Capstone Project

Loan Default Prediction

Machine Learning Model Recommendation

Maggie Hao

October 2022

CONTENT


Problem Definition
Proposed Solution
Insights & Findings
Recommendations

Problem to Solve

High Default Rate (~20%)

- None Performing Assets leads to big default losses
- Consumer lending profitability largely compromised

Rejecting Good Customers

- Losing business as a result of being too conservative
- Lost market share, revenues

Non-Optimal Manual Loan Review Process

Long Review Time

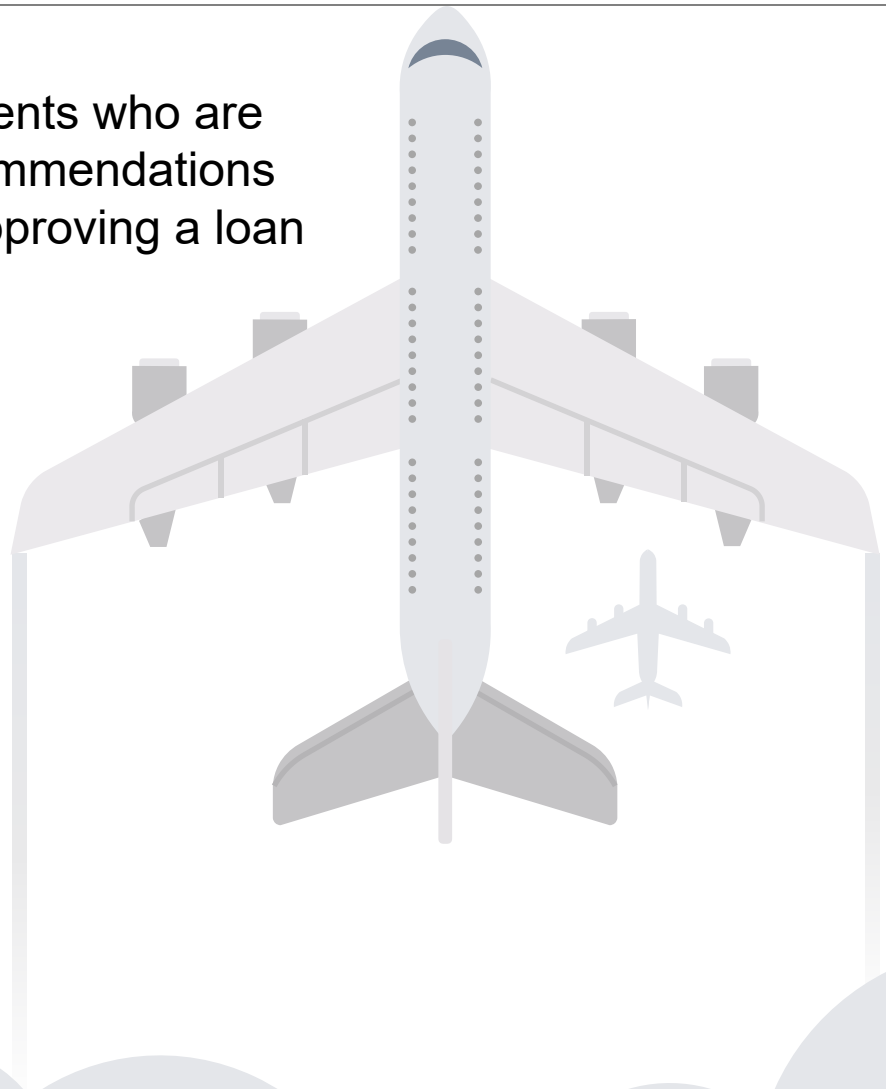
- Loss of clients due to long lead times
- Lower customer satisfaction

Inconsistent Standards

- Variation across different loan officers
- Potential human biases
- No department-wise communication guidelines

Goal: Automate & Improve Loan Review Process with a ML Model

Build an explainable classification ML model to predict clients who are likely to default based on their features, and provide recommendations to the bank on the important features to consider while approving a loan



Expected Key Outcomes

- Unbiased, generalize well on unseen data
- Good prediction results
- Interpretable => bank can act on it
- Supplemental to existing manual process
- ↑ revenues
- ↓ default losses
- Efficient
  - ↓ loan approval lead time
  - ↓ FTEs

Data Overview (EDA)

