PREDICTION
with
MACHINE LEARNING

Problem

Interests from home loans

→ profits

Loan defaults

→ losses

Current decision-making process: manual

- Drawbacks

  - complex

  - effort-intensive

  - prone to human biases and errors

Target

Improve system with predictive modeling

→ simplified

→ faster

→ more efficient

→ free from human biases and errors

Elements and criteria

- client database

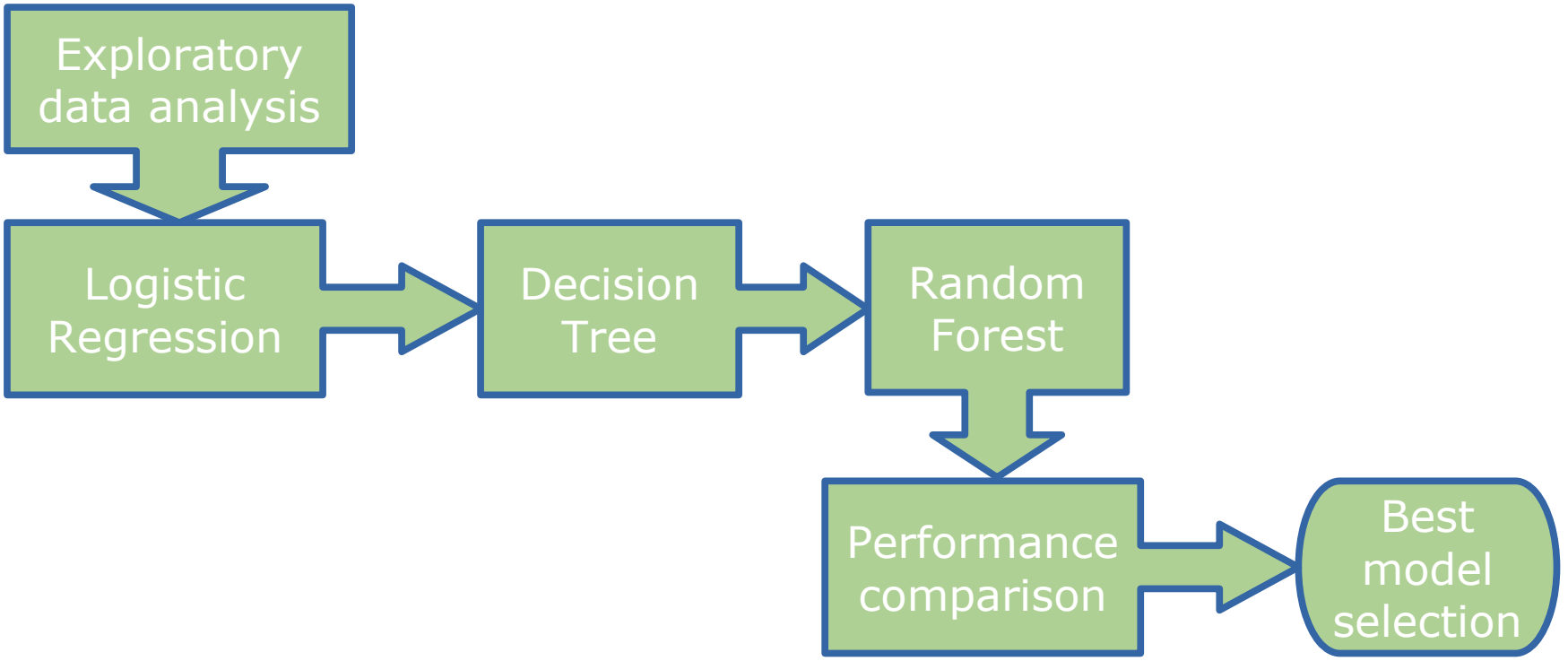
- loan defaults

- predict likelihood of loan default

- importance of features

- interpretable

Solution Approach

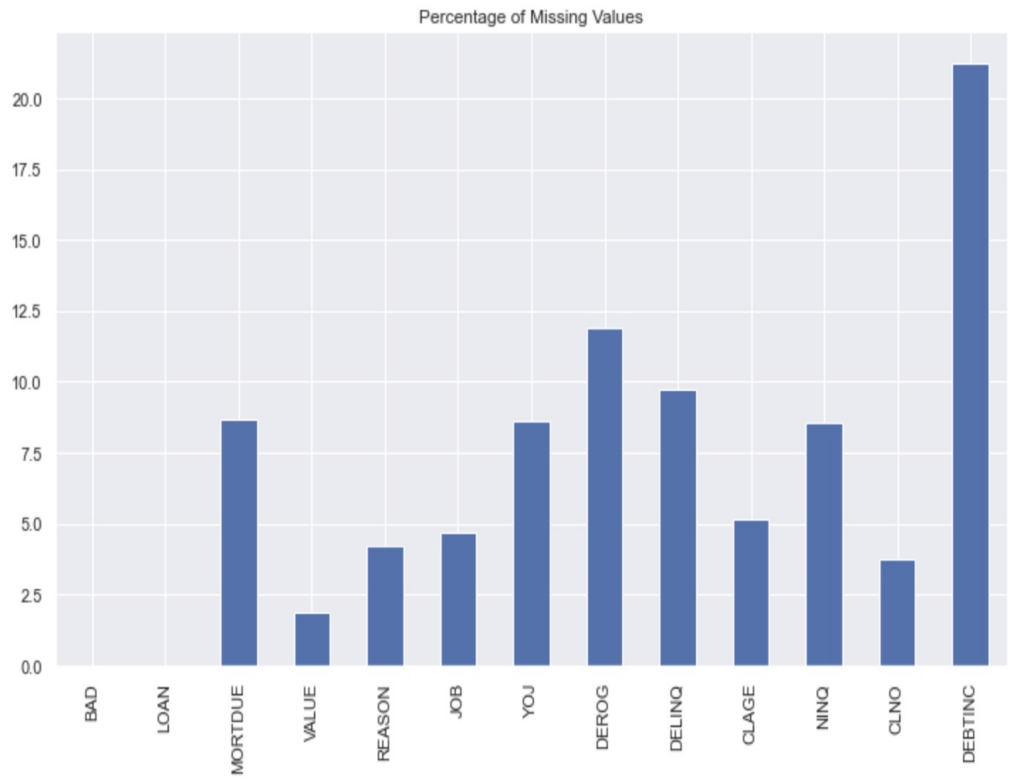


Exploratory Data Analysis

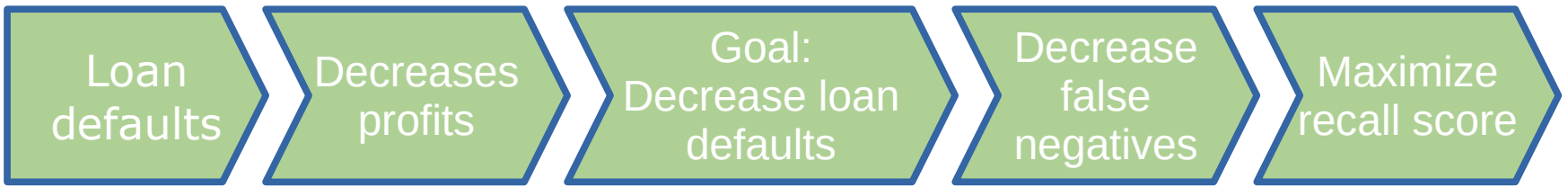
Dataset

- 5,960 observations, 13 columns/features
- Loan default rate: 20%
- Non-defaulters: CLAGE
- Defaulters: NINQ, CLNO, DEBTINC
- Has outliers
- Has missing values

