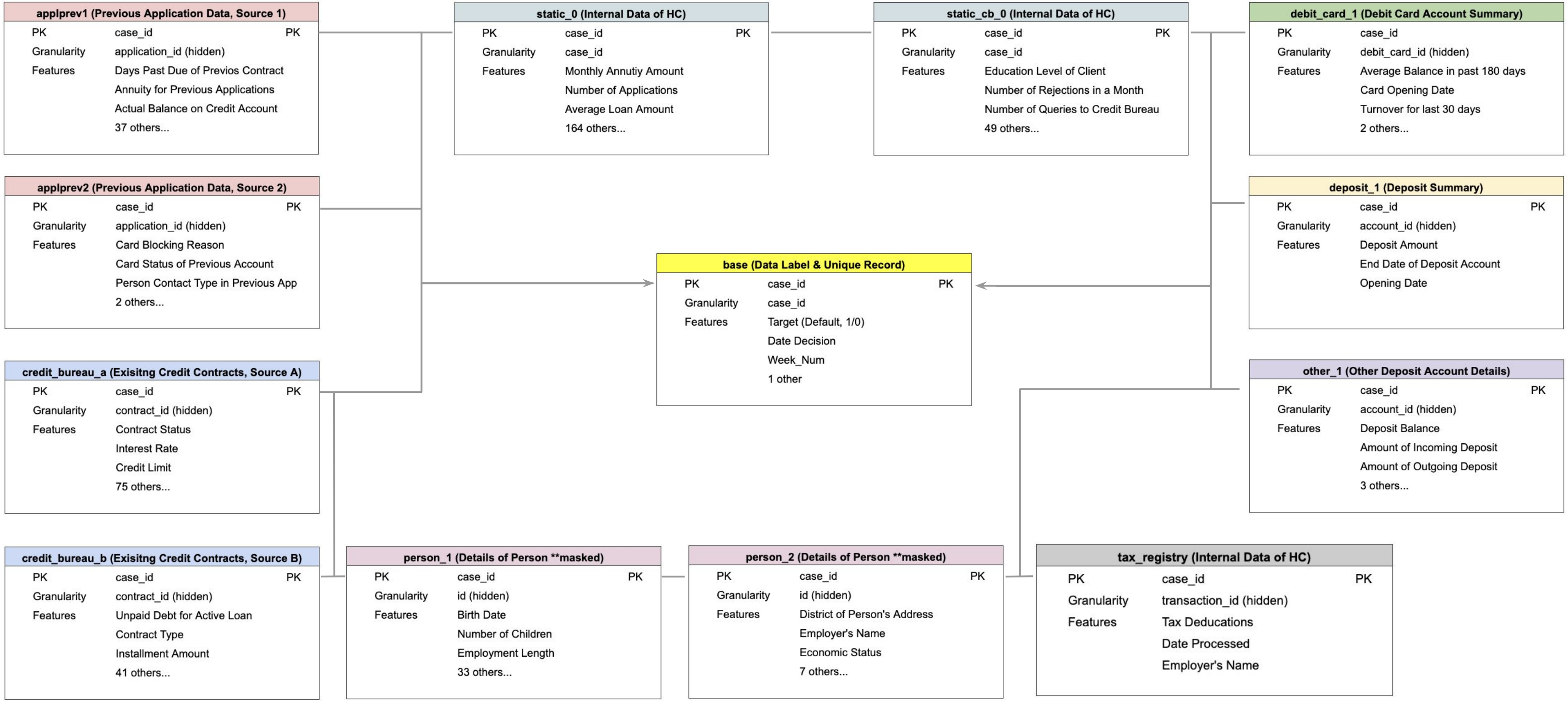
  Assess potential clients' default risks will enable consumer finance providers to accept more loan applications. This may improve the lives of people who have historically been denied due to lack of credit history.

https://www.kaggle.com/competitions/home-credit-credit-risk-model-stability

Model Development & Results

Dataset

Large amount and high dimension of data needed a faster and more automated way of extracting possible features.