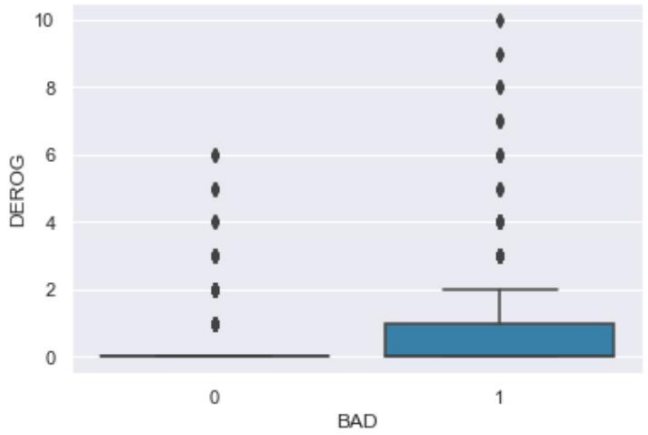
5960 observations, 1 target, 12 feature variables

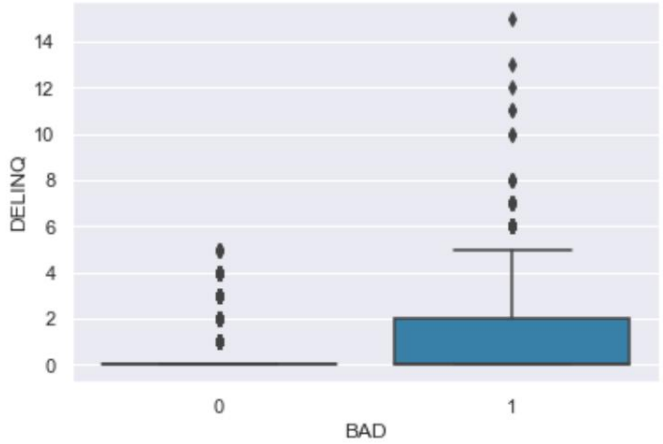
	Definition	y/X	dtype	missing%[1]	skewness, outliers[2]	Significant impact on BAD
BAD	1 = Client defaulted on loan	y	categorical	0.0%	N.A.	
LOAN	Amount of loan approved	X	numeric	0.0%	right, outliers	
MORTDUE	Amount due on the existing mortgage	X	numeric	8.7%	right, outliers	
VALUE	Current value of the property	X	numeric	1.9%	right, outliers	
REASON	Reason for the loan request	X	categorical	4.2%	N.A.	
JOB	Job type of applicant	X	categorical	4.7%	N.A.	√
YOJ	Years at present job	X	numeric	8.6%	right, outliers	
DEROG	# of major derogatory reports	X	numeric	11.9%	right, outliers	√
DELINQ	# of delinquent credit lines	X	numeric	9.7%	right, outliers	√
CLAGE	Age of the oldest credit line	X	numeric	5.2%	right, outliers	
NINQ	# of recent credit inquiries	X	numeric	8.6%	right, outliers	
CLNO	# of existing credit lines	X	numeric	3.7%	right, outliers	
DEBTINC	Debt-to-income ratio	X	numeric	21.3%	right, outliers	√

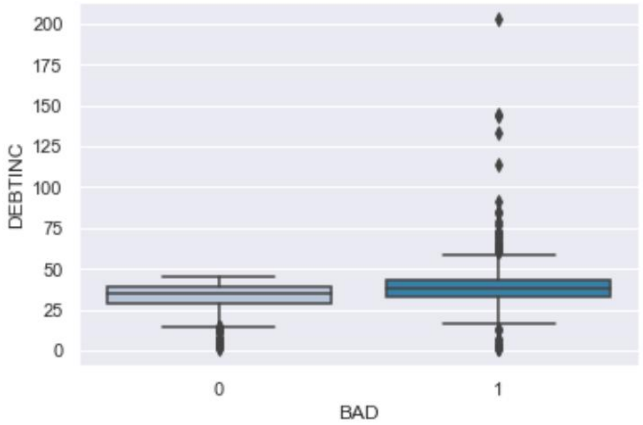
Note: 1. Missing value imputation: mode used for categorical variables, median used for numerical variables

