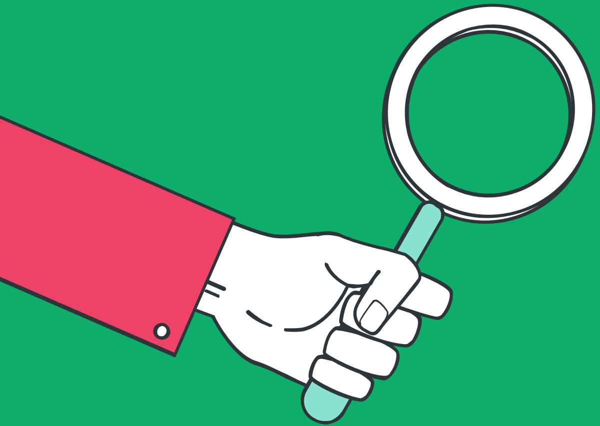
Loan Portfolio Risk using ML

By Marcos Dominguez, Data Scientist



01

Why ML?

How banks benefit as the end-user…

Automation

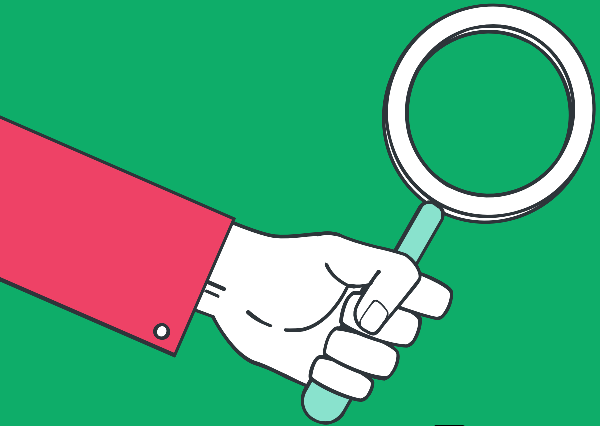
- Efficiency
- Minimize errors

Accuracy

- Prevent defaults
- Reduce loan loss risk

Cutting-Edge

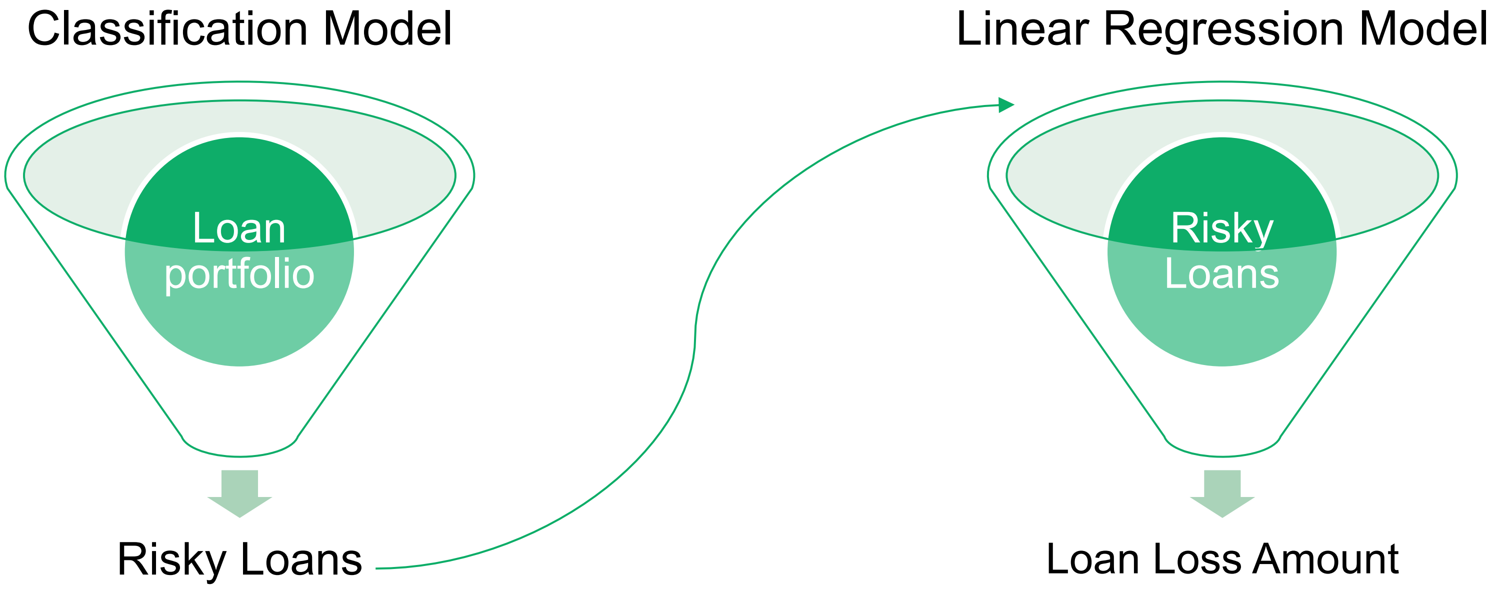
- Outperform competitors
- Increase portfolio

