Loan Risk Category Prediction

Athiya Anindya
ID/X Partners Data Scientist Project Based Internship Program

Agenda

Project Recap
Client brief and problem statement

Problem
The Challenge of Manual Credit Checks & The Opportunity of Automated Credit Checks

Process
Workflow & explanations for each process

Summary
Conclusion & insight from data

Recommendation
Model recommendation, business recommendation and business metrics simulation

Project Recap

Amara is a fast-growing lending company, offering a more convenient, digital-first experience. The number of Amara users is expected to reach 7 million by 2025. With an increasing customer base, Amara's manual background check processes were difficult to scale.

Using automated credit checks to make B2C credit decisions

They requested a model that can predict credit risk category using a dataset from the company which comprises historical list of credits that are accepted and rejected.

The Challenge of Manual Credit Checks

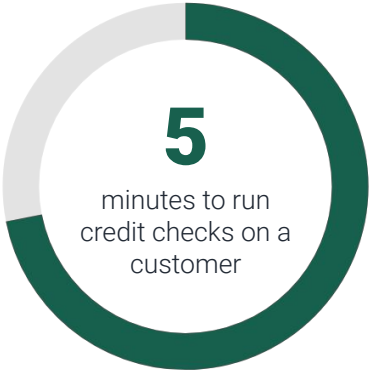
Manual processes tend to be inefficient

In addition to the time associated with manually conducting credit assessment, reliance on manual systems also increases the potential for errors due to the fatigue of human operators.

The Opportunity of Automated Credit Checks

Lessen the cost of overseeing customer credit

Process can be completed in minutes, freeing up valuable time and resources that can be used for final expert judgement and other business activities.



Machine Learning Workflow

Exploring data

Drop features based on knowledge (personal judgment).

Defining label

There is already loan_status feature. We define "bad" & "good" borrowers by this feature.

Data pre-processing

Fill in missing values. Modify string/object data types. Change time to duration. Change data types.

Check correlation

Drop features that have multicollinearity.

Modeling

Algorithms: Logistic Regression, Decision Tree, Random Forest. Evaluation metrics: AUC & KS

Handling imbalance

SMOTE

Split train-test

Train: 80%
Test: 20%

Feature encoding & scaling

Feature encoding on purpose, grade, home_ownership, verification_status and list_status. Scaling all numerical features.

Exploring data

Drop features

More than 60% missing values

mths_since_last_record, mths_since_last_major_derog, annual_inc_joint, dti_joint, verification_status_joint, open_acc_6m, open_il_6m, open_il_12m, open_il_24m, mths_since_rcnt_il, total_bal_il, il_util, open_rv_12m, open_rv_24m, max_bal_bc, all_util, inq_fi, total_cu_tl, inq_last_12m, mths_since_last_delinq

Identifiers

Unnamed: 0, id, member id, url, title, desc, zip_code, emp_title, addr_state, and policy_code

Similar features

sub_grad is similar to grade column. dti_joint is similar to dti. out_prncp_inv is similar to out_prncp. total_pymnt_inv is similar to total_pymnt. funded_amnt_inv is similar to funded_amnt

Only 1 unique value

application_type

Columns that are not relevant to predicting bad borrowers

collection_recovery_fee, pymnt_plan (no description)

Defining label

Bad

Charged Off

Late (31-120 days)

Default

Does not meet the credit policy. Status:Charged Off

Good

Current

Fully Paid

In Grace Period

Does not meet the credit policy. Status:Fully Paid

Late (16-30 days)

Pre-processing

emp_length

Change to numerical data type

term

Change to numerical data type

Handle missing values

impute emp_length with 0
impute 'earliest_cr_line' with mode
impute 'last_pymnt_d' with mode
impute 'last_credit_pull_d' with mode
impute 'annual_inc' with median
impute 'delinq_2yrs' with 0
impute 'inq_last_6mths' with 0
impute 'open_acc' with 0
impute 'pub_rec' with 0
impute 'revol_util' with 0
impute 'total_acc' with 0
impute 'collections_12_mths_ex_med' with 0
impute 'acc_now_delinq' with 0
impute 'tot_coll_amt' with 0
impute 'tot_cur_bal' with 0
impute 'total_rev_hi_lim' with 0