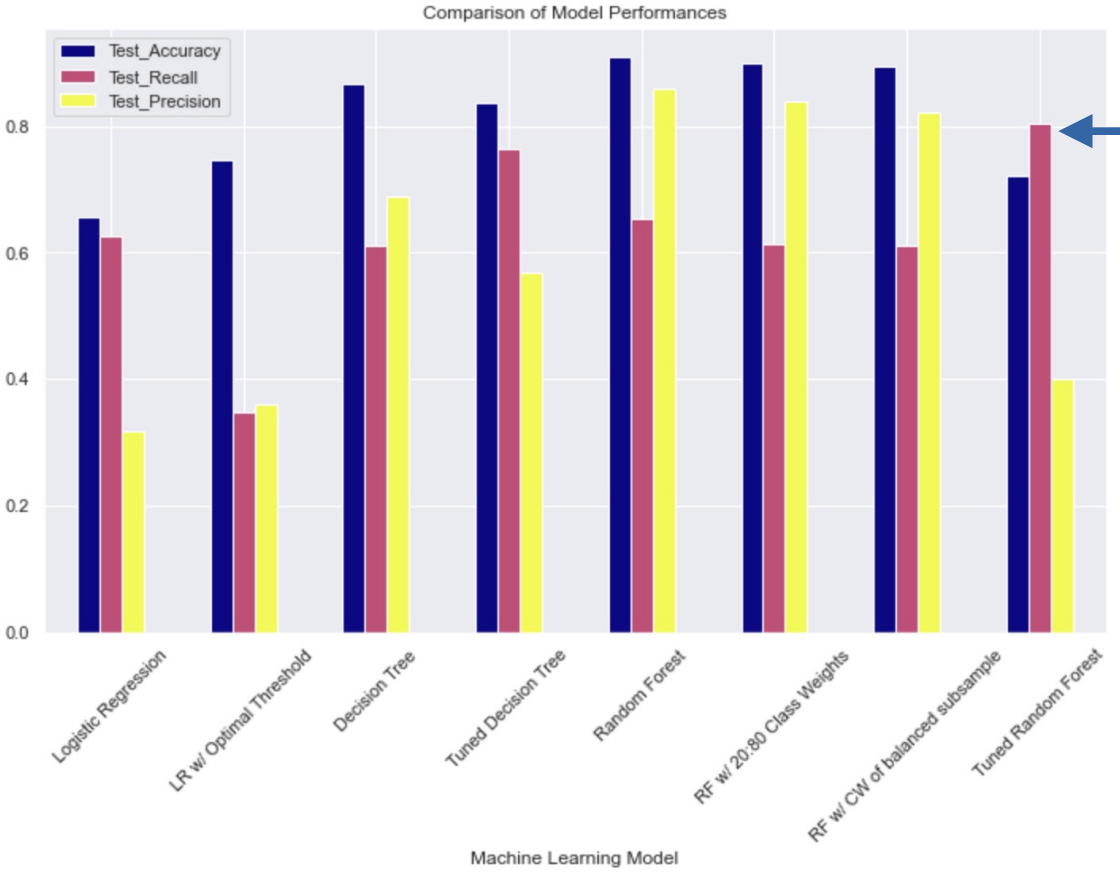
Note: Meanings of data acronyms in Appendix