Feature Engineering

Common Data Type Handling and Aggregation functions were applied on both numerical and categorical functions – Pipeline and Aggregator classes.

class Pipeline
(Data Type Handling)

class Aggregator
(Aggregating Low Granularity Data)

Numerical Columns
(int, float)

Extract Mean, Max, Median, Sum, Last First, Std

Categorical Columns
(string/objects)

Extract Last, First

Dates

Get Difference from Decision Date (Convert to Numeric)

1.53M IDs
774 Vars

HOME CREDIT

Study Base

Data was split into Train (80%) - Test (20%) set. Data is highly imbalanced.

	Size	Non-Defaulters	Defaulters	%Default
Training Set (80%)	1,221,327	1,182,886	38,441	3.15%
Test Set (20%)	305,332	295,779	9,553	3.13%
Total (100%)	1,526,689	1,478,665	47,994	3.14%

Target Labels: Default = 1, otherwise 0.

Feature Selection

Columns with zero to low usability, high complexity, and high correlation with other features will be dropped from the base table

Missing Values

- Features with greater than 95% Null values were excluded from the dataset

No Variance

- Features with Zero Variance were excluded from the dataset
- Numeric: std = 0
- Categorical: 1 unique value

Complex Variables

- Categorical Features with greater than 50 unique values were excluded from the dataset

Correlated Variables

- Highly Correlated (>0.9) Features were reduced based on Importance
- Used Correlation, Chi-square Test, and Anova F-Value

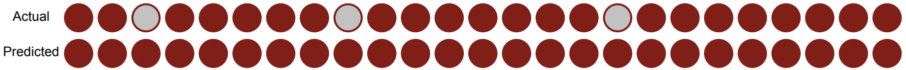
1.5M unique cases, 455 features

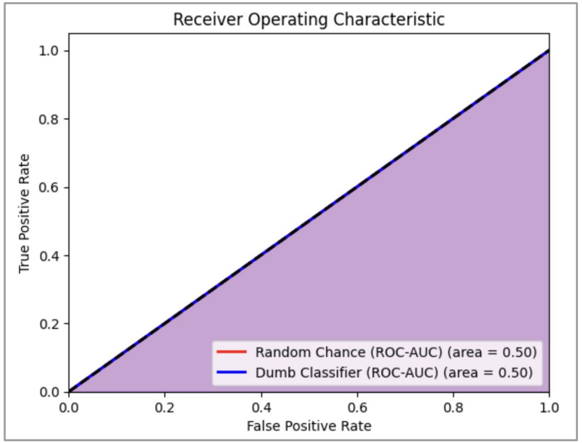
Baseline Models

A set of simple and non-informative models were created to provide a benchmark for the more optimized and sophisticated models.

Dumb Classifier

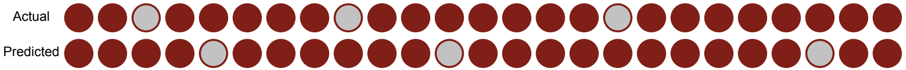
Predicts the majority class. Used to benchmark imbalanced data



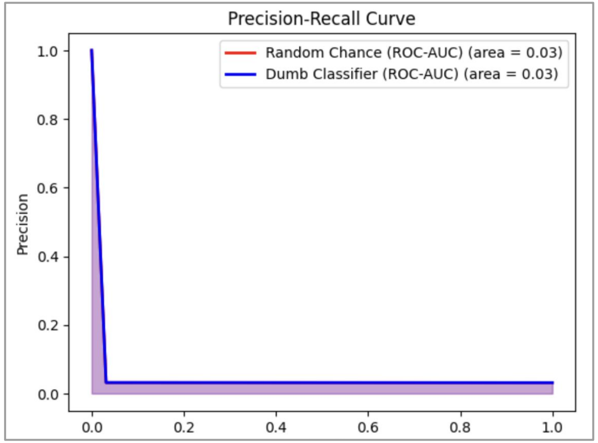


Random Chance Classifier

Randomly predicts the target variable based on event rate



Metrics	Dumb Classif	Random Chance
Accuracy	96.85%	93.96%
Precision	0.00%	3.19%
Recall	0.00%	3.13%
F1 Score	0.00%	3.16%
ROC-AUC	0.5000	0.5002
PR-AUC	0.0315	0.0315



