IÉSEG SCHOOL OF MANAGEMENT

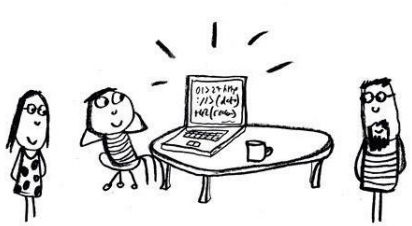
Statistical & Machine Learning

Minh PHAN

Group project: In Class Kaggle competition

AGENDA

HAPPINESS IS



...when your code runs without error.

- TEAM PRESENTATION
- DATA SUMMARY AND PROCESSING
- METHODOLOGY
- RESULTS
- SOLUTION IMPLEMENTATION

4 FROMAGE PIZZA

Fernando Delgado

Used to work in finance as a Global Business Services Trainee, but the hunger for data (and pain au chocolat) led him to…. France and an MSc in Big Data Analytics. He is an expert in data wrangling, predictive modelling and out-of-the box thinking. When he's not asking “how can we improve our Kaggle score“, he enjoys eating Mexican food, travelling and playing music.

Sofie Ghysels

Has a background in HR, international trade and communications and has worked in various HR roles at multinational companies in Belgium and France. She has a knack for organization and a passion for reporting and data visualization tools. When not scheduling in team meetings or sharing Belgian chocolates with her colleagues, Sofie can be found watching reality series, walking or hanging out with friends.

Nour Azar

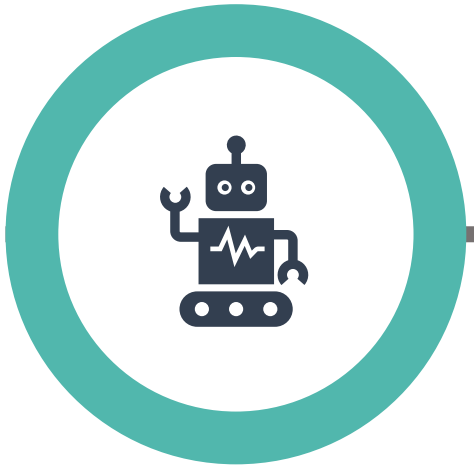
Has more than five years of experience as an accountant and external auditor in highly demanding environments. She is known for her critical thinking skillset and problem-solving abilities. Nour has a passion for machine learning and enjoys getting her hands dirty during all aspects of the modelling pipeline. Nour did community volunteering in her home country Lebanon and in her free time you might see her eating delicious food and getting outside.

BUSINESS PROBLEM

- **Portuguese Bank Telemarketing Campaign**
- **Classification Goal → Subscribe or Not**
- **Success of Bank Telemarketing**

IÉSEG SCHOOL OF MANAGEMENT

DATA SCIENCE PROBLEM



Problem Type

Classification Supervised
(Yes/No 1 or 0)