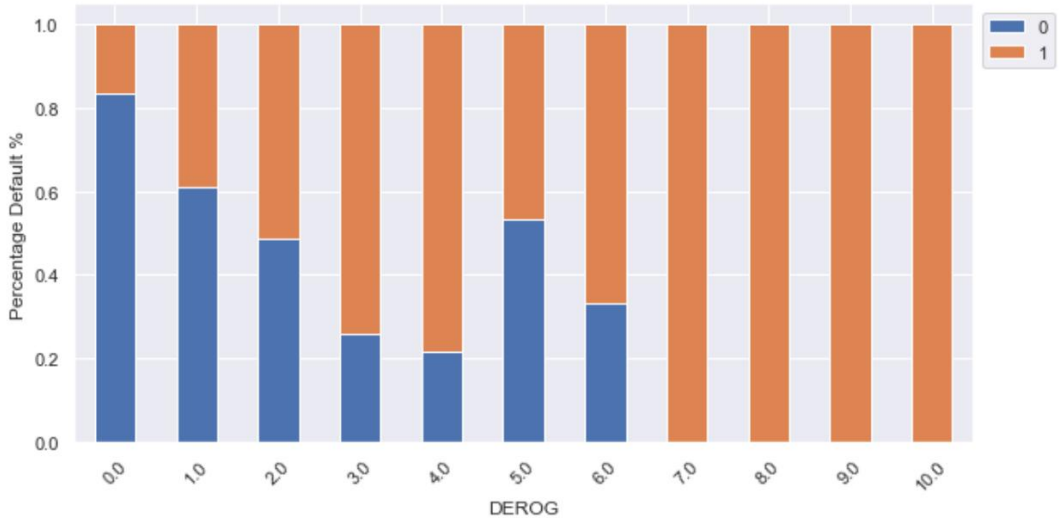
2. Outlier treatment is done for the dataset used in non-tree based algorithms

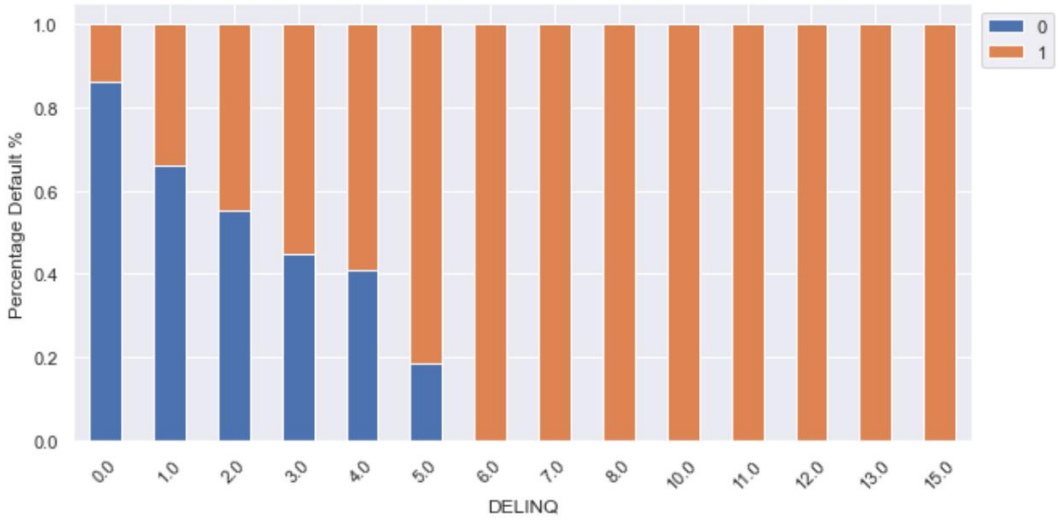
Important Factors from EDA











Solution Summary

Proposed Model: ***Tuned Random Forest model***

RandomForestClassifier(class_weight='balanced', criterion='entropy', max_depth=7, max_features=0.9, min_samples_leaf=20, n_estimators=500)

Reasons for recommendation:

- Best overall performance among 7 models
- Good and balanced prediction performance: accuracy > 85%, recall ~75%, precision ~65%.
- Not overfitting
- Interpretable at the aggregate level - feature importance

Important Factors:

- not providing any income-to-debt ratio
- income-to-debt ratio
- # of delinquency credit lines
- length of credit history
- # of major derogatory reports

