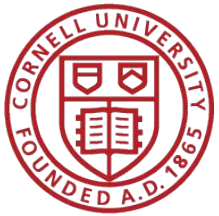


Reducing Risk and Increasing Customer Trust

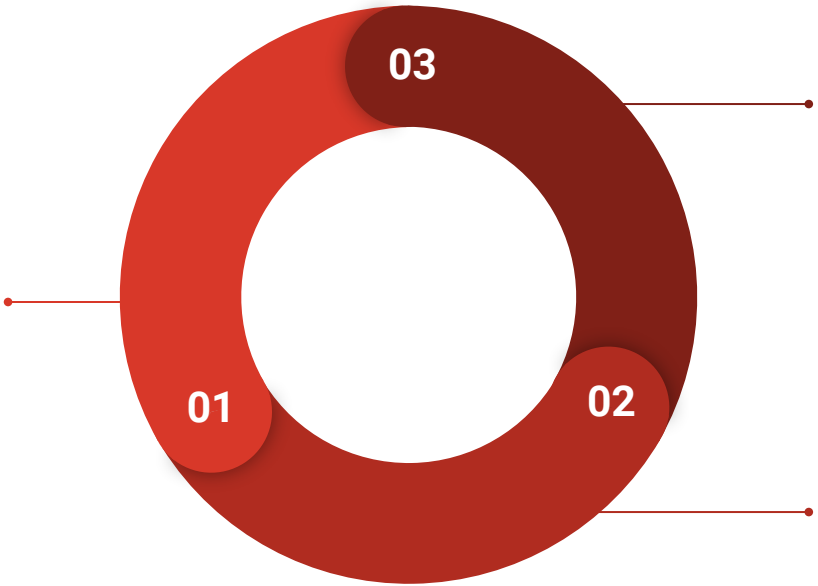
Using Machine Learning to Predict Loan Defaults

Benjamin Hu, Hanna Yen, Kristi Zhang

Can We Predict Loan Defaults?

What is defaulting?

Defaulting on a loan is the process of failing to keep up with the loan repayments agreed upon. It significantly lowers the credit score of the client and causes the bank to incur losses.



Why do these decisions matter?

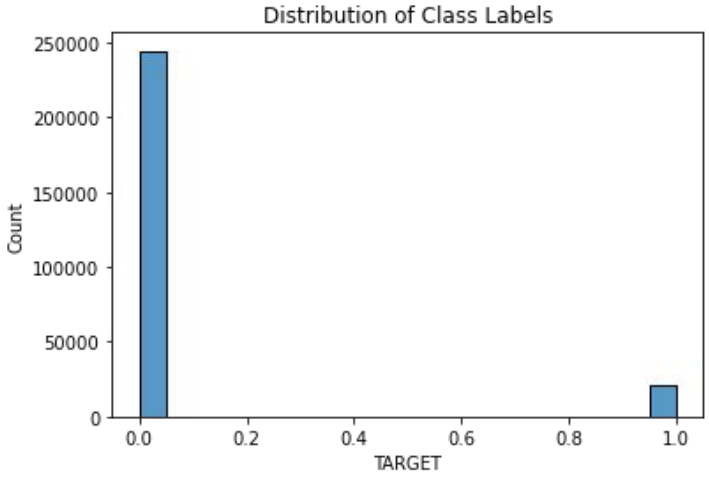
Predicting whether a client will default can reduce the risk our bank incurs upon approving loans, but also building customer trust by only approving customers we know will be able to take on the responsibility of a loan.

Decision Problem

Given a customer's application for a loan, can we accurately predict whether or not the customer will default on the loan?