Hyperparameter Tuning

Given a hyperparameter space, a Bayesian-based search algorithm was used to look for the parameters that will yield optimal results.

ML Algorithm

LightGBM & CatBoost

- Gradient Boosting Decision Trees
- Accepts Categorical Values
- Handles Missing Values
- Faster Training Times

Objective Function

Maximize Avg PR-AUC (5-Fold Cross Validation)

Search Algorithm

Hyperopt - Tree of Parzen (max of 250 trials per algo)

```
space = {
    'max_depth': hp.quniform("max_depth", 3, 10, 1),  #Max depth 10
    'gamma': hp.loguniform('gamma', np.log(1), np.log(100)),
    'reg_alpha': hp.uniform('reg_alpha', 0, 10),
    'reg_lambda': hp.uniform('reg_lambda', 0, 10),
    'colsample_bytree': hp.uniform('colsample_bytree', 0.5, 1),
    'min_child_weight': hp.loguniform('min_child_weight',np.log(1), np.log(100)),
    'n_estimators': hp.quniform('n_estimators', 50, 250, 1),
    'learning_rate': hp.loguniform('learning_rate', np.log(0.01), np.log(0.2)),
    'subsample': hp.uniform('subsample', 0.5, 1),
    'random_state': 42,
    'verbose': -1
}
```

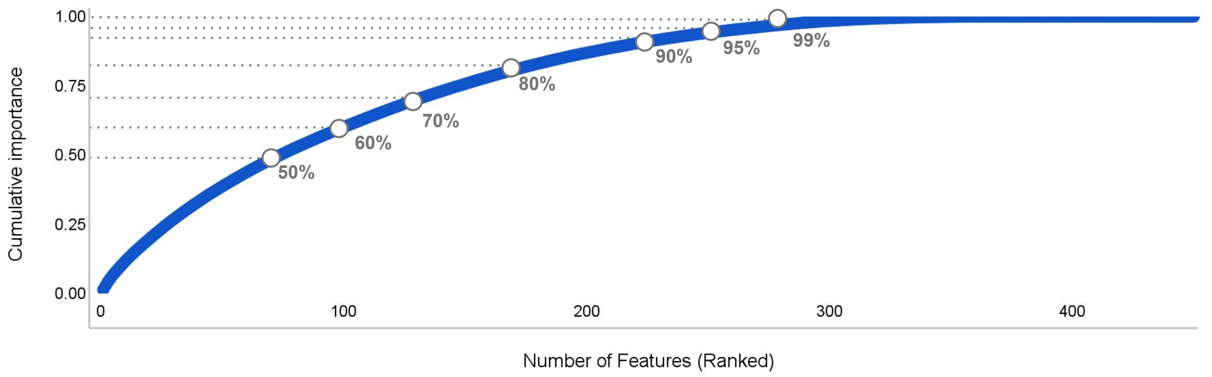
Average of 5-Fold Cross Validation (Training Dataset - 80%)			
model_name	ROC AUC	PR AUC	Gini
LGB_Tune34	0.84925	0.18499	0.69850
LGB_Tune28	0.84923	0.18492	0.69846
LGB_Tune3	0.84956	0.18449	0.69911
LGB_Tune2	0.84909	0.18426	0.69819
LGB_Tune33	0.84851	0.18343	0.69702
LGB_Tune23	0.84866	0.18339	0.69732
LGB_Tune26	0.84862	0.18276	0.69725
LGB_Tune25	0.84778	0.18271	0.69556
LGB_Tune22	0.84774	0.18259	0.69548
LGB_Tune27	0.84726	0.18207	0.69453
LGB_Tune5	0.84733	0.18148	0.69467
LGB_Tune20	0.84778	0.18144	0.69556
LGB_Tune21	0.84819	0.18111	0.69638
LGB_Tune24	0.84786	0.18049	0.69572
LGB_Tune6	0.84447	0.17763	0.68894
LGB_Tune14	0.84529	0.17756	0.69057
CatBoost_Tune12	0.84359	0.17745	0.68718
CatBoost_Tune0	0.84425	0.17736	0.68850
LGB_Tune12	0.84364	0.17728	0.68728
CatBoost_Tune6	0.84302	0.17720	0.68604
CatBoost_Tune14	0.84400	0.17672	0.68800
LGB_Tune30	0.84382	0.17652	0.68764
CatBoost_Tune15	0.84173	0.17499	0.68347
LGB_Tune10	0.84192	0.17495	0.68384
CatBoost_Tune10	0.84176	0.17164	0.68352