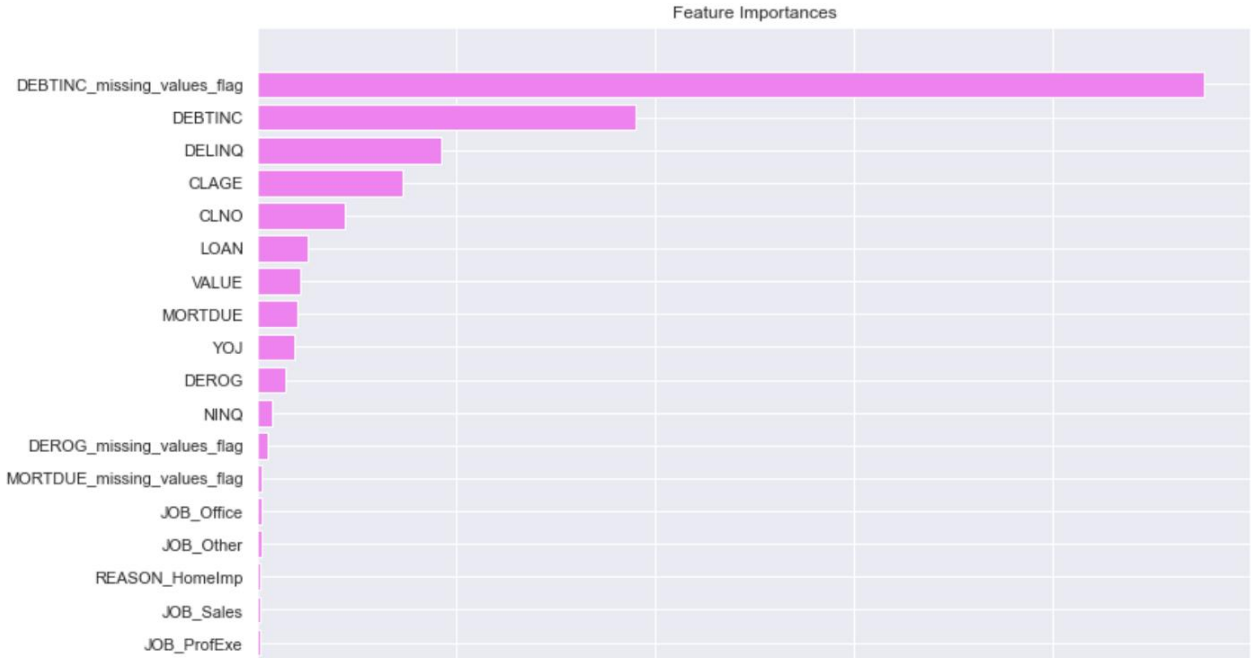
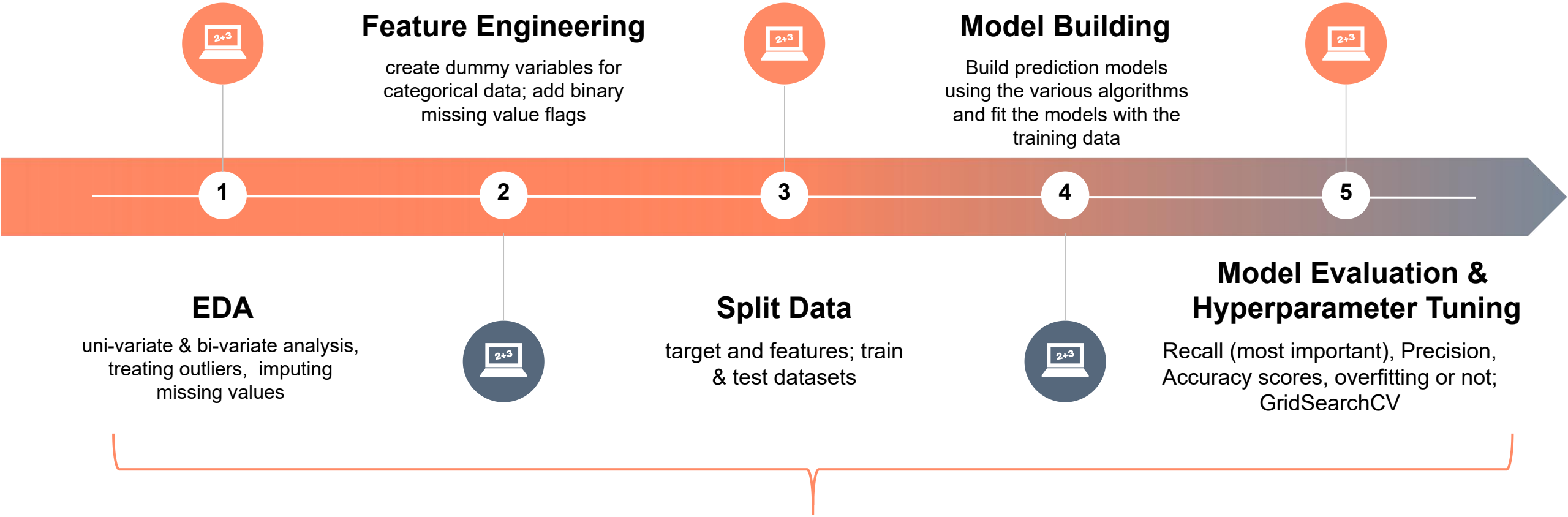