The Dataset

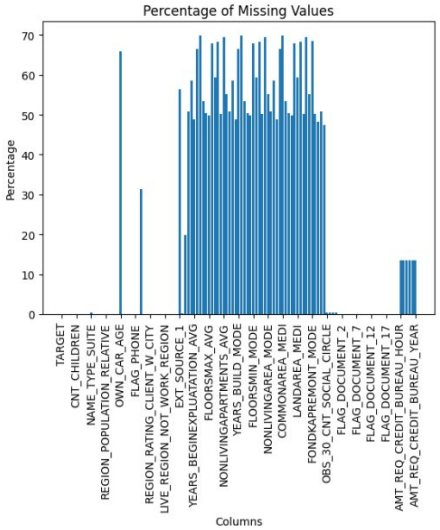
- Labeled data:
  - 1 if the customer defaulted
  - 0 if they did not
- 307, 511 data points
  - Each is a unique applicant
- 122 features
  - e.g. applicant age/gender/income, type of loan, whether they own a car/house, whether they have children, housing information
- Many missing entries



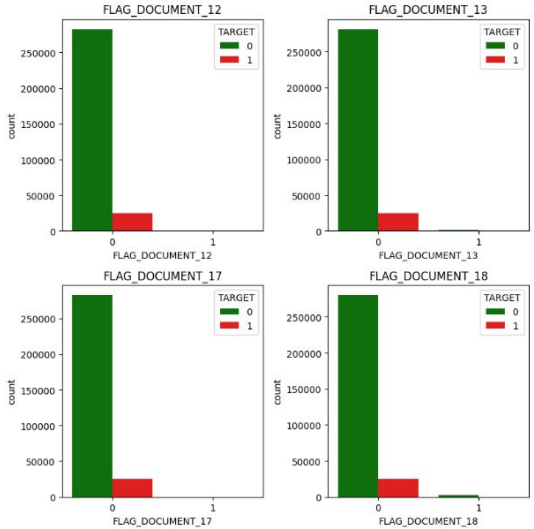
Important observation: the data is heavily imbalanced! 92% of the data has a label of 0.

Exploratory Data Analysis

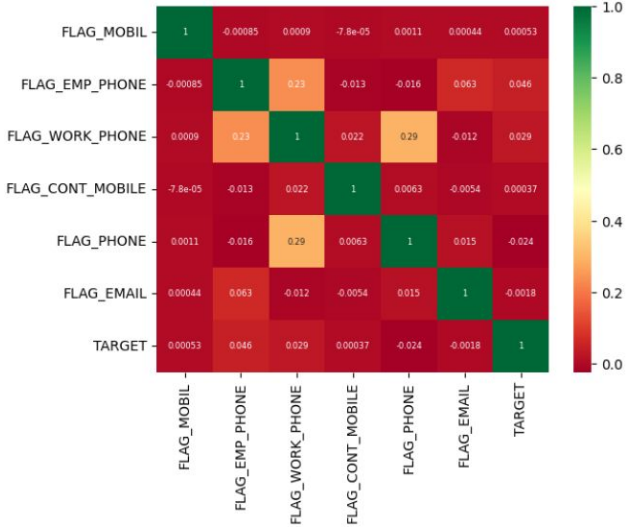
- Identifying useful features and removing the rest of the features