Convert date to duration

Issue_d, earliest_cr_line, last_pymnt_d, last_credit_pull_d

Check correlation

Drop features with correlation > 0.7

'installment', 'total_rec_prncp', 'last_pymnt_d', 'revol_bal'

Feature encoding

Grade

Before: A, B, C, D, E, F, G
After: 6, 5, 4, 3, 2, 1, 0

Home_ownership

Before: 'RENT', 'OWN', 'MORTGAGE', 'OTHER', 'NONE', 'ANY'
Method: One-hot encoding

Verification_status

Before: 'Not Verified', 'Verified', 'Source Verified'
After: 0, 1, 1

Purpose

Before: 'credit_card', 'car', 'small_business', 'other', 'wedding', 'debt_consolidation', 'home_improvement', 'major_purchase', 'medical', 'moving', 'vacation', 'house', 'renewable_energy', 'educational'
Method: One-hot encoding

Initial_List_status

Before: 'f', 'w'
After: 0, 1

Feature scaling

Standardization

All numerical values

Split train & test data

Train: 80%
Test: 20%

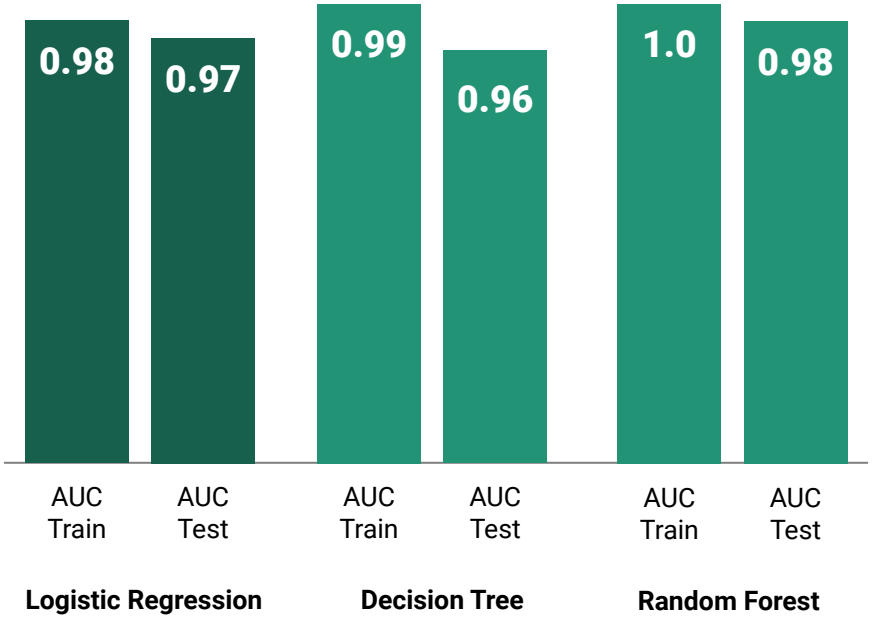
Imbalance resampling

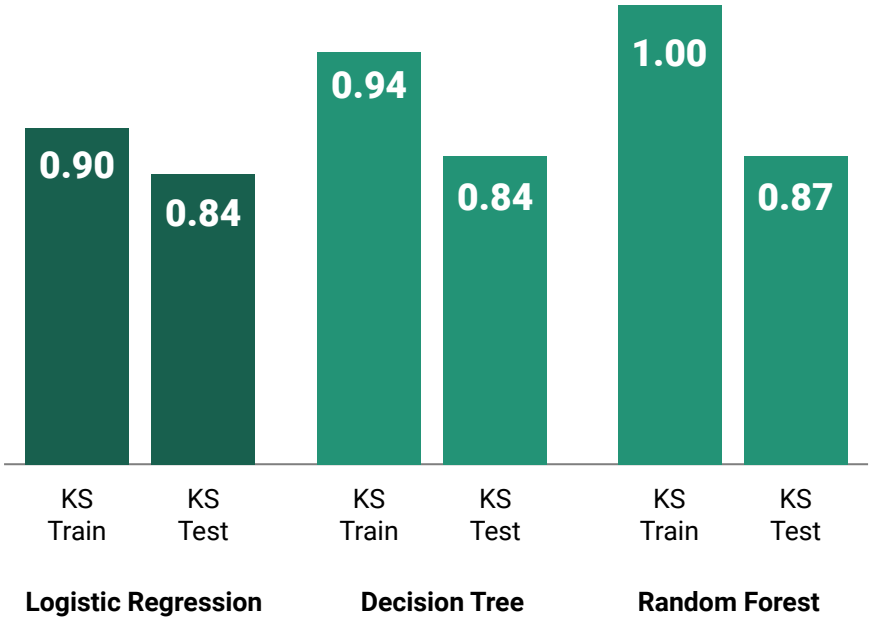
Method: SMOTE

Before:
bad
0 332250
1 40778

After:
bad
0 332250
1 332250

Modeling


0.98
0.97
0.99
0.96
1.0
0.98
AUC Train
AUC Test
AUC Train
AUC Test
AUC Train
AUC Test
Logistic Regression
Decision Tree
Random Forest


0.90
0.84
0.94
0.84
1.00
0.87
KS Train
KS Test
KS Train
KS Test
KS Train
KS Test
Logistic Regression
Decision Tree
Random Forest

Feature Importance

Logistic Regression

Summary