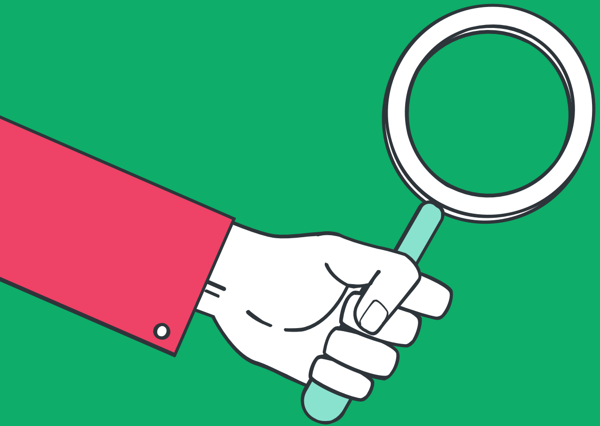


02

Prediction Pipeline using ML

Prediction pipeline in a nutshell…





03

Data

Loan Portfolio

Size

~100k loans

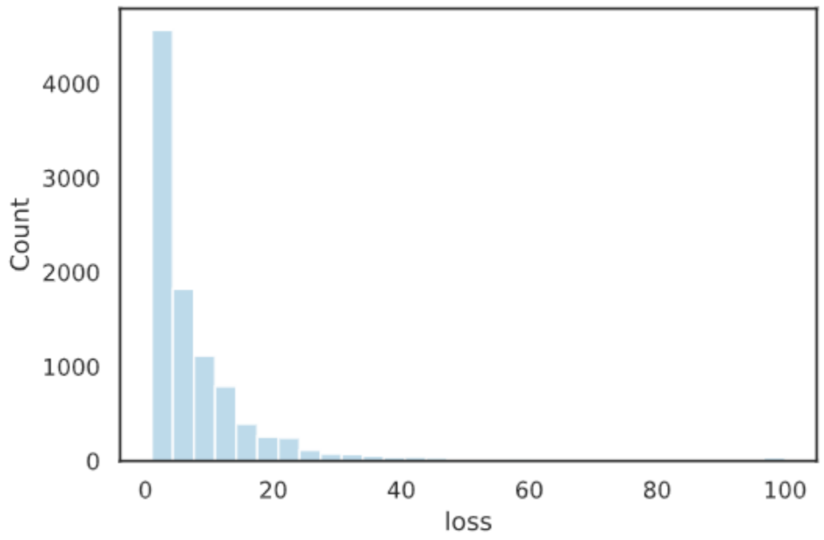
Features

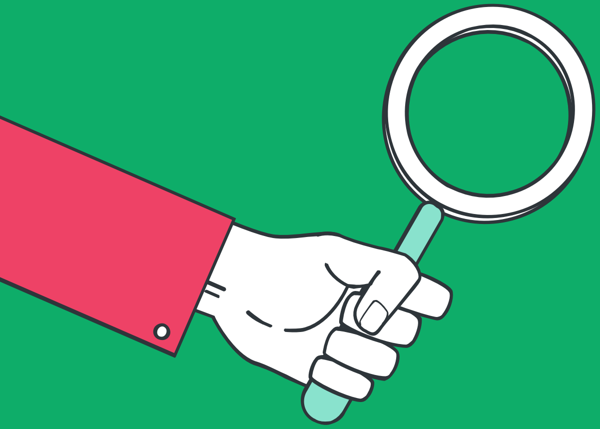
770 anonymized, numerical columns

Target

Loan loss: 0-100%