Null Percentage

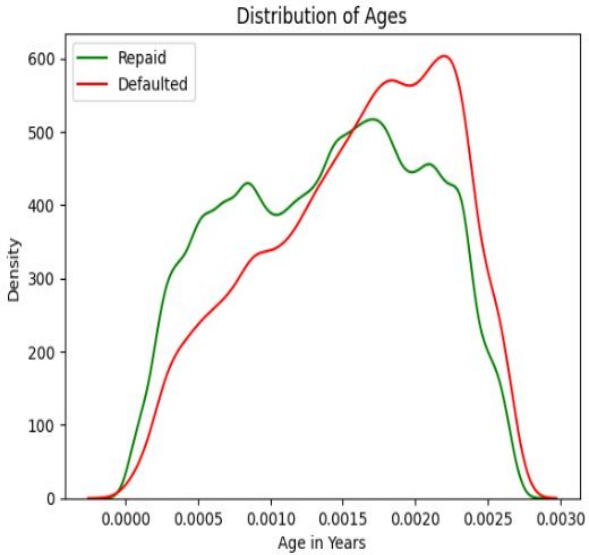


Distribution

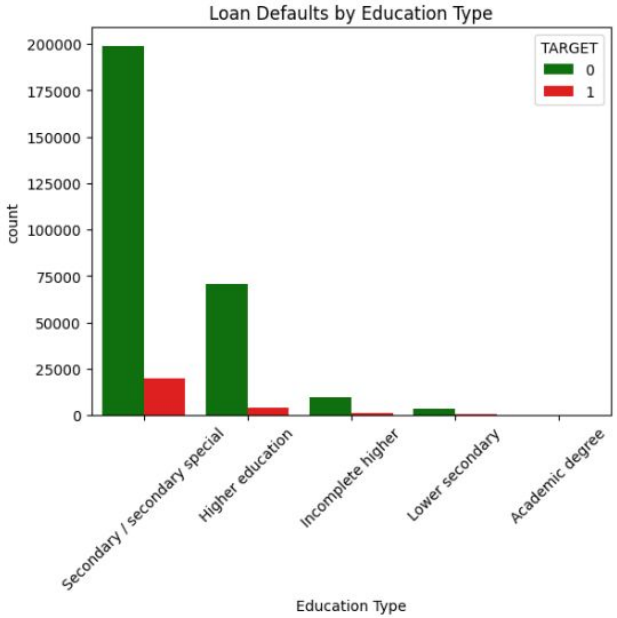


Correlation

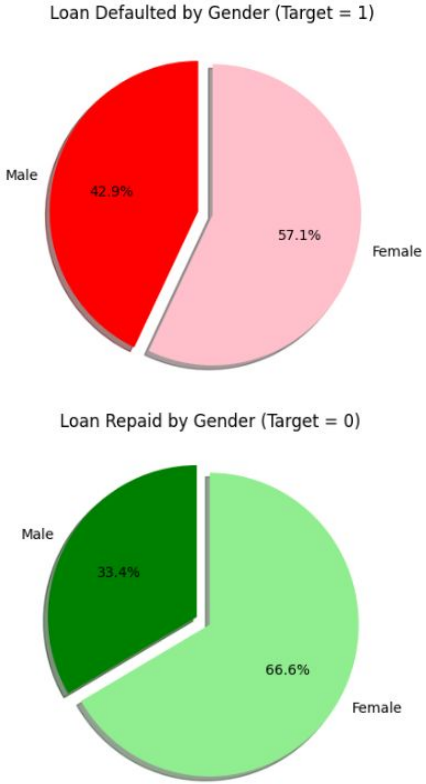
Exploratory Data Analysis



Age



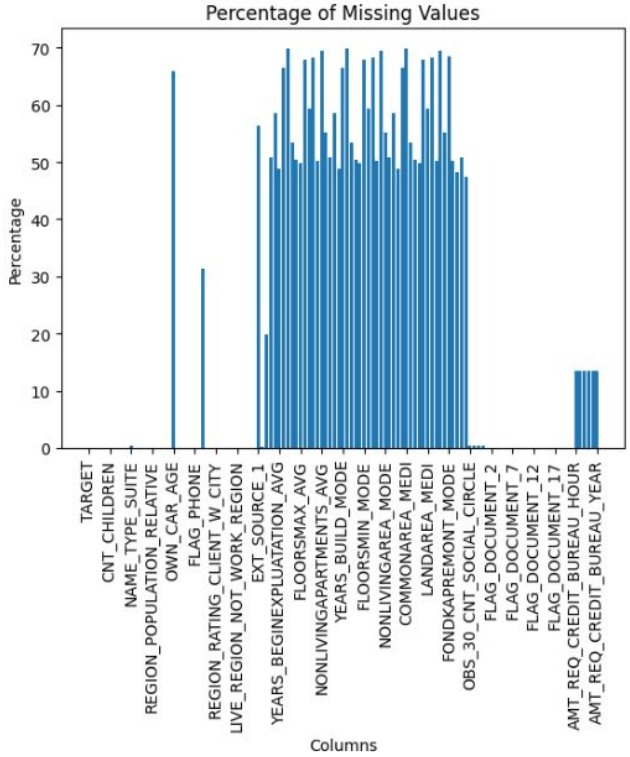
Education



Gender

Feature Engineering

- One-hot encoding of categorical variables
- Missing data:
  - EXT_SOURCE_1/2/3
    - Previous score was never assigned
    - **Solution:** impute missing entries with the mean
  - Other missing data with little occurrences
    - Drop records with a NaN
- Normalizing data
  - All features that did not have 0-1 values
  - Improves performance and training stability of model since all features are on the same scale

