Home Credit Default Risk

By Sangwon Baek and Simone Rittenhouse

What is the purpose of this ADS?

Inclusivity

Seeks to provide loans for those with little to no credit history

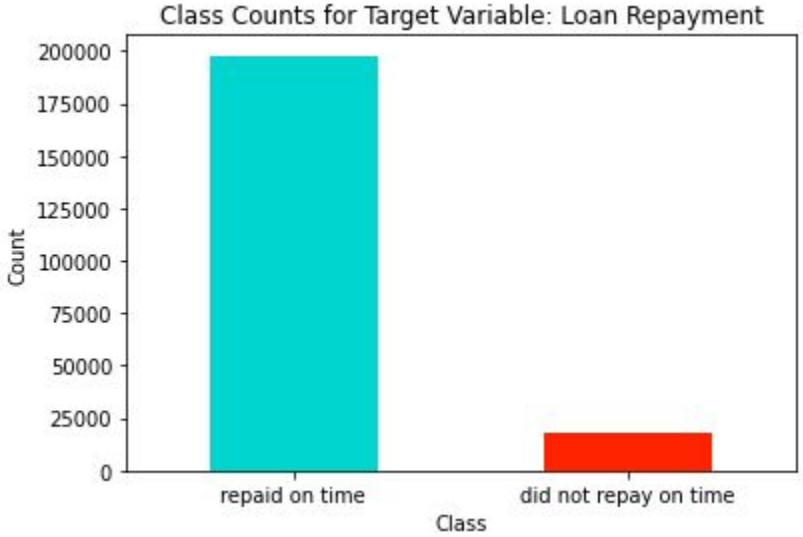
Simplicity

Used as a teaching tool for anyone interested in pursuing data science

Accuracy

Seeks to achieve high AUC for successful prediction

How Many People Defaulted?



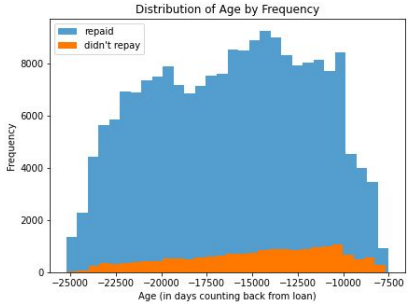
What are the Inputs and Outputs?

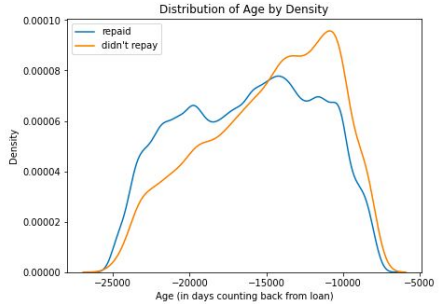
Input

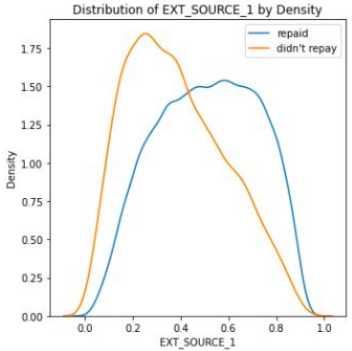
- 122 features
- 16 categorical variables
- 54.92% had missing values
- Insufficient metadata

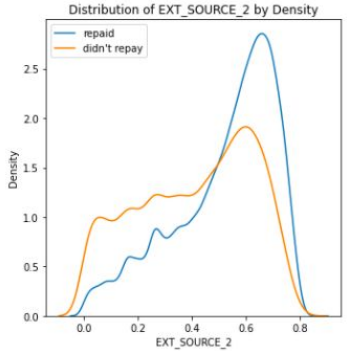
Output

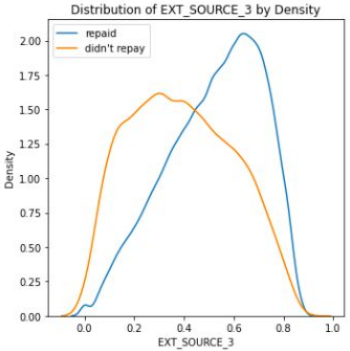
- Probability of default
- 0 = repayment
- 1 = default