Figure: Tuned Random Forest Feature Importance

Solution Design



Final Step: compare all models and pick the best overall model

Analysis - model performance comparison

	Model	Train_Accuracy	Test_Accuracy	Train_Recall	Test_Recall	Train_Precision	Test_Precision
0	Logistic Regression	0.889022	0.892617	0.615385	0.607843	0.781679	0.806691
1	LDA	0.885427	0.885347	0.669471	0.649860	0.732895	0.743590
2	QDA	0.853068	0.850112	0.561298	0.551821	0.653147	0.645902
3	Decision Tree	1.000000	0.880313	1.000000	0.655462	1.000000	0.720000
4	Tuned Decision Tree	0.823826	0.820470	0.843750	0.792717	0.537108	0.533962
5	Random Forest	1.000000	0.911633	1.000000	0.703081	1.000000	0.828383
6	Random Forest with Class Weight Specified	0.999760	0.912192	0.998798	0.691877	1.000000	0.840136
7	Tuned Random Forest	0.883030	0.865772	0.818510	0.742297	0.668959	0.641646

Overall, the best performing model that is recommended for deployment is the ***Tuned Random Forest*** model, balancing all evaluation factors: accuracy > 85%, recall ~75%, precision ~65%.

- Recall close to 75%: the model can identify ~75% of all defaulters and reject them.
- Precision close to 65%: bank would lose less legitimate customers compared to the Tuned Decision Tree model.
- Not overfitting
- Explainable to some extend with features importance chart

Insights & Findings

Improved Decision Quality

- ***Reduced False Negative*** leads to less default loss: the model will identify 75% (i.e., class 1 recall) of all the true defaulters and reject their applications. This should have been an significant increase from previous pure human approval results.

- ***Reduced False Positive*** leads to less lost revenues: among all the true good customers who would not default, the model will be able to approve over 90% (i.e., class 0 recall) of them. Equivalently, the model has a class 1 precision of 65%, meaning the model is only making 35% mistakes in all rejected customers.

Interpretability

- Key factors: DEBTINC_missing_values_flag, DEBTINC, DELINQ, CLAGE, DEROG.

Benefits:

- Assist internal communication among loan approval teams, management, and sales
- Loan sales team can use it to explain the banks' decisions to a certain extent to their potential customers, or to advise their potential customers on actions that can increase their chances of obtaining the loans in the future. ↑ customer satisfaction and future revenues.
- Marketing and sales team can use it to improve marketing / sales initiatives targeting the right segments of customers. ↑ the conversion rate and interest revenues.

Limitations & Risks

- Model performance limitations:
  - Currently not 100% accurate, innate uncertainty;
  - Possible latent variables;
  - Static model;
  - Not fully interpretable at individual borrower level

- Implementation challenges:
  - Training and model maintenance costs
  - Potential resistance if the implementation is not handled well

- Potential reputational / regulation risks

Recommendations

Deploy the Tuned RF Model

As a supplemental tool and to integrate the model prediction in the loan review process; update credit review guidance & SOP

Bank should not approve application if customer checks multiple boxes:

- If no debt-to-income ratio provided
- If debt-to-income ratio > 45%-50%
- If delinquent credit lines >= 3
- If major derogatory reports >= 2

Conduct legal & compliance review

Review the proposed policies and new workflow in light of the legal and regulation requirements to ensure it is proper and compliant

Internal training

Conduct presentations and training sessions on the model to broader stakeholders, including the consumer credit loan review teams, compliance, sales, etc.

Draft communication guidance

create a general communication internal and external guidance on the decision rules and customer engagement

Model maintenance and improvement

- Monitor the performance overtime and enhance the model with feedback. e.g., post-transaction evaluation to study the error cases
- Explore other variables (e.g., total income, co-applicant, education, property area)
- Increase the training data size
- Further hyperparameter tuning on the existing models, e.g., Optuna
- Try other algorithms such as KNN and neural network (SHAP framework)
- Try Boosting algorithms (e.g., AdaBoosting, XGBoosting, GradientBoosting)

Refine marketing and sales initiatives

Use the insights on desirable customer profiles to improve marketing and sales programs and increase conversion rates

Thanks.

Maggie Hao

Oct 8, 2022