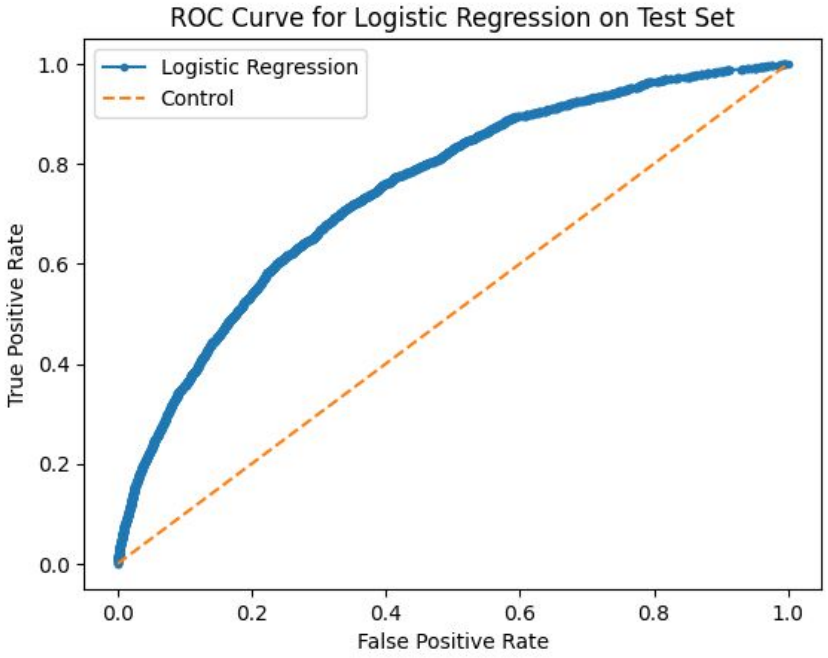


Model Formulation and Evaluation

- K-fold cross validation
  - Split data into k = 10 different subsets
  - Helps to better estimate performance
- F1 score
  - Recall: how many of the positive class samples present in the dataset were correctly identified by the model.
  - Precision: how many of the “positive” predictions made by the model were correct.
  - Harmonic mean between precision and recall, better metric for imbalanced data
- AUC ROC score
  - Unaffected by class imbalance
  - Reflects model’s performance over all classification thresholds
  - Reflects discriminatory power