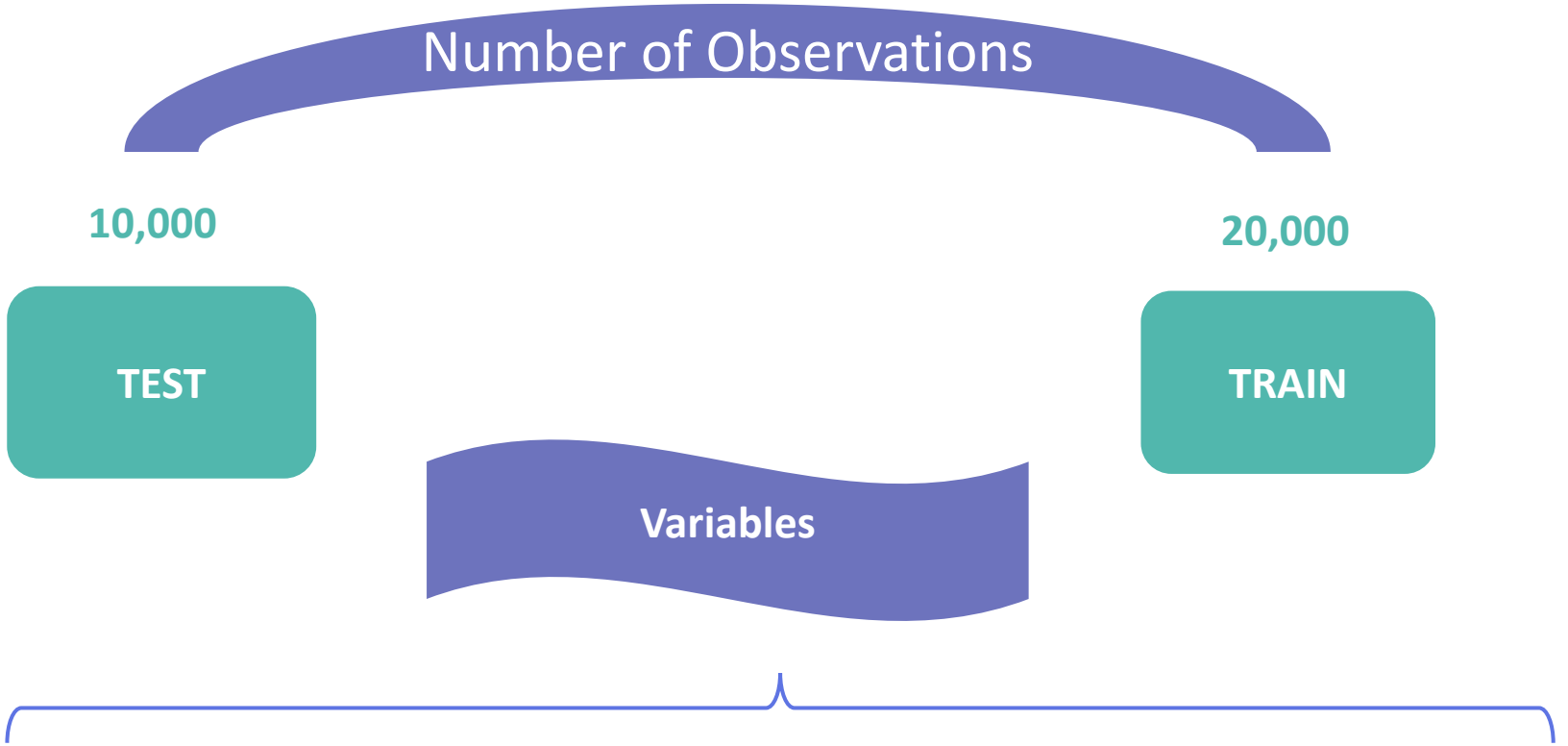
Data-Science Tools

Logistic Regression
Random Forest
Gradient Boosting
K Nearest Neighbors
Support Vector Machine

Expected Results

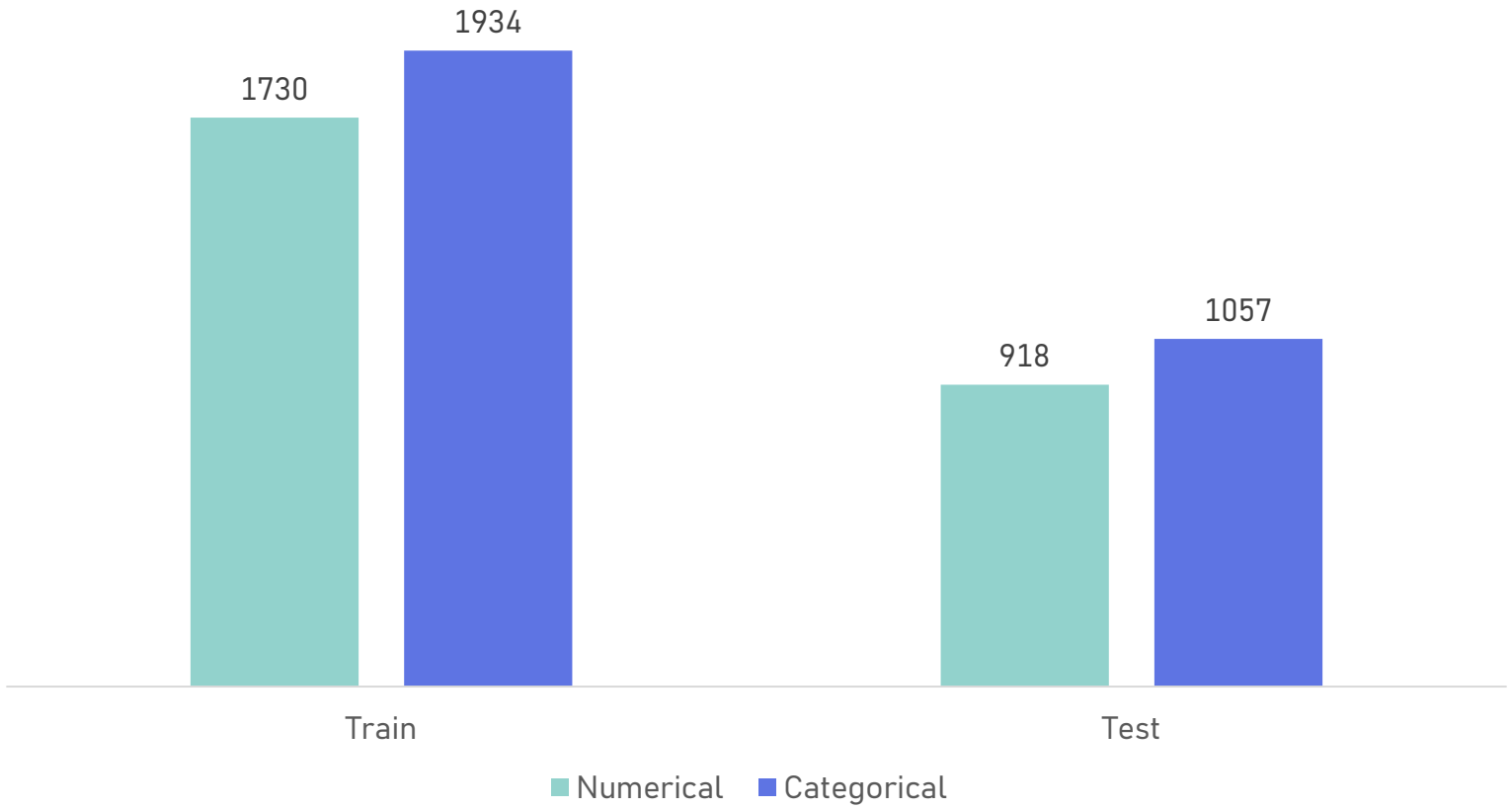
Model evaluation with AUC to perform predictions: probability of subscribe