1

The best fit model uses Logistic Regression algorithm and results in AUC = 0.97 and KS = 0.84.

2

Feature importance

Total_pymnt: Payments received to date for total amount funded
Funded_amnt: The total amount committed to that loan at that point in time.
Out_prncp: Remaining outstanding principal for total amount funded

Model recommendation

1

For higher interpretability, consider creating a model with Feature Selection using Information Value and Feature Engineering using Weight of Evidence.

2

Perform hyperparameter tuning.

Business Recommendation

Customer segmentation

Create customer segments to adjust marketing strategy.

Adjusted loan conditions

Amara can determine loan conditions based on customer's most important features.

Business Simulation

Credit application

Customers have to fill in manual multi-page application form.

Data verification & reference check

3 business days

Manual check

1 business day

Document submission

Credit report, bills, income statement, ec.

Credit worthiness analysis

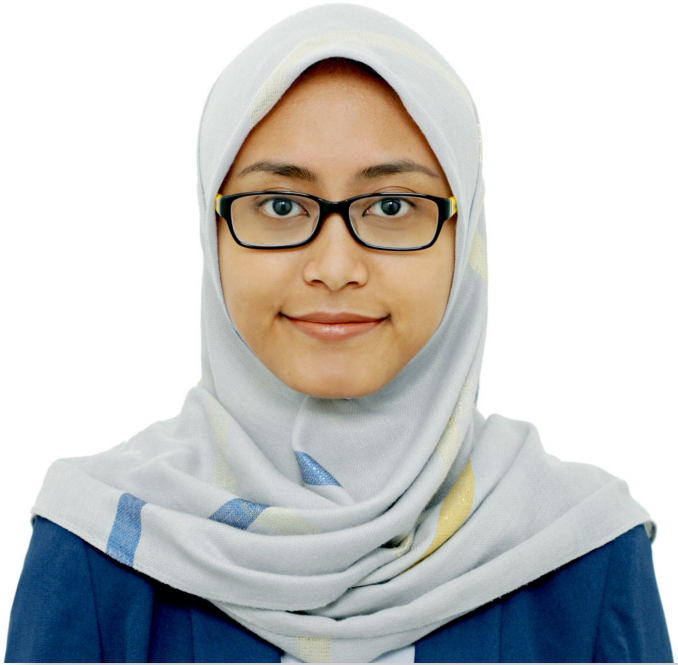
3 business days

5 minutes

Reduced time!

Credit decision result

Determine decision, net term and credit limit



Intern Profile

Athiya Anindya

Email address: athiyanindya@gmail.com

Project link: https://github.com/learnindya/Loan-Credit-Risk

Linkedin: https://www.linkedin.com/in/athiyanindya/