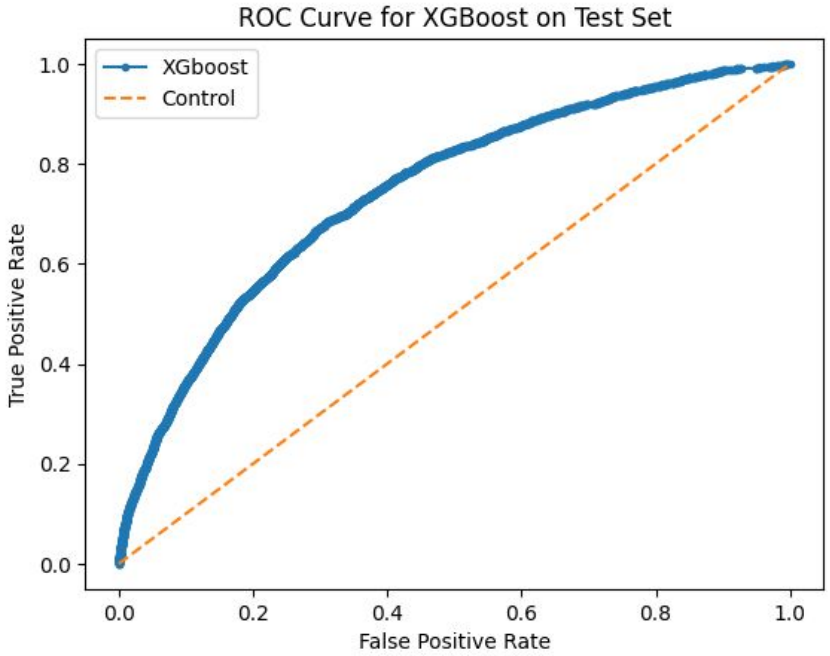
Model 1: Logistic Regression

	Training	Test
Accuracy	69.4%	68.54%
F1 Score	0.257	0.246
ROC AUC Score	0.753	0.747
False Negative Rate	0.319	0.305
False Positive Rate	0.3120	0.314



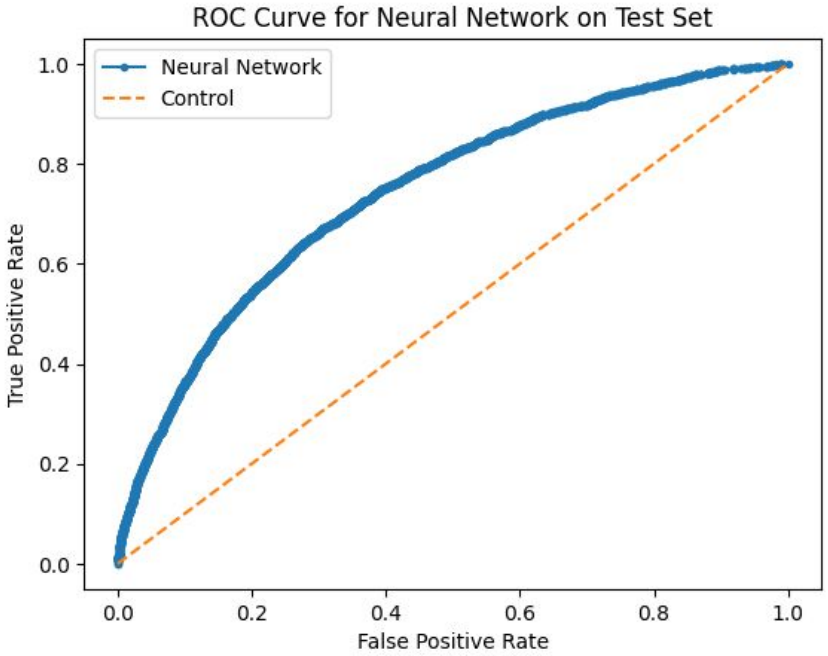
Model 2: Boosted Decision Tree

	Training	Test
Accuracy	77.31%	73.54%
F1 Score	0.359	0.260
ROC AUC Score	0.874	0.745
False Negative Rate	0.1808	0.3826
False Positive Rate	0.2308	0.2350