DATA SUMMARY



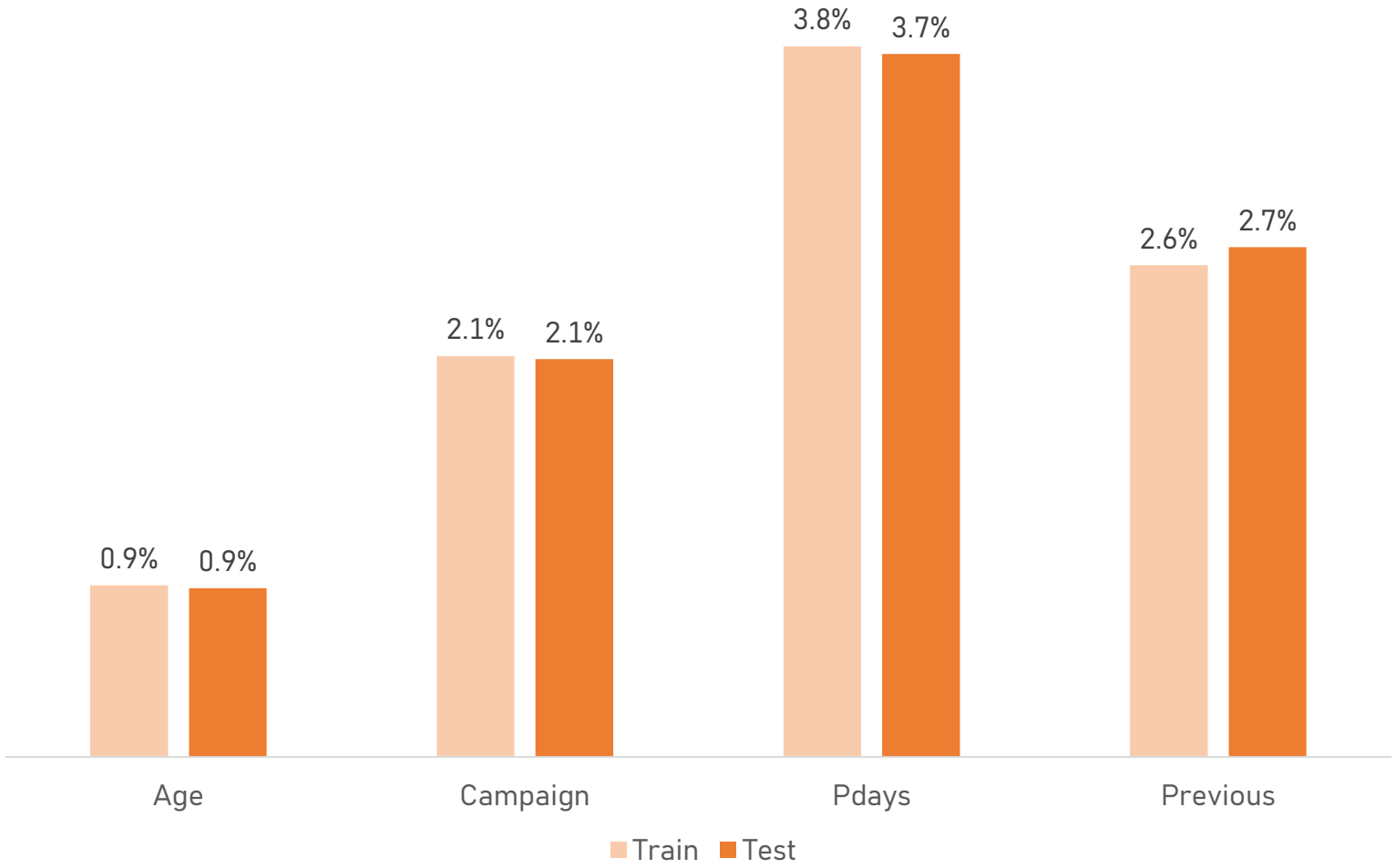
DATA SUMMARY

Missing Values by Variable Type



DATA SUMMARY

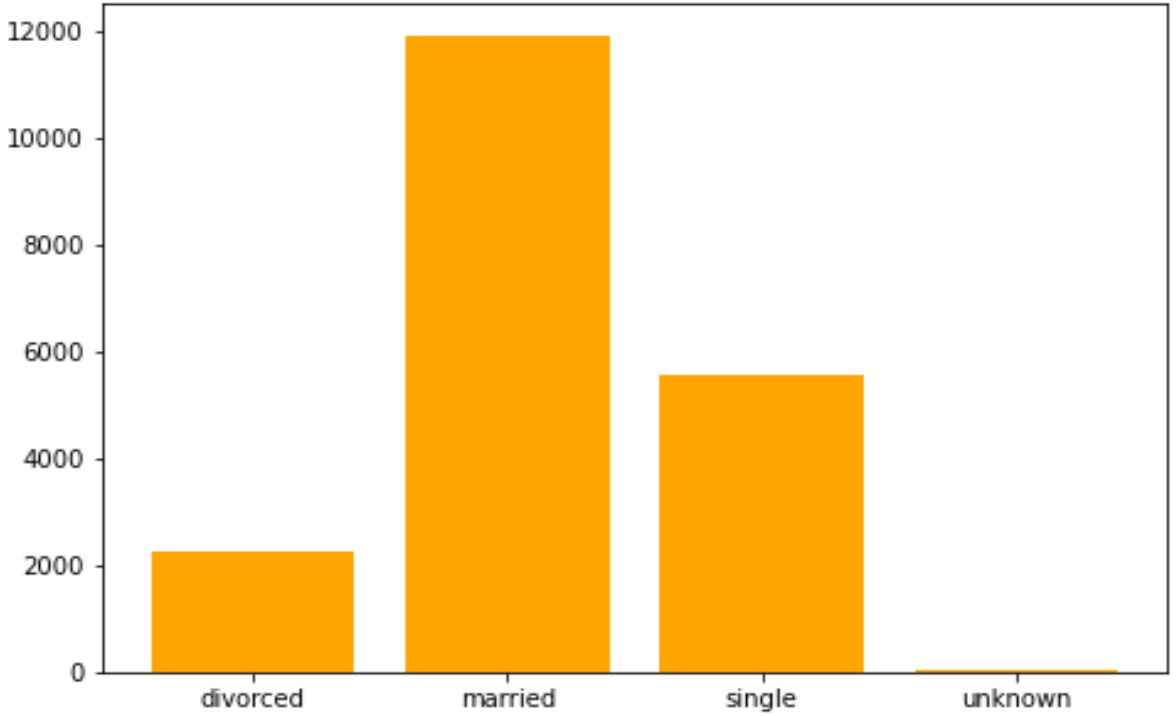