Imbalanced Target





04

Methodology

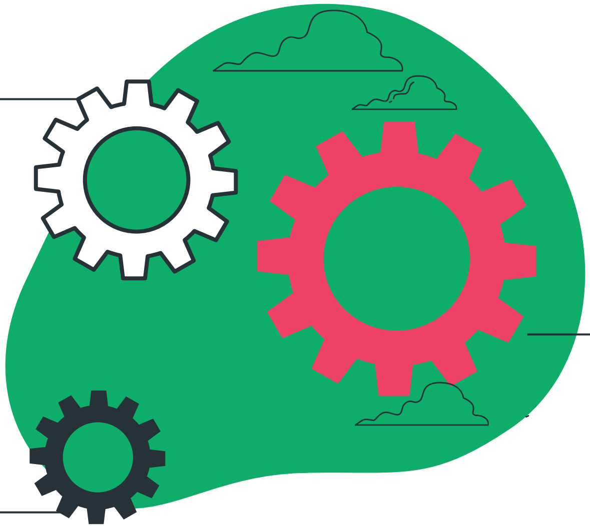
Methodology

EDA

Data cleaning, NaNs, data formatting

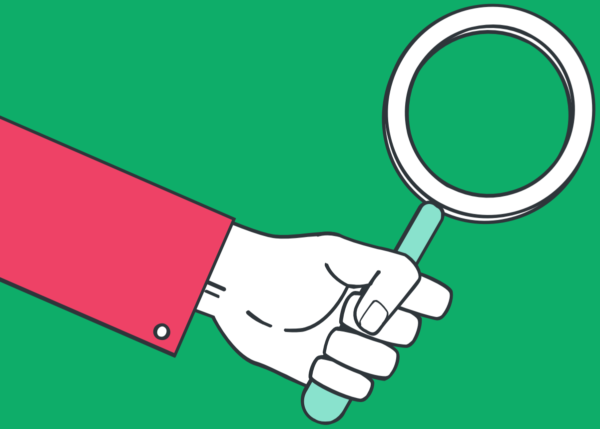
Model Selection

Undersampling, hyperparameter tuning, feature engineering



Prediction Pipeline

XGBoost for both classification, regression



05

Results

