

Loan Default Predictor

Anton Shesternev

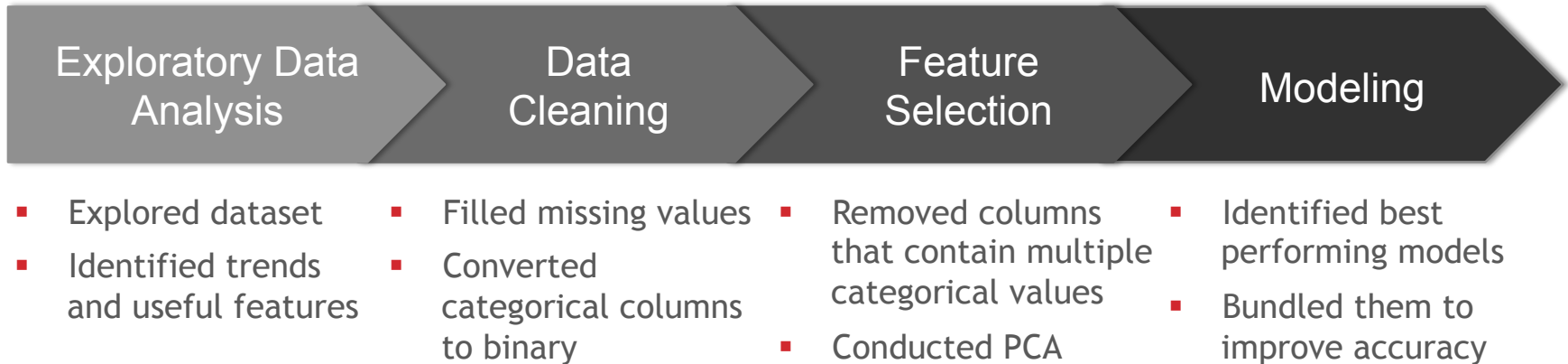
March 27, 2019

Project Goal and High-Level Approach

Goal:

- Develop a model that could identify loans that will go in default with high degree of accuracy
- The model could be used to assess default risk for existing loans and for loan adjudication

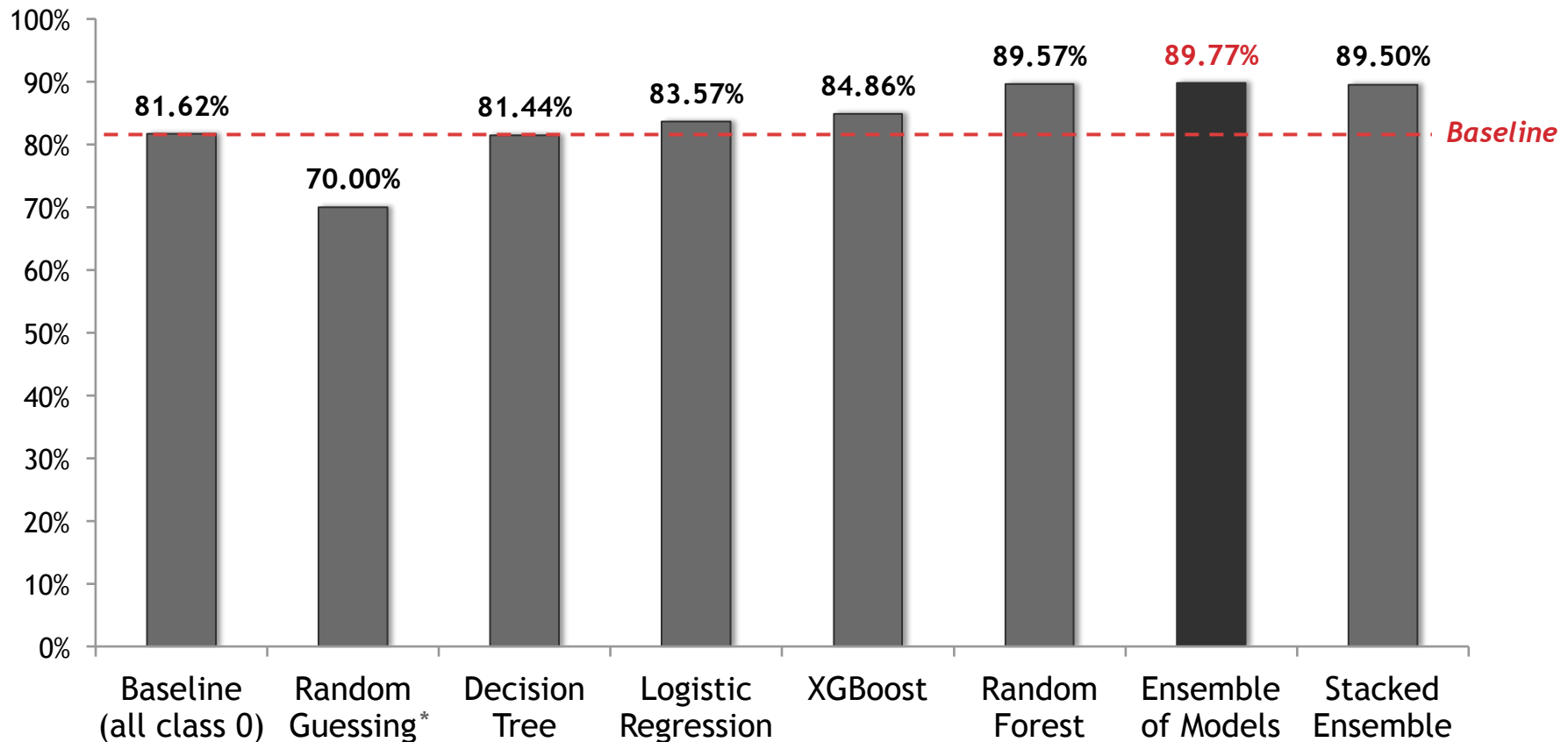
High-Level Approach:



Outcome

Accuracy Score:

- Achieved prediction accuracy of **89.77%**
- Accuracy score is ~ 8.2% above the baseline - see details below

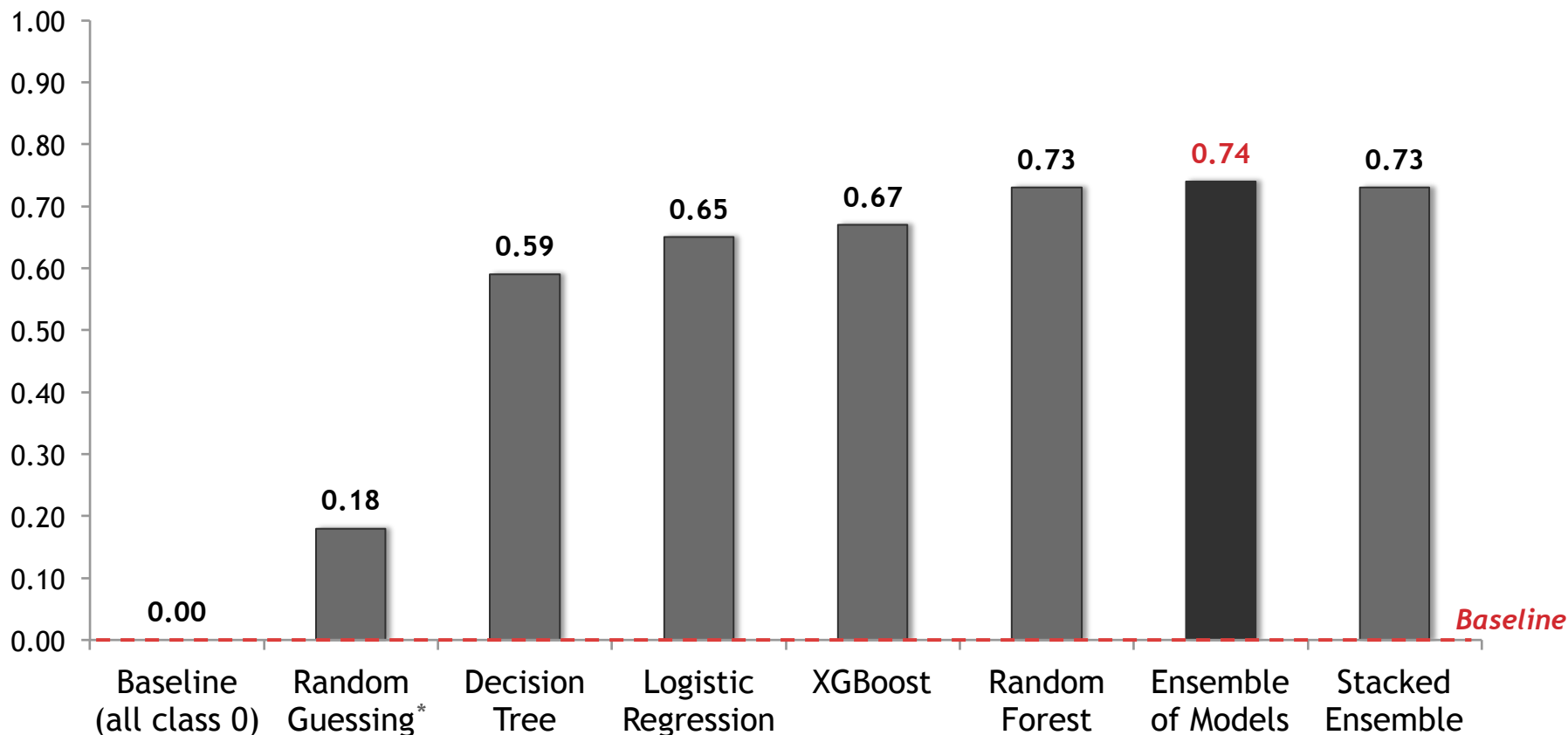


* ~ 80% of guesses are class 0 (paid in full) and ~ 20% of guesses are class 1 (in default)

Outcome

F1 Score (another accuracy measure):

- Achieved F1 score of **0.74** (lowest possible value is 0 and highest is 1)
- Baseline F1 score is 0 - see details below



* ~ 80% of guesses are class 0 (paid in full) and ~ 20% of guesses are class 1 (in default)

Appendix

Appendix

Data Source:

- “Lending Club Loan Data” data set from kaggle.com
- Data set contains complete loan data for all loans issued through the 2007-2015, including the current loan status and payment information
