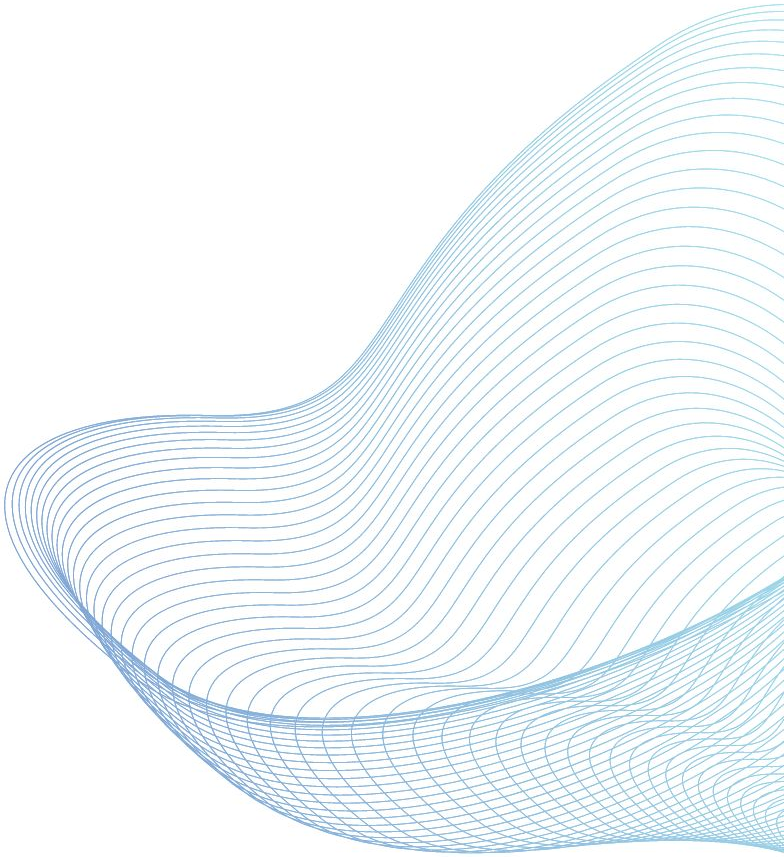
Machine learning/
Loan Default Prediction



By Supasit Vitooraporn

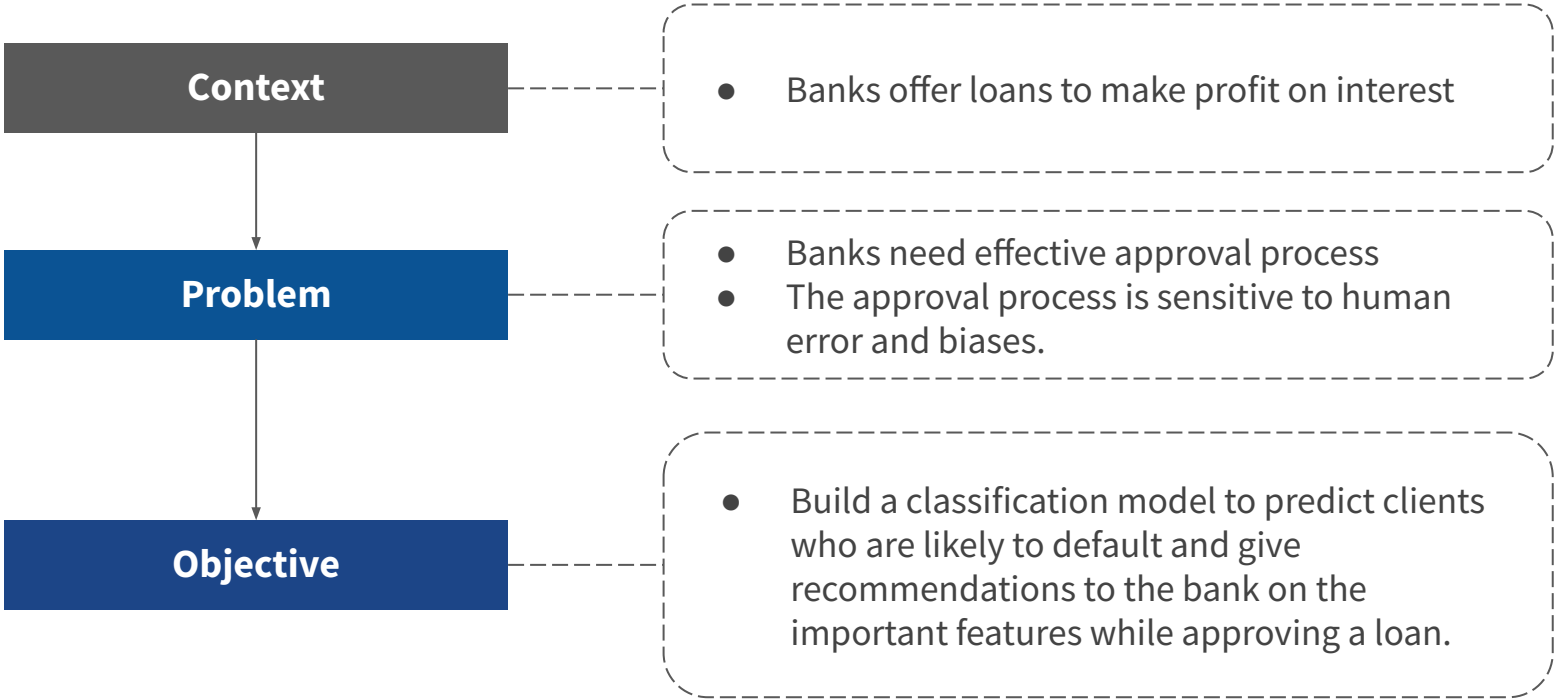
Loan Default Prediction Dataset:

Home Equity loan (HMEQ)

- The Home Equity dataset (HMEQ) contains baseline and loan performance information for **5,960 recent home equity loans**
- **The target (BAD) is a binary variable** that indicates whether an applicant has ultimately defaulted or has been severely delinquent.
- This adverse outcome occurred in **1,189 cases (20 percent).**
- **12 input variables** were registered for each applicant.

Problem Definition:

Overview



Important Metrics:

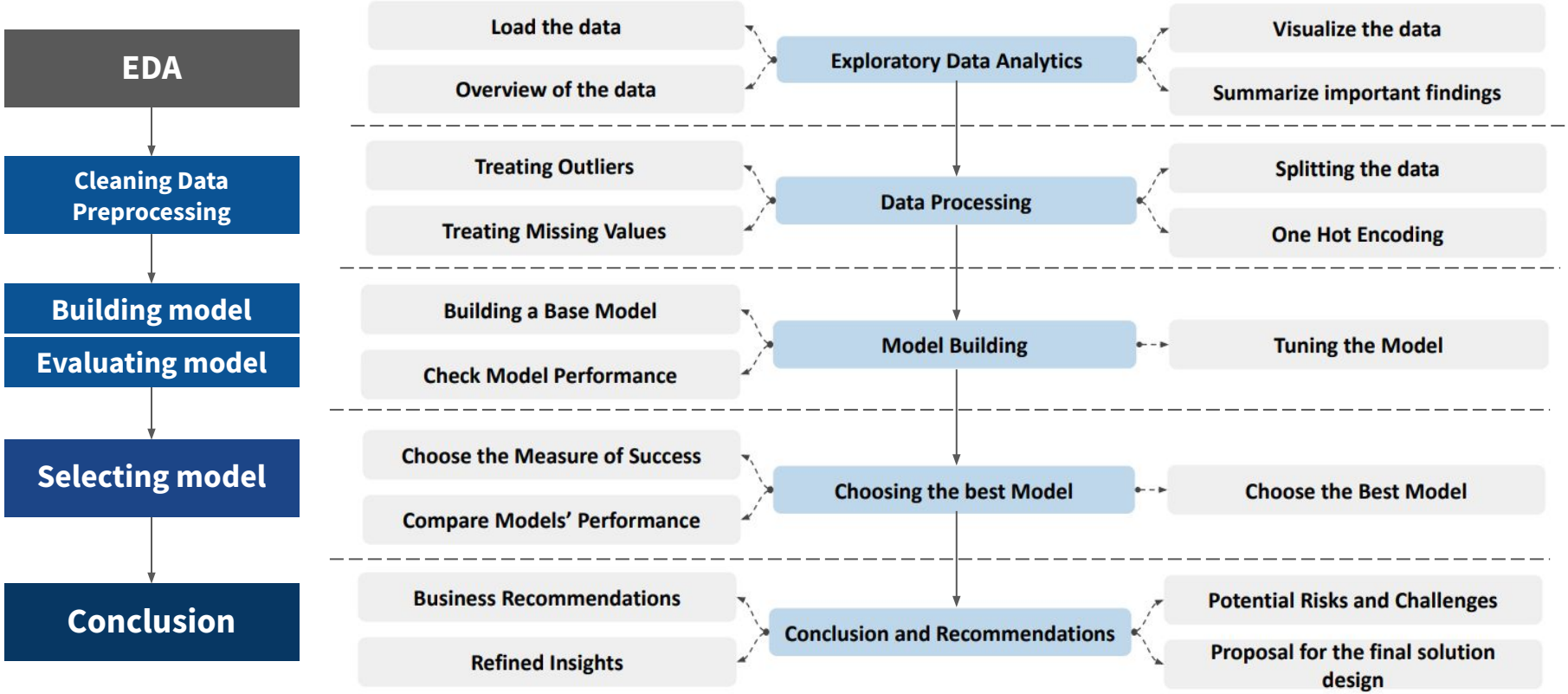
Home Equity loan (HMEQ)