Outliers in Numerical Variables



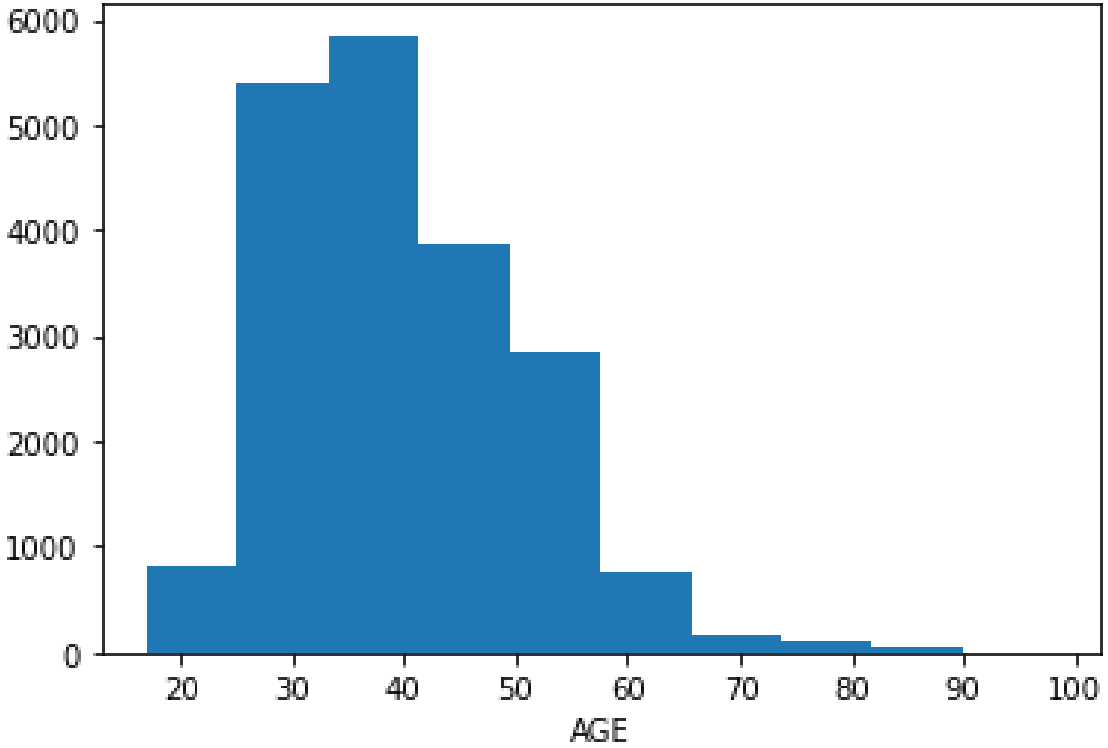
We decide not to remove outliers since they are similar on both sets