Optimizing Features

Using the importance scores of the features from the optimized model, we experiment on the effect of recursive feature elimination.

Recursive Feature Elimination

- ➔ Slightly Better Performance at Top 95% Features (Removed 192 Vars)
- ➔ Continued Decrease in Performance below 90%
- ➔ Some features might have added noise and complexity to the model



%Cum Imp	50%	60%	70%	80%	90%	95%	99%	100%
n_features	73	99	129	166	218	258	312	450
ROC-AUC	0.8508	0.8534	0.8551	0.8556	0.8565	0.8561	0.8566	0.8561
PR-AUC	0.1830	0.1867	0.1887	0.1897	0.1907	0.1919	0.1903	0.1896

HOME CREDIT

Experimenting with PCA

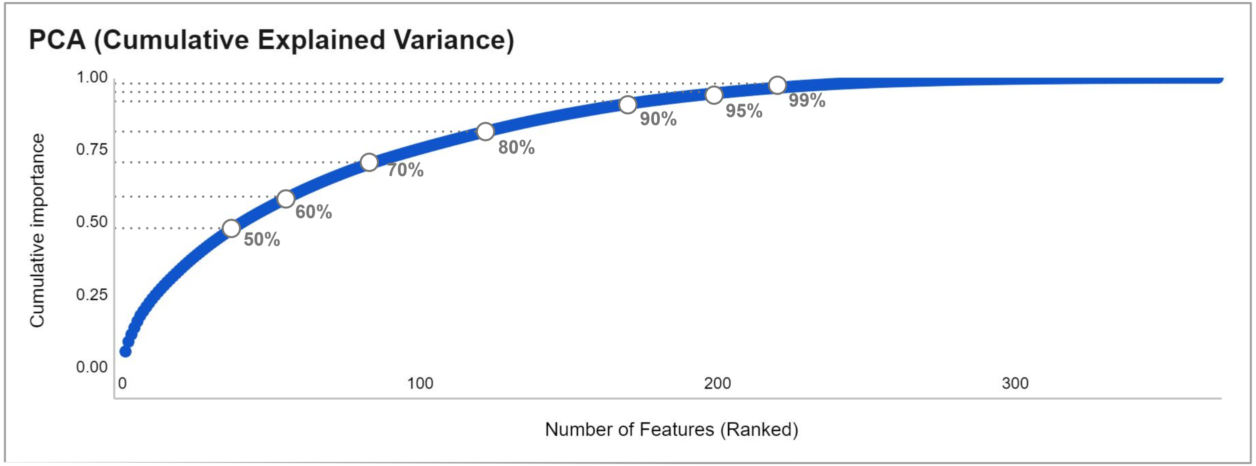
Using the PCA explained variance, we experiment on the effect of varying the n_components of the dataset.