- Best performance metric that best fits the business objectives: **Recall and Accuracy**
  a. **Recall**
     i. measures the ability of a model to correctly identify the actual positive cases. A high recall means that the model can effectively identify borrowers who are actually at risk of defaulting on their loans.
  b. **Accuracy**
     i. measures the overall correctness of a model's predictions. High accuracy means that the model is making correct predictions for both positive and negative cases.

- The bank wants to **maximize the recall.** The greater the recall score, the **higher chance to minimize the False Negative** case that results in defaulted loans.

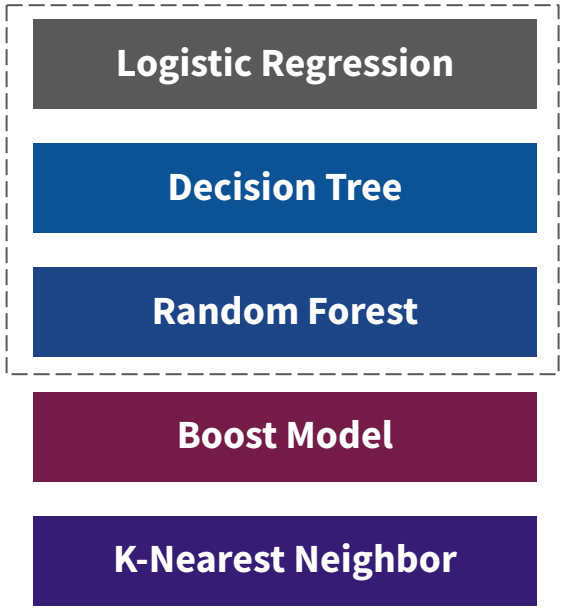
Solution Approach:

Steps to find the optimal outcome



Proposed Model Algorithms:

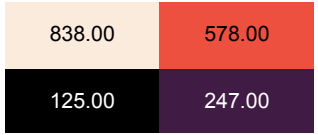
Model Solutions



Model performances:

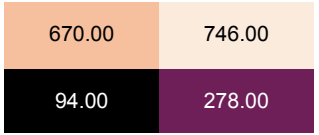
Logistic Regression

Baseline model



Accuracy	Precision	Recall	F1-score
0.61	0.58	0.63	0.56

Tuned Model



Accuracy	Precision	Recall	F1-score
0.53	0.57	0.61	0.51

Decision Tree

(Overfitting)

Accuracy	Precision	Recall	F1-score
0.85	0.78	0.76	0.77

Accuracy	Precision	Recall	F1-score
0.87	0.80	0.82	0.80

Random Forest

(Overfitting)