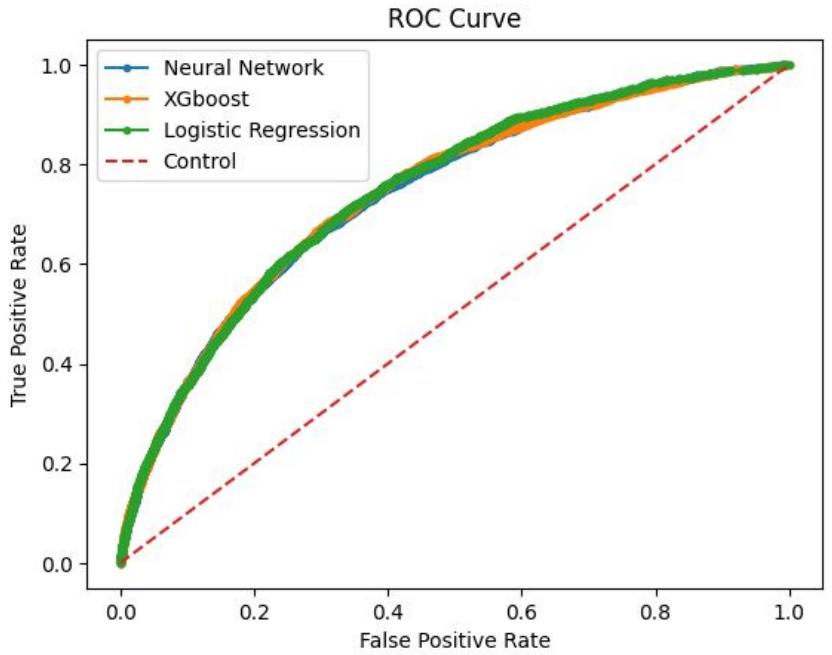
Model 3: Neural Network

	Training	Test
Accuracy	72.93%	72.63%
F1 Score	0.287	0.257
ROC AUC Score	0.783	0.743
False Negative Rate	0.3230	0.3736
False Positive Rate	0.2555	0.2656



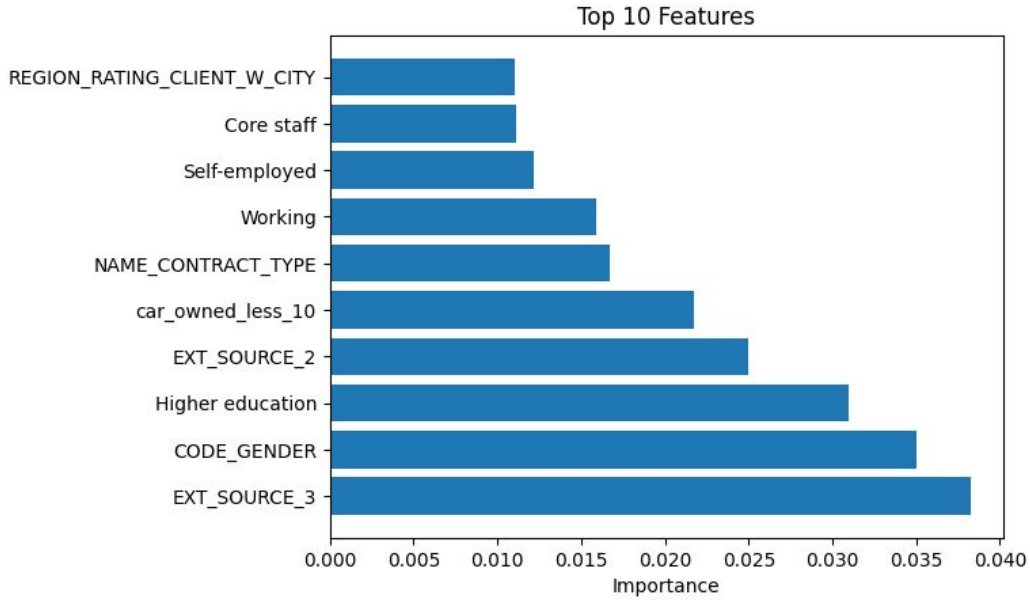
Interpreting Results: Best Model

- Boosted Tree!
  - Highest F1 score
  - Highest training AUC score
  - Lowest FPR
  - Lowest training FNR
  - Marginally second highest test AUC score
    - Overfitting may be present



Interpreting Results: Feature Importance

- **Most important:**
  - **Previous score**
  - **Gender**
  - **Education**
- Other predictive features:
  - Residence rating
  - Occupation
  - Income source
  - Type of loan
  - Age of car



Future Work

- If model is overfitting → reduce depth of tree, reduce number of features
- If false positives are less costly than false negatives → adjust prediction threshold
  - Predict defaults more aggressively
  - i.e. Adjust threshold from .5 to .4
- Utilize previous application information if possible
  - Cross compare information to identify trends
    - e.g. increasing/decreasing income, frequency of loan requests, etc.

Conclusion

Accurately predicting loan defaults is a very difficult task. While our model may not be sufficient as a standalone decision maker, it has enough predictive power to be an effective consulting tool for our bank's employees.

We recommend integrating the model into the bank's decision making pipeline in order to better protect both the bank and its customers regarding the distribution of loans.