- The data set is a matrix of ~ 890,000 observations (loans issued through Lending Club peer-to-peer lending platform) and 74 columns/features

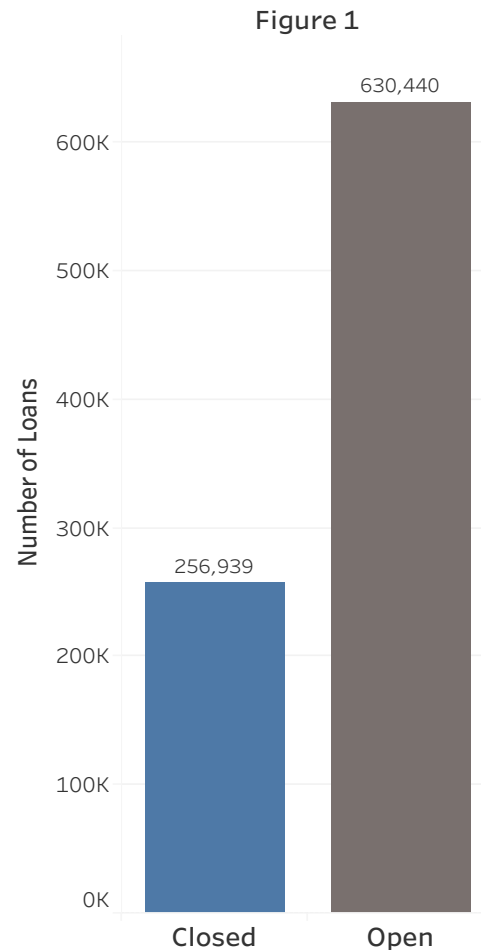
Methodology:

- Used data on ~ 257,000 closed loans (i.e. loans that are either paid off, in default or charged off) to train models and make predictions
- Data on open loans that have not been paid in full and have not gone in default was not used, as the final status of these loans is unknown

Appendix

There is data on the total of 887,379 issued loans in the dataset:

- 630,440 loans or ~ 71% of loans are open
- 256,939 loans or ~ 29% of loans are closed (see Figure 1 below)



Appendix

Out of 256,939 **closed** loans:

- 209,711 or ~ 82% of loans were paid in full and
- 47,228 or ~ 18% of loans have gone in default (see Figure 2 below)

