Loan Default Prediction

MIT Professional Education Capstone Project

Jun 13 2024

Ilia Kolesnikov
i.a.kolesnikov@gmail.com

Agenda

01

Overview of the Problem

02

Approach to the Solution

03

Key Findings and Insights

04

Recos and Next Steps

Overview of the problem

Interests from home loans is a significant part of banks profits

Significant financial risks in cases of unpredicted defaults and missed opportunities

Manual process is prone to errors, time-consuming

Process should provide a justification for any adverse behavior

Approach to the solution

Random Forest Classifier

The top performer model with an accuracy of 92% and an AUC of 0.965

Hybrid Approach with human loop

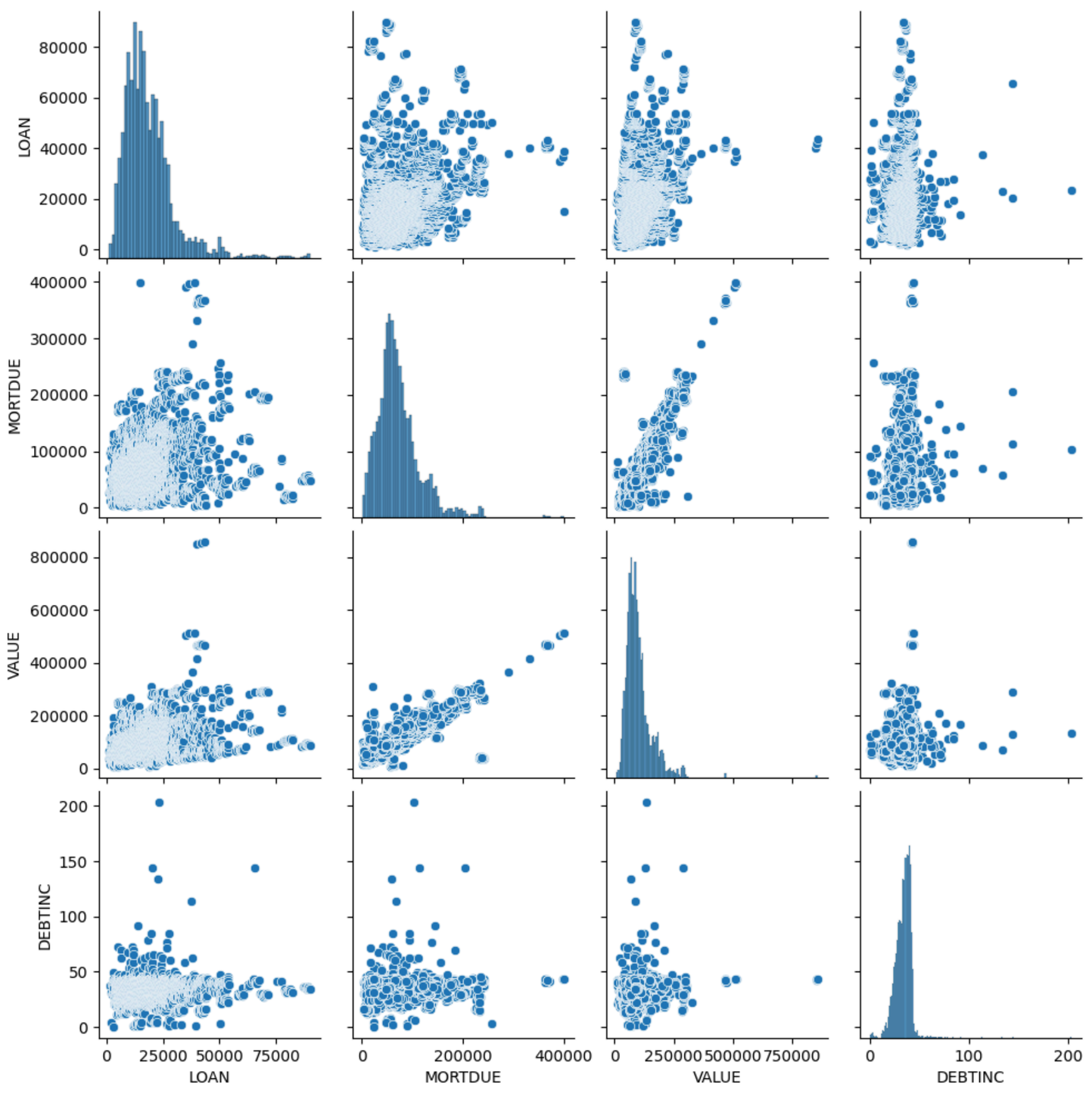
Combine the model's predictions with human expertise (only in cases when model doesn’t predict default)