DATA SUMMARY

Marital Status Distribution

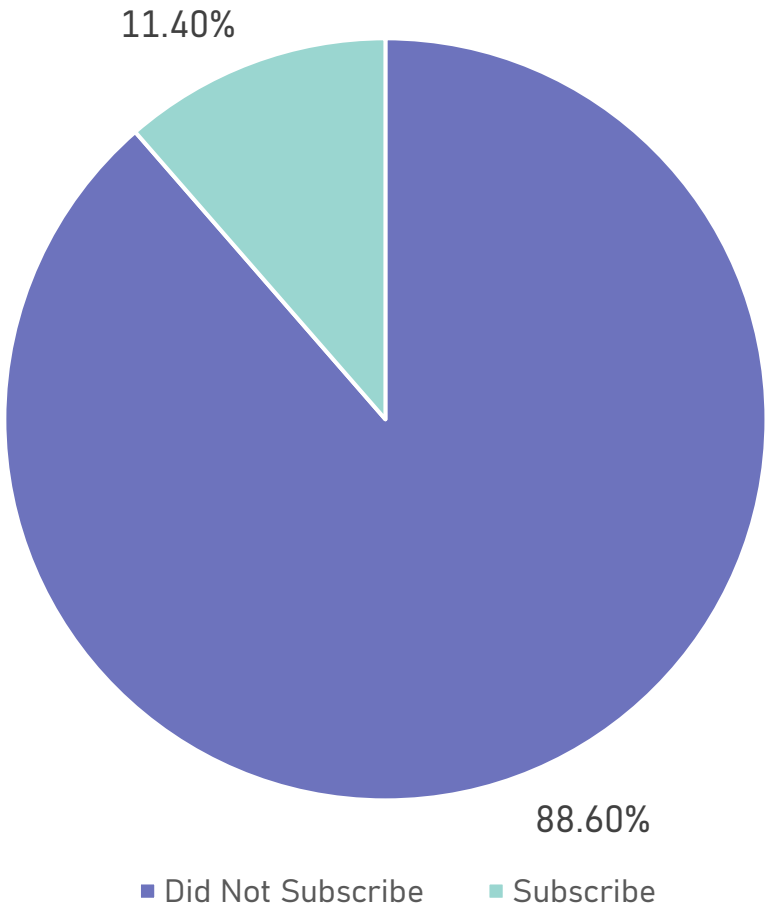


Age Distribution



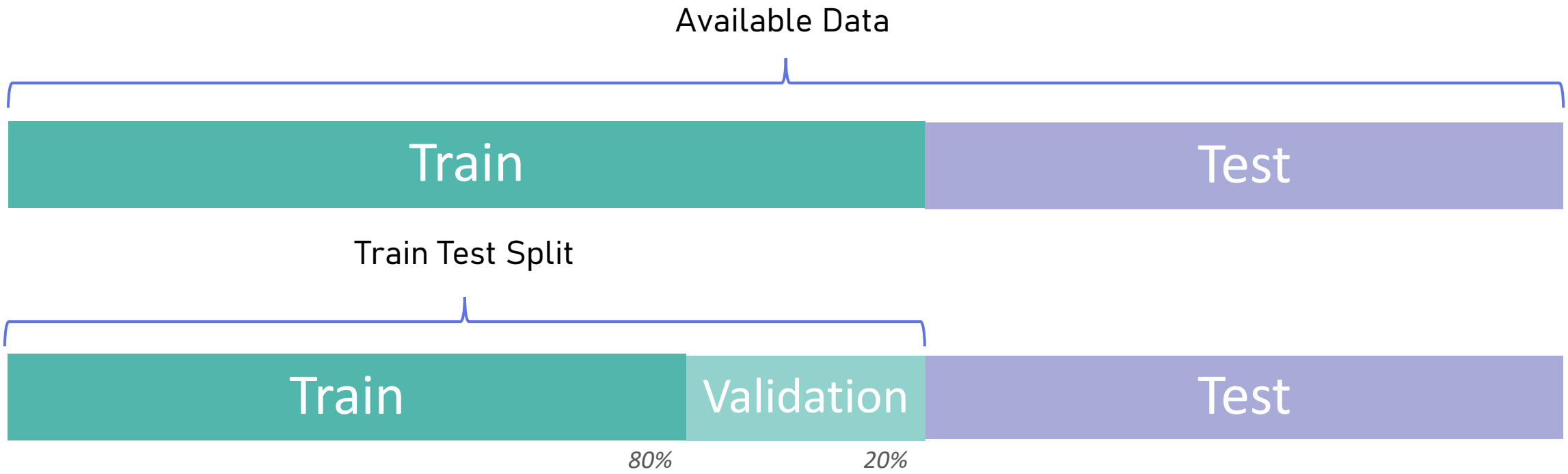