PCA Dimensionality Reduction

➔ Overall Performance is worse than original dataset

➔ Reducing Principal Components continuously decreased performance

➔ Unlike RFE, PCA's ranking is not influenced by relationship with target

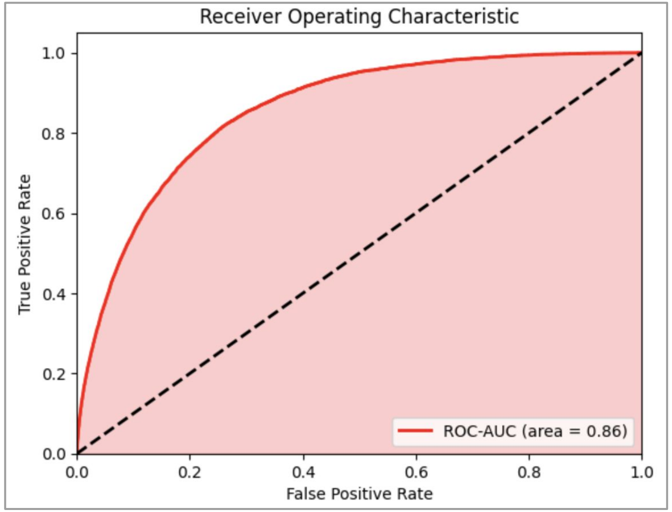


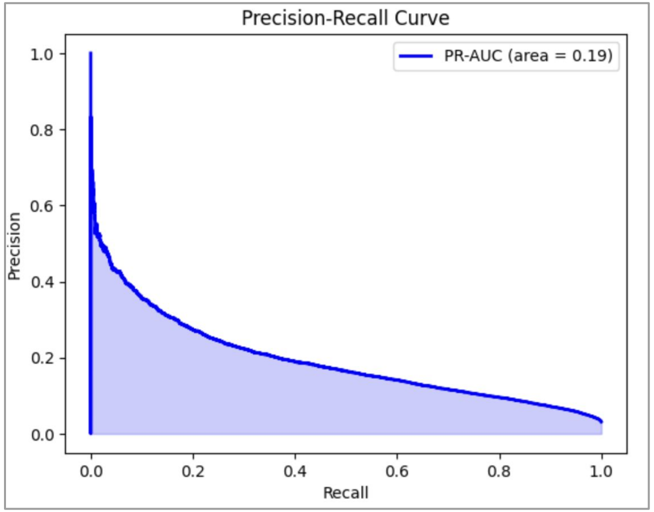
%Cum Imp	50%	60%	70%	80%	90%	95%	99%	100%
n_features	122	140	164	198	245	285	351	450
ROC-AUC	0.8198	0.8223	0.8232	0.8247	0.8259	0.8274	0.8289	0.8335
PR-AUC	0.1467	0.1496	0.1511	0.1522	0.1529	0.1534	0.1550	0.1580

Model Performance

The best performing model is selected from all iterations. Optimizing thresholds/cutoffs for prediction will also be determined.

LightGBM Tuned Model, 285 Features





	ROC-AUC	PR-AUC	GINI
Dumb Classifier	0.500	0.032	0.000
Random Chance Classifier	0.500	0.032	0.000
LightGBM Tuned Model	0.860	0.190	0.720