Accuracy	Precision	Recall	F1-score
0.89	0.86	0.80	0.82

Accuracy	Precision	Recall	F1-score
0.89	**0.83**	**0.84**	**0.83**

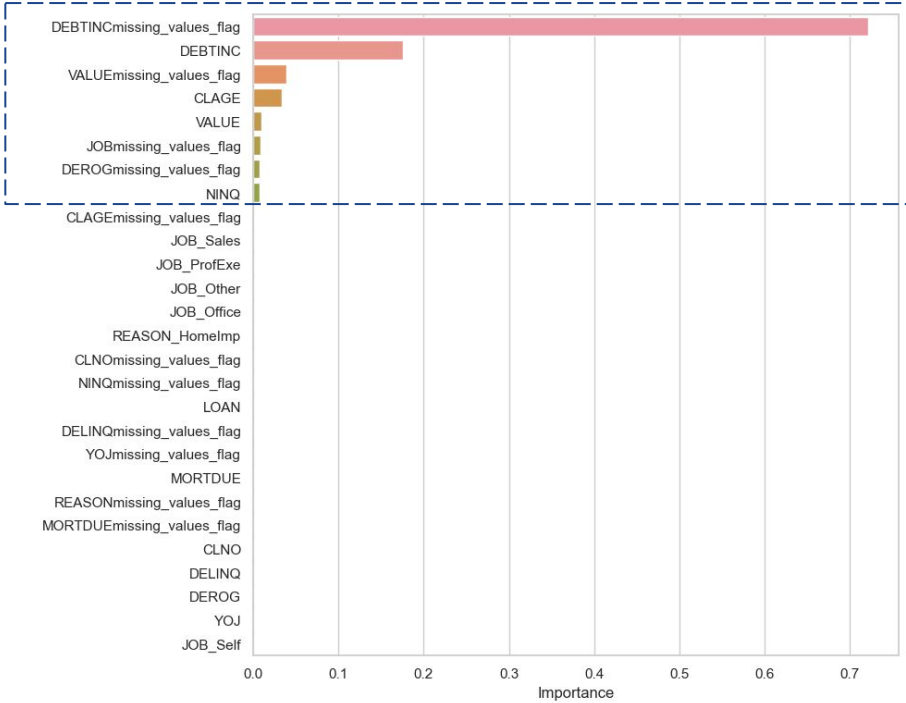
Comparison of techniques:

	Precision	Recall	Accuracy	Pros	Cons
Logistic regression model	0.584796	0.627893	0.606823	-	Extremely low recall
Logistic regression model using 0.6 threshold	0.584796	0.627893	0.606823		
Logistic regression model using 0.45 threshold	0.584796	0.627893	0.606823		
Decision Tree	0.779942	0.757495	0.852908	**Interpretable**	**Low recall**
Tuned Decision Tree	0.795514	0.815872	0.86689	**Interpretable**	**Lower recall**
Random Forest	0.860074	0.797783	0.893177	High accuracy	Lower recall
Balanced Random Forest	0.856855	0.782714	0.888143	High accuracy	Lower recall
Tuned Random Forest	**0.829884**	**0.838915**	**0.889262**	**Highest recall***	**Non-interpretable**
XGBoost	0.865233	0.797852	0.894855	Highest accuracy	Lower recall
KNN	0.786685	0.572592	0.814318	-	Lowest recall
Tuned KNN	0.681039	0.605317	0.798658	-	Extremely low recall