Continuous Monitoring and Improvement

Data from human loop, new defaulters data and performance metrics to be used for continuous monitoring and improvement of the model

Key Findings and Insights

- Dataset has missing values, outliers and correlated variables. It should be preprocessed
- There is a trade-off between model performance and interpretability as the best model is less interpretable one
- Data Science model could provide significant and immediate benefits
- Human loop is important for implementation of data science model

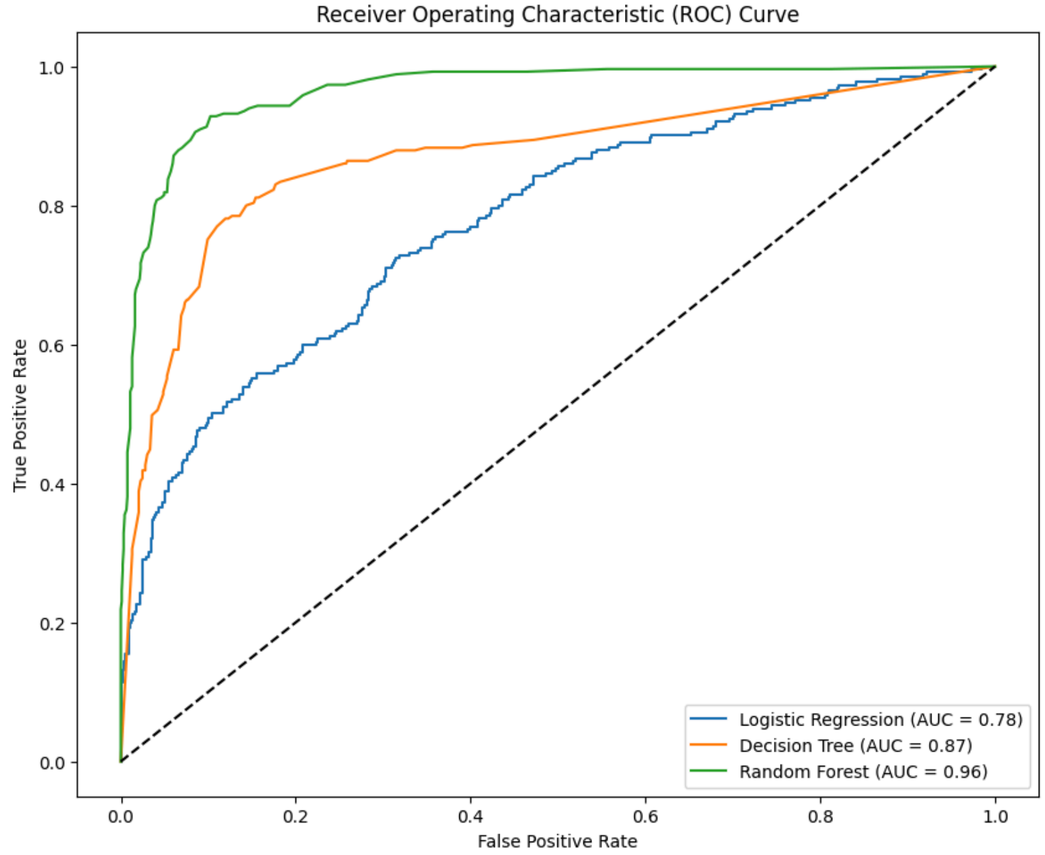


Key Recos

- Immediate deployment of Random Forest Classifier
- Additional actions to improve interpretability:
  - Provide personnel with data about features importance
  - Implement SHaply Additive exPlanations to show reason of decision in each case
- Implement hybrid approach:
  - If model identifies default, it should be immediately imlemented as its precision for 1-class is high (0.9)
  - If model doesn't identify default, additional check from the loan specialist should be done. Loan specialist could refuse potential client and dissaprove the loan. The specialist’s decision should be recorded, providing clear justifications for any overrides.
  - Data from the specialist reviews should be analyzed to improve model performance and reduce biases over time
- Continous monitoring and improvement
  - Data from this human loop and new defaulters data should be used to continously monitoring and improvement of the model. All model metrics should be also monitored and model hyperparametres should be fine-tuned in order to maintain model relevancy and improve its performace

Classification report

	precision	recall	f1-score	support
0	0.92	0.98	0.95	927
1	0.90	0.70	0.79	265
accuracy			0.92	1192
macro avg	0.91	0.84	0.87	1192
weighted avg	0.92	0.92	0.91	1192



Thank you!

Ilia Kolesnikov
i.a.kolesnikov@gmail.com