Performance Metric



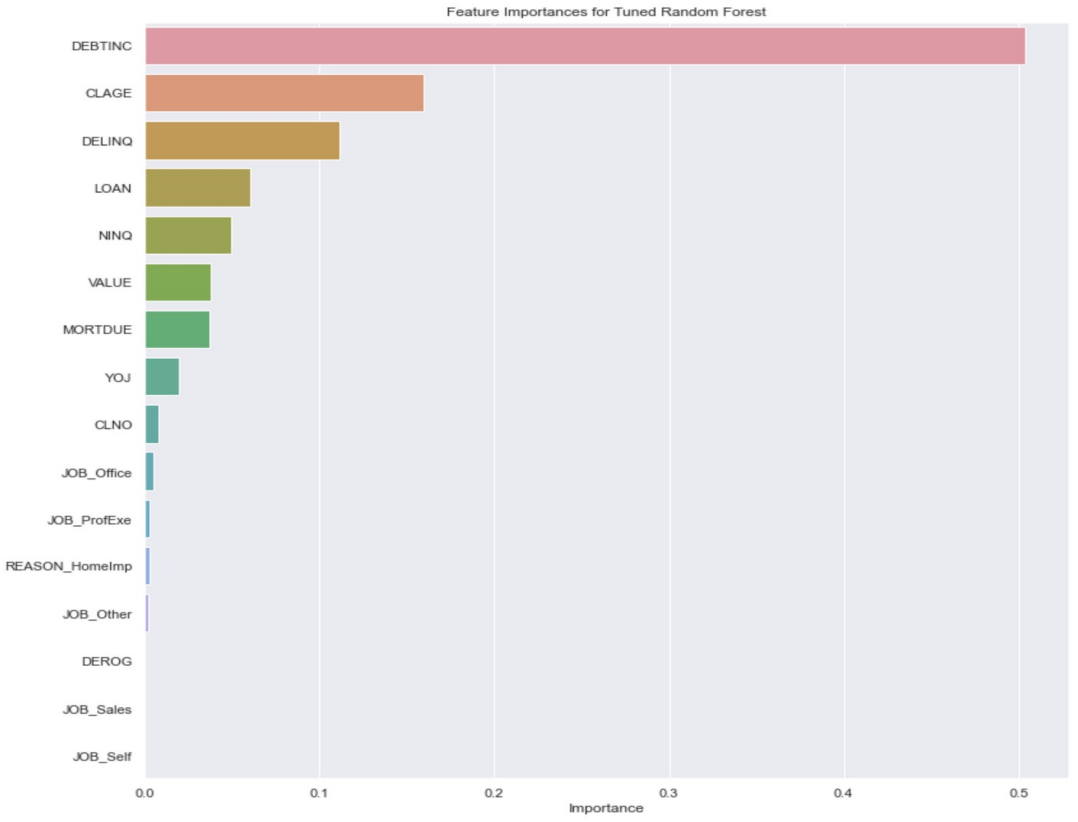
Performance Comparison



Highest recall score

Note: Comparison Table is in Appendix

Features Importance



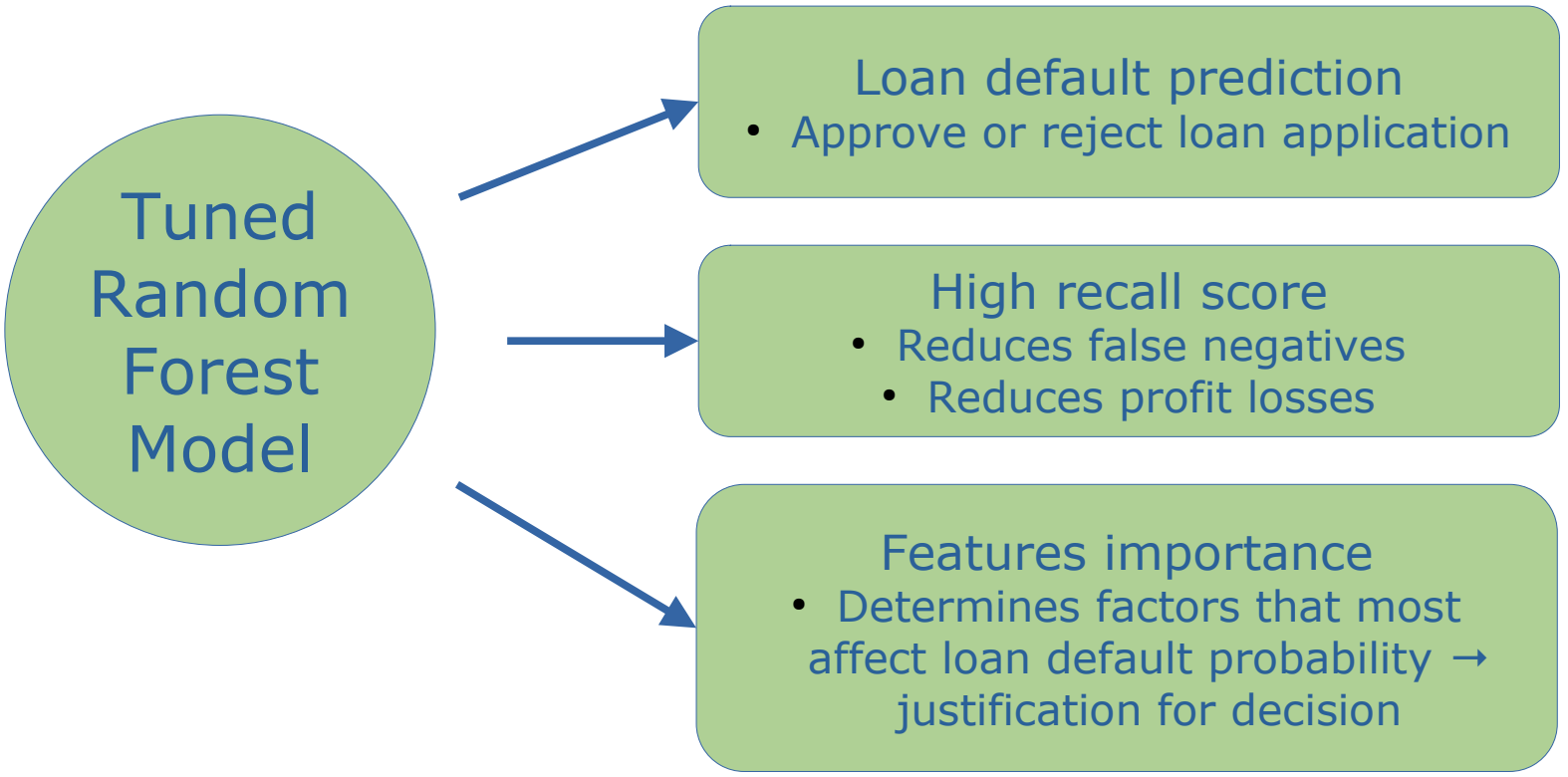
DEBTINC

- debt-to-income ratio
- has the most effect on the likelihood of loan defaults
- the higher the DEBTINC, the greater the chances of defaulting on the loan

CLAGE

DELINQ

Proposed Model



Limitations

- Dataset – missing values
- Hyperparameter tuning – need for optimization
- Tradeoff between model performance and model interpretability
- Features importance – values to classify credit risk not explicit
- Precision decreases as recall is increased - false positives increase

Recommendations

- **Comprehensive data gathering** - lessens missing values, better data accuracy
- **Can tune further to increase recall** - adjust values of and/or add more hyperparameters
- **With new client data** - train, test, evaluate, tune model - achieve optimal model version
- **Performance metrics**
  - assess regularly as dataset and needs change
  - high recall score - ongoing goal to reduce false negatives (loan defaults)
  - depending on overall bank vision, balance the effects of precision, recall, accuracy
- **Features importance**
  - can change after tuning, prioritization heeded
  - give due attention, esp. debt-to-income ratio, during loan application appraisal
- **Explicit features' cut-off values** - bank has to set based on client database or use banking industry standards

In a nutshell

Tuned Random Forest – proposed model

- Predicts clients who are likely to default on their loan
- Gives important features to consider during loan approval – interpretable, provides justification for a decision
- Has high recall – addresses false negatives, reduces loan defaults
- Improved decision-making process

  → simplified

  → faster

  → more efficient