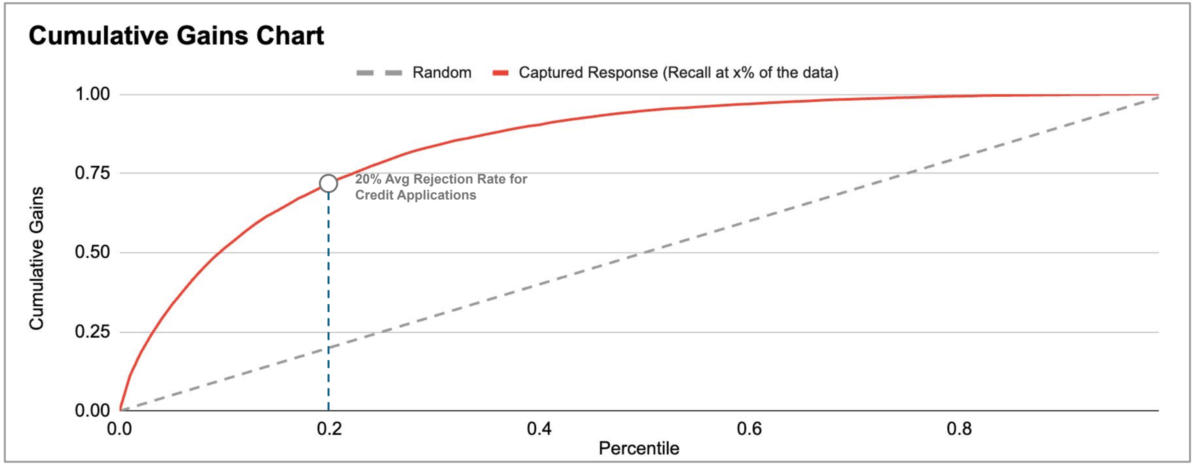
	GINI
Kaggle Leaderboard	0.67303
Best Submission	0.63763

Model Performance

The best performing model is selected from all iterations. Optimizing thresholds/cutoffs for prediction will also be determined.

Captured Response

- ➔ 72% of Defaulters are captured at 20% of the base.
- ➔ Trade-off of precision
- ➔ Maximize Returns based on Risk & Returns



Percentile Threshold	3%	5%	7%	10%	15%	20%	30%	50%
Accuracy	95.37%	93.97%	92.47%	90.09%	85.83%	81.37%	72.11%	52.80%
Precision	24.97%	21.00%	18.59%	16.07%	13.18%	11.24%	8.73%	5.93%
Recall	23.94%	33.56%	41.60%	51.38%	63.19%	71.85%	83.69%	94.79%
F1	24.45%	25.84%	25.70%	24.49%	21.81%	19.44%	15.81%	11.16%
Lift	7.98	6.71	5.94	5.14	4.21	3.59	2.79	1.90

Data Strategy

Business Impact & Use Case

WHAT INFORMATION ARE NECESSARY TO APPROVE LOANS?

Top 20 features are a good balance between credit data and non-credit data.