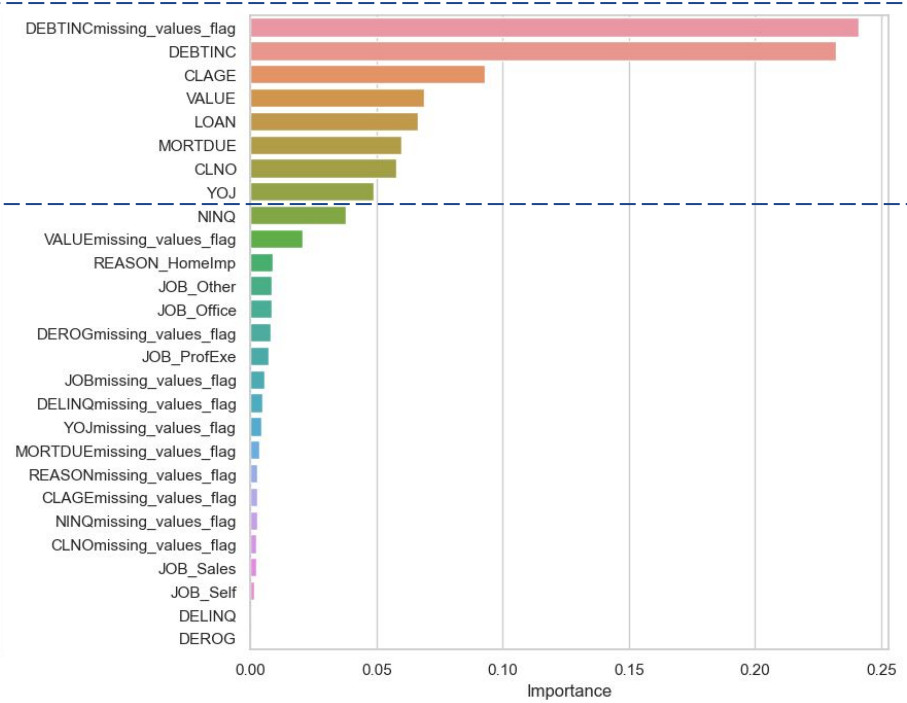
Importance Features:

Model Solutions

Tuned Decision Tree

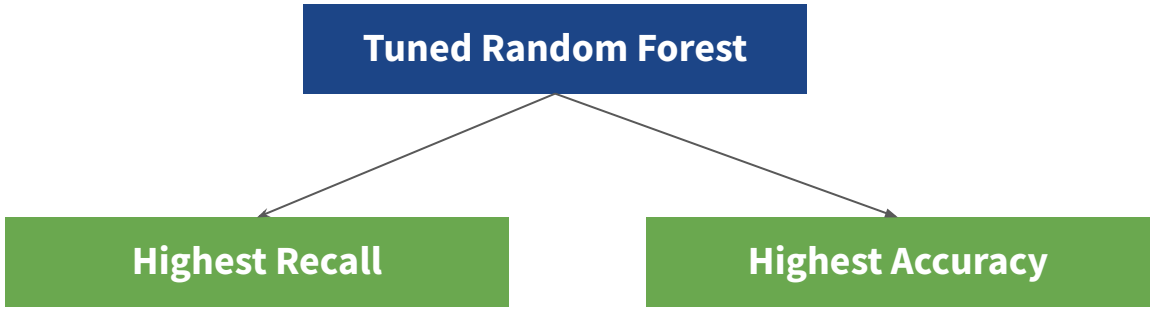


Tuned Random Forest



Model selection:

Model Solutions



- Best model: **Tuned Random Forest** (Highest Recall, Second highest Accuracy).
  a. Accuracy: 0.89
  b. Recall: 0.84
  c. Precision: 0.83

- Both Decision tree and random forest shared the same most importance features.

Executive Summary:

Home Equity loan (HMEQ)

- A decision tree model can predict loan defaulters **74%** of the time they come to ask for a home loan.

- **Debt-to-income ratio** is the most important driver for defaulting or repaying a loan. The bank should consider this as a top feature to consider while processing a loan, but also the one with the most missing data (21.3%) which is similar to the proportion of defaulted customers (20%)

- The bank should consider **CLAGE, VALUE, LOAN, MORTDUE, and CLNO** as a set of features that have a high impact on defaulting on a loan.

- It is recommendable to explore the possibility to create an alternative business process to manage and take decisions on those clients with no Debt/Income ratio available.

Recommendations:

Home Equity loan (HMEQ)

- Best model: **Tuned Random Forest** (Highest Recall)
- Best performance metric that best fits the business objectives: **Recall and Accuracy**
- Check the possibility to create an alternative business process to manage and take decisions on those clients with no Debt/Income ratio available.
- Explore other machine learning techniques such as engineering features, dropping columns, support vector machine, neural networks, among others.
- Create a pilot test with new model and compare the results with the current manual process before completing the transition to the new model.
- Check if there is a way to complete the missing values in the dataset.

Thank you