DATA SUMMARY

Target Variable Distribution

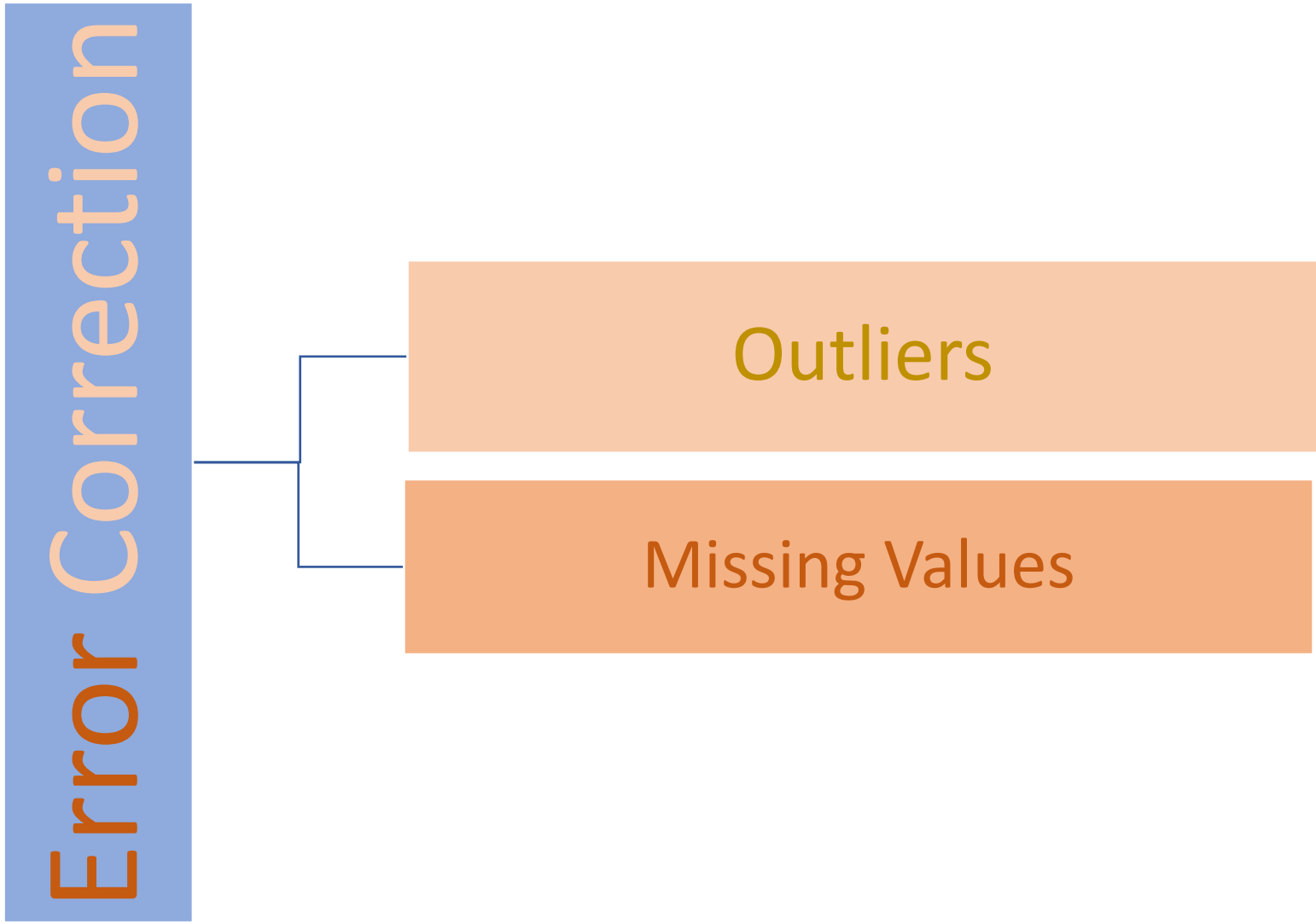


Target variable not equally distributed.