Performance Metrics

Classification

- Recall: 0.94
- Precision: 0.29
- ROC AUC: 0.92

Regression

- $R^2: 0.73$
- MAE: 0.29

Actuals vs Predicted

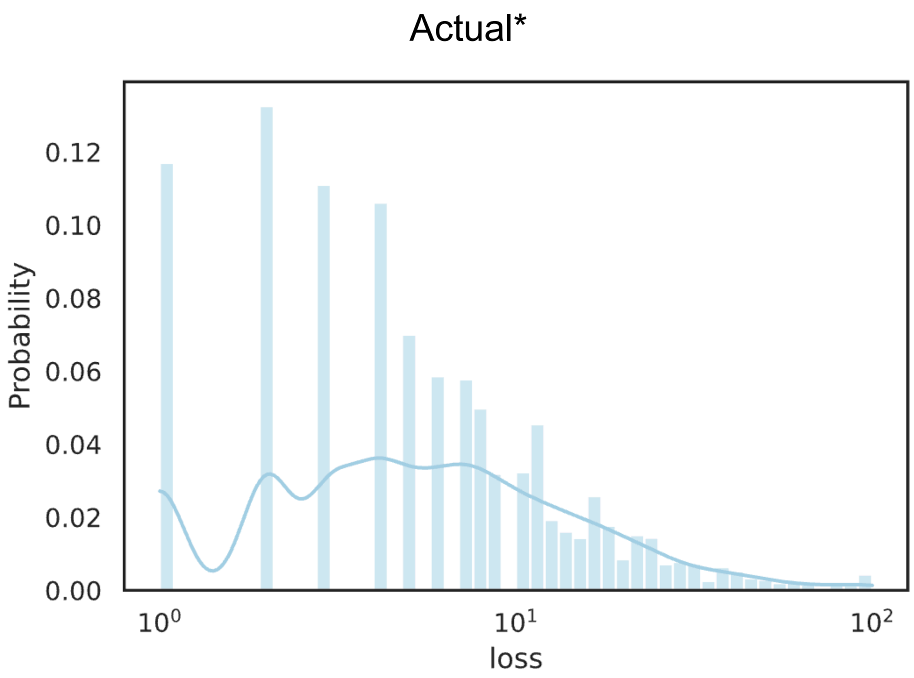
Risky Loans

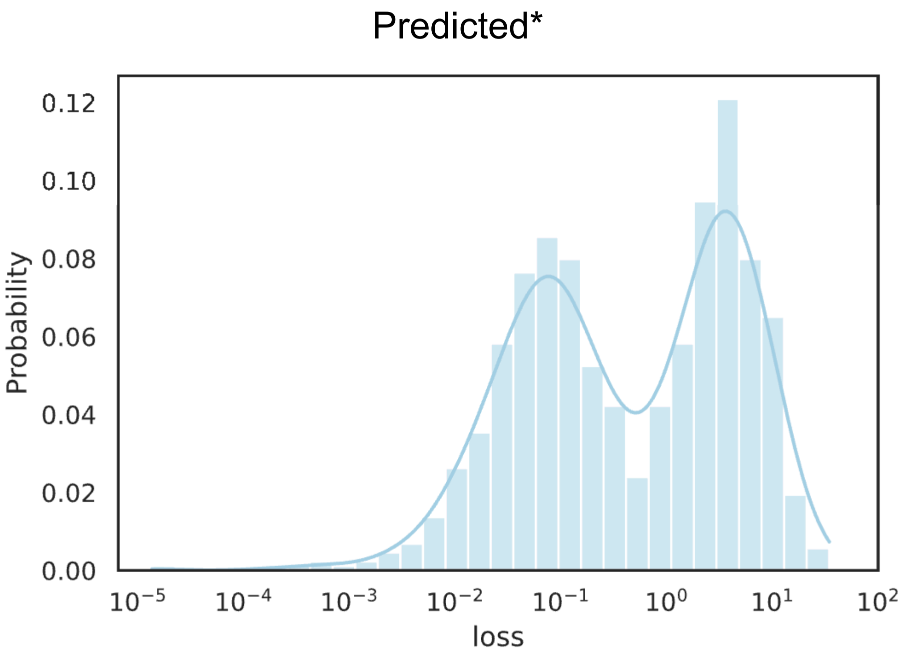
Actual	Predicted
9,783 loans 9.25% of portfolio	6,388 loans 6.06% of portfolio