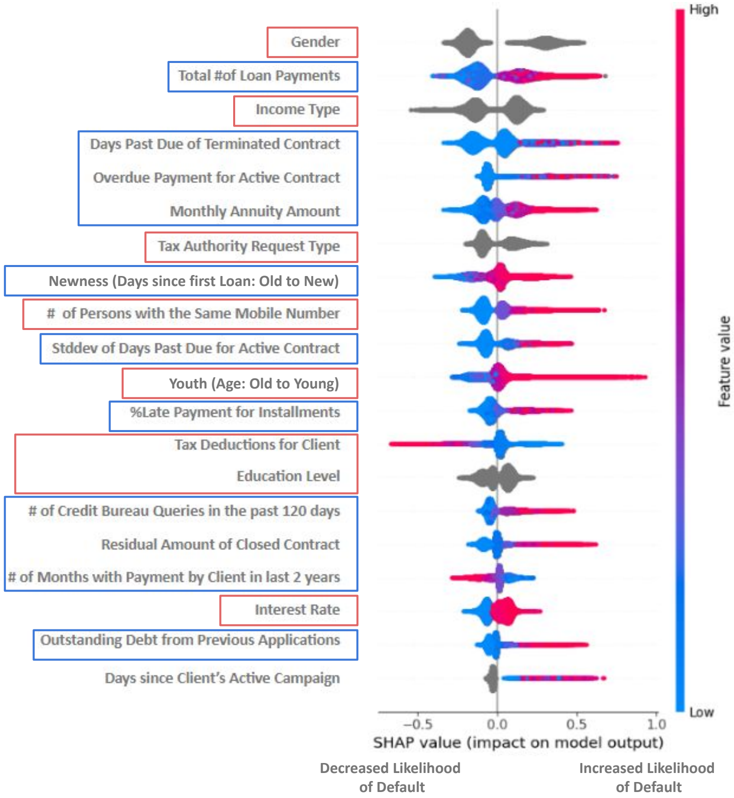
Little to None Credit History

- Income
- Income Type
- Age
- Education
- Tax Payments

With Credit History

- Promptness of Payments
- Consistency of Payments
- Delinquency History
- Annuity Amount
- Loan Tenure

...on top of requirements for little to no credit history



Defaulters Profile

Payment behavior of experienced loaners are determinant of default while income, age, and application frequency are determinant for new applicants.

WHAT TYPES OF APPLICANTS HAVE HIGH LIKELIHOOD TO DEFAULT ON LOANS?

Little to None Credit History

- Private Employees
- Duplicate/ Redundant Applications
- Younger Loaners
- Lower Tax Bracket

With Credit History

- Late Payers
- Delinquent Loaners
- Irregular Payers
- Larger Annuity Amounts
- Newer Loaners

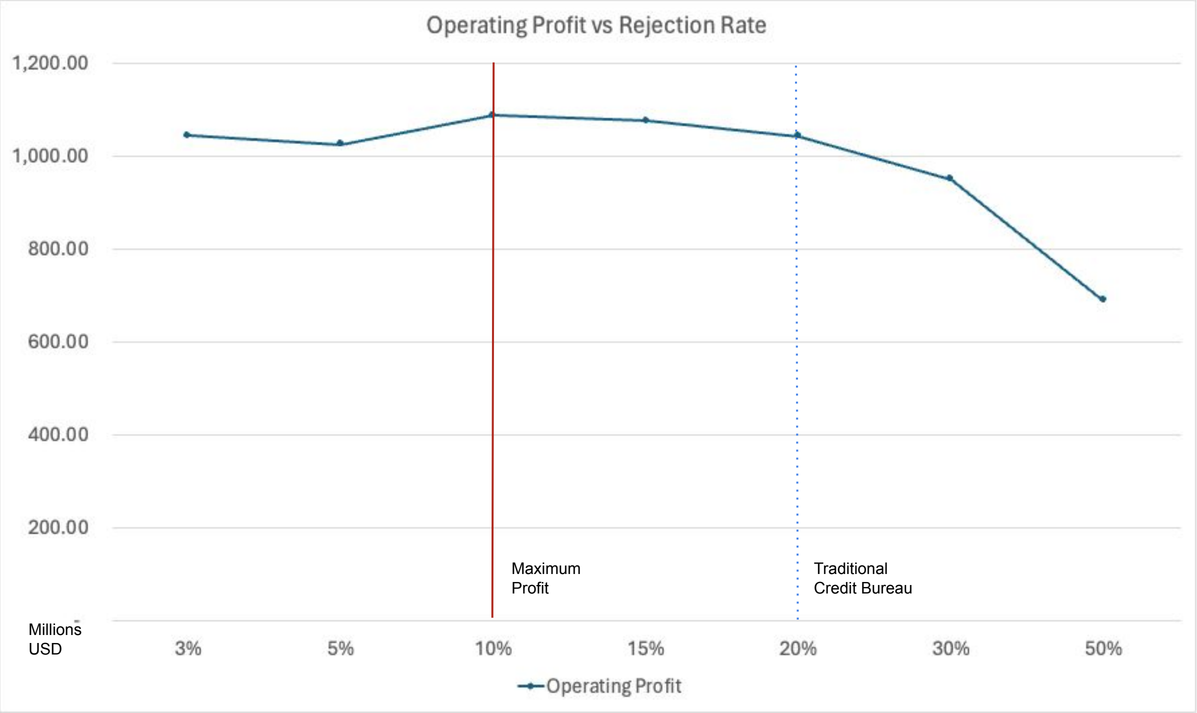
	Features	Defaulters	Good Loaners
☆	Gender	Male	Female
	Payment Frequency	16x	12x
☆	Income Type	Private Sector Employee	Other, Retired Pensioner
	Days Past Due of Terminated Contract	7.37	0.28
	Monthly Annuity Amount	3,288.40	3,142.20
	Tax Authority Request Type	DEDUCTION_6	PENSION_6
	Years since First Credit	1.5 Years	2 Years
☆	Persons with the same Mobile Number	2	1
☆	Std Dev of Days Past Due	0.45	0
☆	Age	38 yrs	43 yrs
	%Late Payments on Installation	29%	8%
☆	Taxes	5,800.00	8,520.60
	Education Level	MASKED CATEGORIES	
☆	# of Credit Bureau Queries	2	1
	Residual Amount of Closed Contract	23,339.93	10,449.90
	Months with Payment by Client	8	10
	Interst Rate	0.30	0.28
☆	Oustanding Debt from Previous Application	11,695.25	7,837.31