  → free from human biases and errors

Appendix

Performance Comparison Table

Machine Learning Model	Test_Accuracy	Test_Recall	Test_Precision
Logistic Regression	0.657159	0.627451	0.318182
LR w/ Optimal Threshold	0.746085	0.347339	0.359420
Decision Tree	0.867450	0.610644	0.689873
Tuned Decision Tree	0.836689	0.764706	0.567568
Random Forest	0.909396	0.652661	0.859779
RF w/ 20:80 Class Weights	0.899329	0.613445	0.839080
RF w/ CW of balanced subsample	0.895973	0.610644	0.822642
Tuned Random Forest	0.721477	0.803922	0.401399

The Tuned Random Forest Model has the highest recall score among all the models that were built

Data Description

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.

- **BAD:** 1 = Client defaulted on loan, 0 = loan repaid
- **LOAN:** Amount of loan approved.
- **MORTDUE:** Amount due on the existing mortgage.
- **VALUE:** Current value of the property.
- **REASON:** Reason for the loan request. (HomeImp = home improvement, DebtCon= debt consolidation which means taking out a new loan to pay off other liabilities and consumer debts)
- **JOB:** The type of job that the loan applicant has, such as manager, self, etc.

- **YOJ:** Years at present job.
- **DEROG:** Number of major derogatory reports (which indicates a serious delinquency or late payments).
- **DELINQ:** Number of delinquent credit lines (a line of credit becomes delinquent when a borrower does not make the minimum required payments 30 to 60 days past the day on which the payments were due).
- **CLAGE:** Age of the oldest credit line in months.
- **NINQ:** Number of recent credit inquiries.
- **CLNO:** Number of existing credit lines.
- **DEBTINC:** Debt-to-income ratio (All of your monthly debt payments divided by your gross monthly income. This number is one way lenders measure your ability to manage the monthly payments to repay the money you plan to borrow.)