Loan Loss

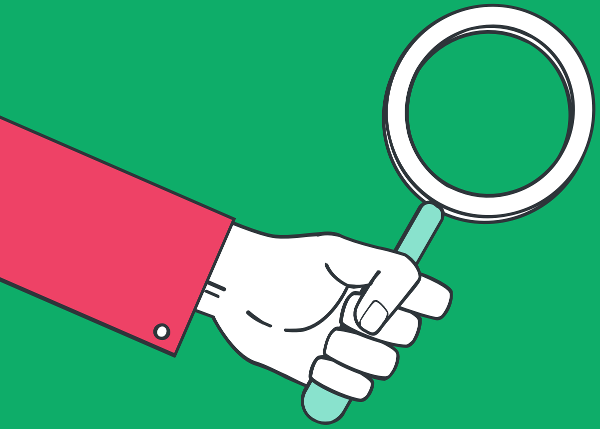
Actual	Predicted
Avg: 8.62% Min: 1.0% Max: 100.00%	Avg: 1.65% Min: 0.43% Max: 34.50%

Probability Density Functions: Actual vs Predicted





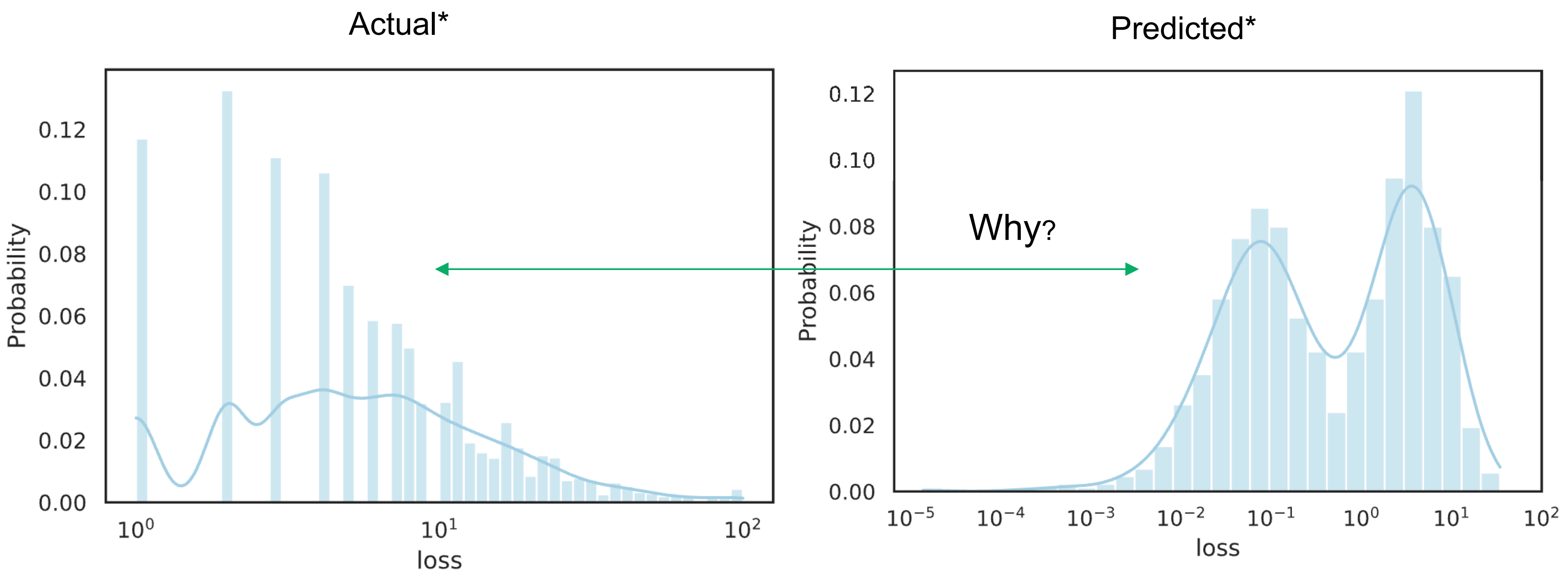
*Standardized using log_{10}



06

Next Steps

Research differences in distributions…



*Standardized using log_{10}

Acknowledgements

Thank you!

Questions? Contact me:

md.ghsd@gmail.com

github.com/mdominguez2010

Appendix

1. Model Confusion Matrix