☆ Interesting insights

Impact & Strategy

Simulated Gains/Loss for the Test Data using the Predictions. Find Optimal Rejection Criteria based on Risk Appetite.

- Traditional credit companies has 20% rejection rate.
- Profit can be increased by **43M** (~0.4%) by relaxing rejection rate to 10%.



Recommendations

1

Model Enhancements

- Experiment with Handling Imbalanced Datasets
- Separate Models for those with Credit History and those without
- Explore other Algorithms

2

Consider Market Nuances

- Demographics might be skewed for each location
- Different Markets (i.e. jurisdictions) have different financial legal frameworks.
- Identify location with good credit scores for expansion

3

Dynamic Product Tiers

- Adjust Credit Limit, Fees, and Max Loan Terms according to default risk and other significant features

HOME CREDIT

WE ARE LT4.

THANK YOU!

Impact & Strategy

Simulated Gains/Loss for the Test Data using the Predictions. Find Optimal Rejection Criteria based on Risk Appetite.

Rejection Criteria	Actual Default Rate	Approved Loans	Performing Loans	Interest Earned (Assume 10% EIR)	Non-Performing Loans (Assume Uncollectible)	Gains/Loss	
Reject Top 3% High Probability to Default	2.40%	14,640,507,000	14,259,040,000	1,425,904,000	381,467,000	1,044,437,000	
Reject Top 5% High Probability to Default	2.20%	14,312,962,100	13,980,840,000	1,398,084,000	332,122,100	1,042,178,600	
Reject Top 10% High Probability to Default	1.70%	13,516,131,300	13,275,980,000	1,327,598,000	240,151,300	1,087,446,700	Highest Profit
Reject Top 15% High Probability to Default	1.40%	12,718,963,300	12,539,970,000	1,253,997,000	178,993,300	1,075,003,700	
Reject Top 20% High Probability to Default	1.10%	11,928,779,400	11,791,780,000	1,253,997,000	136,999,400	1,042,178,600	Avg Credit Card Rejection Rate
Reject Top 30% High Probability to Default	0.71%	10,340,142,990	10,263,300,000	1,026,330,000	76,842,990	949,487,010	
Reject Top 50% High Probability to Default	0.33%	7,161,089,380	7,136,630,000	713,663,000	24,459,380	689,203,620	
Rejection Criteria using Model Scores	%Default Rate of Approved Applicants	Sum of All Loan Amounts/ Approved by chosen criteria	Sum of all Approved Loan Amounts that did not default	Interest Earned (assumed 10% EIR) from Performing Loans	Sum of all Approved Loan Amounts that defaulted (considered as loss if uncollectible)	Net of Interest Earned - Non-Performing Loans	

HOME CREDIT