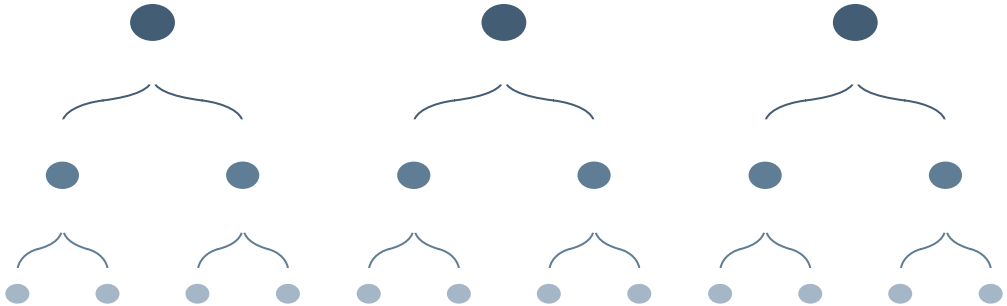








How did the ADS use these inputs?



Random Forest Classifier with 100 trees

Original training data split into a training/test set to obtain target labels for training and analysis.

AUC = 0.71 **Accuracy = 92.05%**

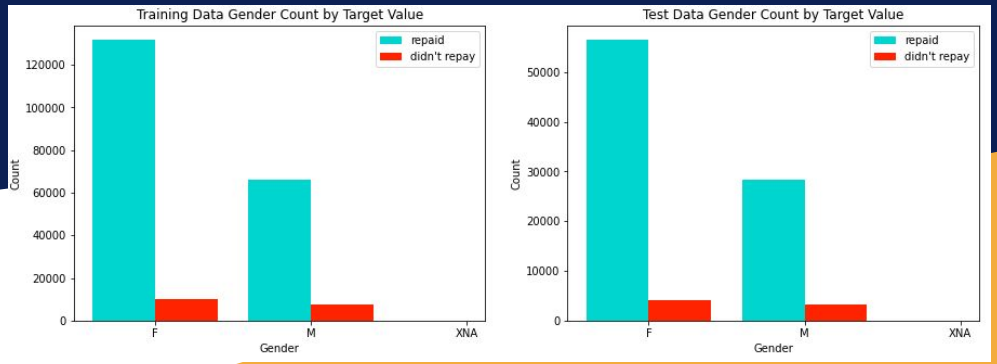
Who did we find to be risky?

Confusion Matrix of Random Forest:

	Predicted Repayment	Predicted No Repayment
Actual Repayment	84911	3
Actual No Repayment	7330	10

7 Men, 6 Women

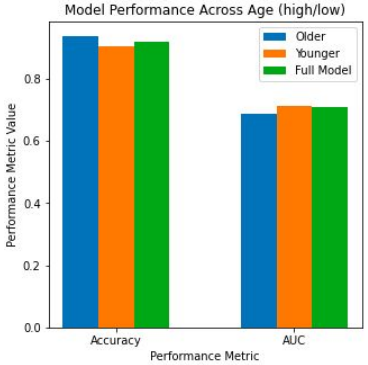
Average Age of 0.19 Compared to
Training Set Average Age of 0.48

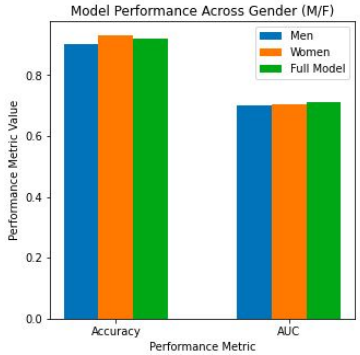


Are these predictions really fair?

01 Performance Across Subpopulations

Lower accuracy for men and younger people



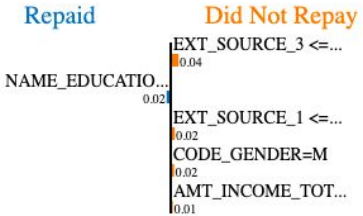


02 Fairness Metrics

FPR, FNR, Disparate Impact, Mean Difference

03 SP-LIME

Gender globally important feature



Feature	Value
EXT_SOURCE_3	0.19
NAME_EDUCATION_TYPE=Higher education	True
EXT_SOURCE_1	0.52
CODE_GENDER=M	True
AMT_INCOME_TOTAL	0.00

Summary

This ADS has several issues and should not be used publicly

Issues

1. Insufficient metadata
2. Unknown privacy protection strategy
3. High false negative rate
4. Unstable predictive power
5. Problems of fairness and bias against groups
6. Use of discriminatory features

Thank You!