LendingClub Scoring model

ANDREA Koltai & DANIEL Torres

Goal: Predictive model to select defaulted loans

Dataset: LendingClub Personal loans from years 2014*
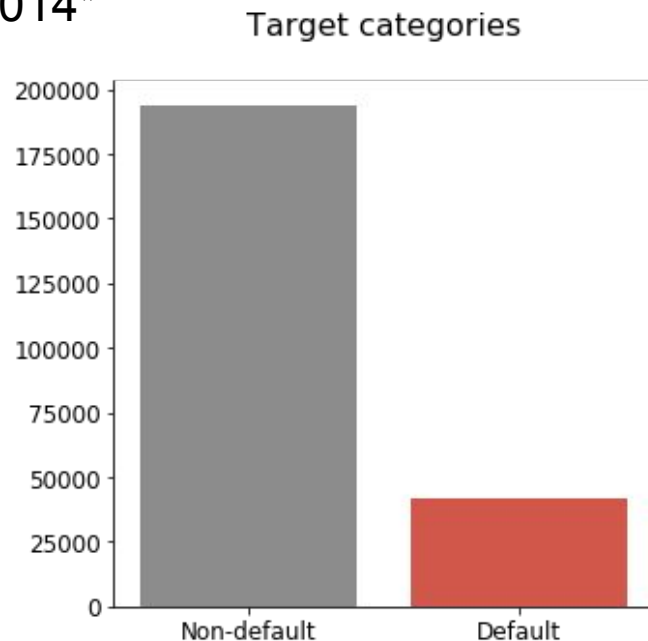
Features: after feature selection 49

(44 continuous, 5 categorical)

Observations: 235 629 (17.7% defaulted)

Model: Logistic regression

Metric for optimization: F1-score

Target categories

Feature correlation heatmap

Confusion **Matrix** (with SMOTE)

	Predicted non default	Predicted default
Actual non default	99 924	46 233
Actual default	11 312	18 760

F1-score: 0.39   **Cut-off:** 0.52   **Recall:** 0.60   **Precision:** 0.28
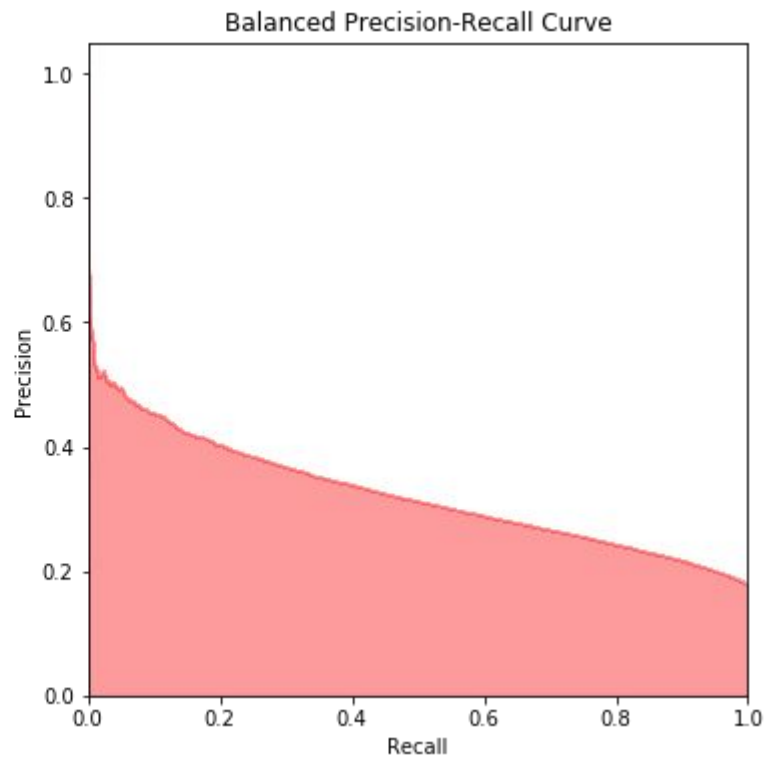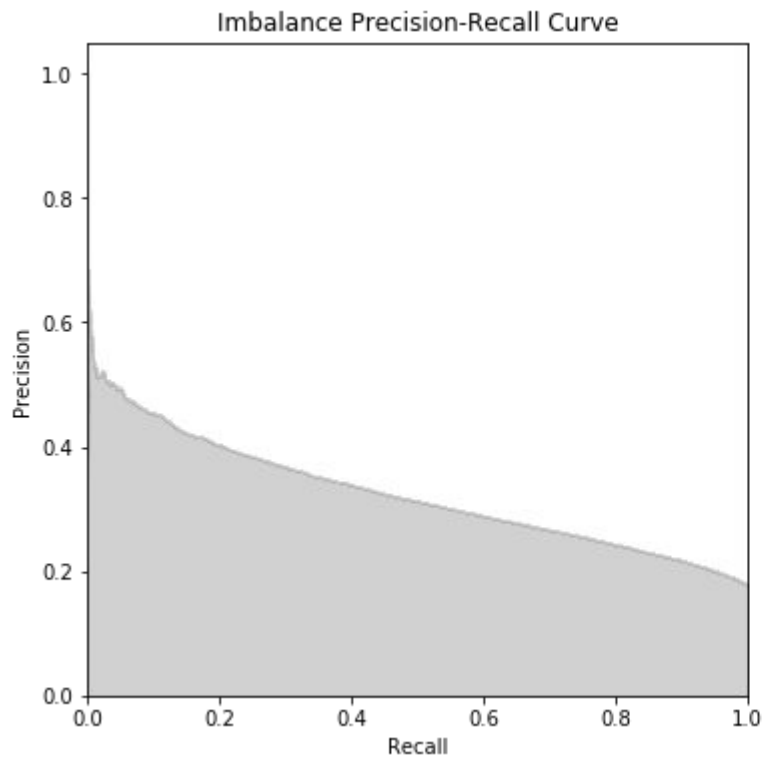
F1-score*: 0.39   Cut-off*: 0.19   Recall*: 0.60   Precision*: 0.29

* w/o imbalanced strategy

Precision-Recall Curve

Conclusion - future task

By smote to offset imbalanced data, our precision is slightly better than randomly guessing

Applying advanced models (random forest, neural network)

Feature engineering: creating more features to improve efficiency

Metrics: calculating with real cost weights for optimizations