PRE-PROCESSING DATA



PRE-PROCESSING DATA



PRE-PROCESSING DATA

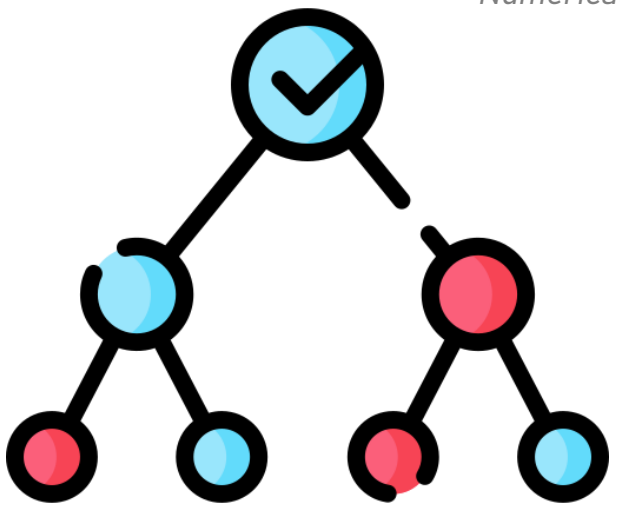
Ordinal Encoding for Categoricals

Education
basic.4y
basic.6y
basic.9y
high.school
illiterate
professional.course
university.degree
unknown

Education
1
2
3
4
5
6
7
8

Decision Tree-Based Remapping

Numericals and Categoricals



FEATURE ENGINEERING

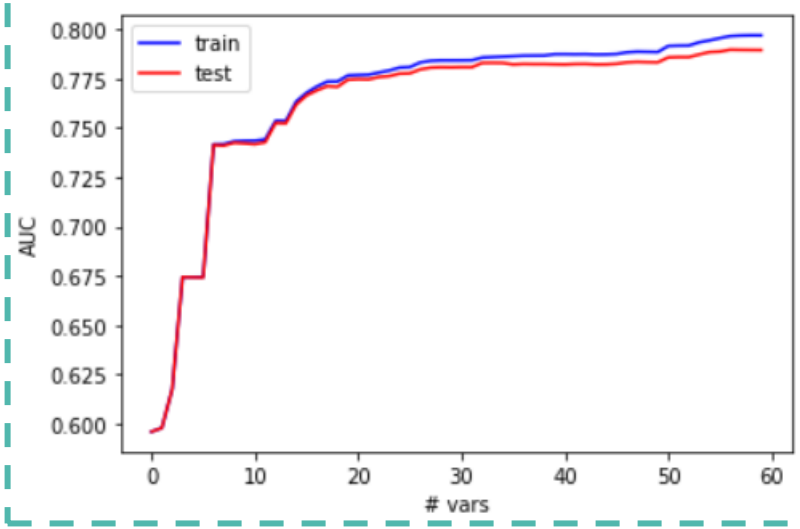
Since all variables are anonymized, we can't create features based on meaning:

- Age
- Cons.conf.idx

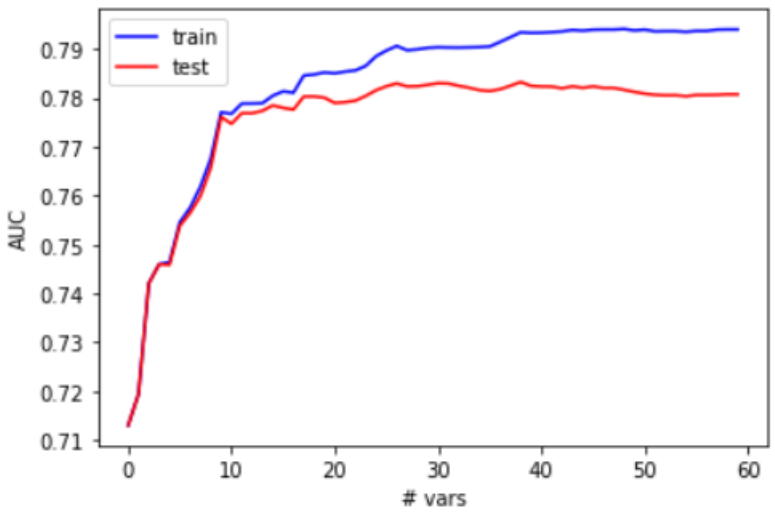
VARIABLE SELECTION

With Fisher Score Methodology, out of 194 Variables we select the best 55:

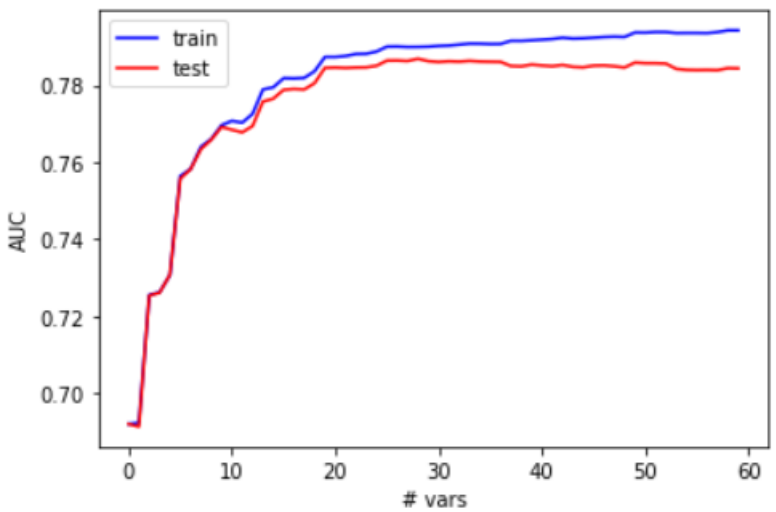
Decision-tree based re-mapping



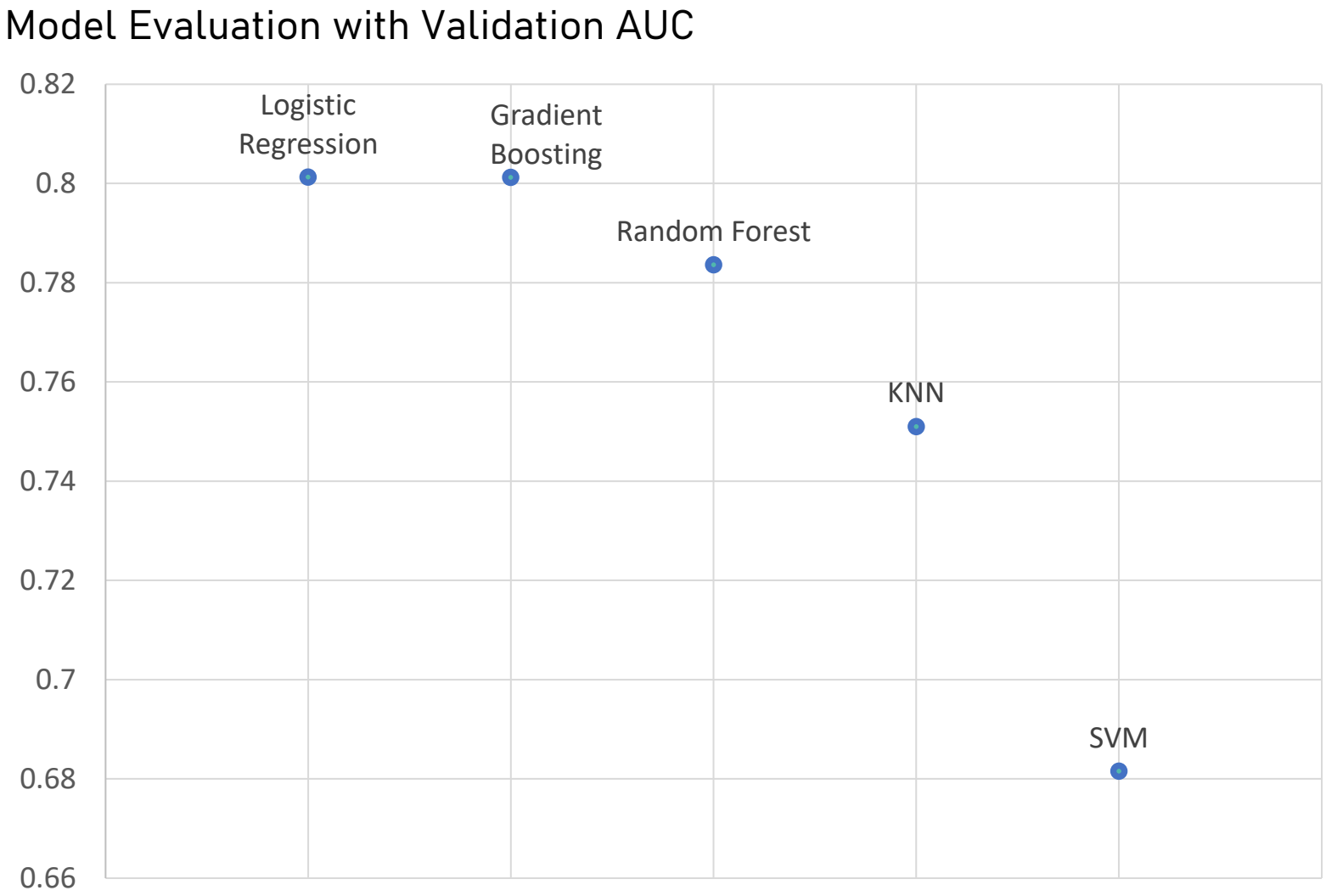
Equal Frequency Discretization



Equal Width Discretization



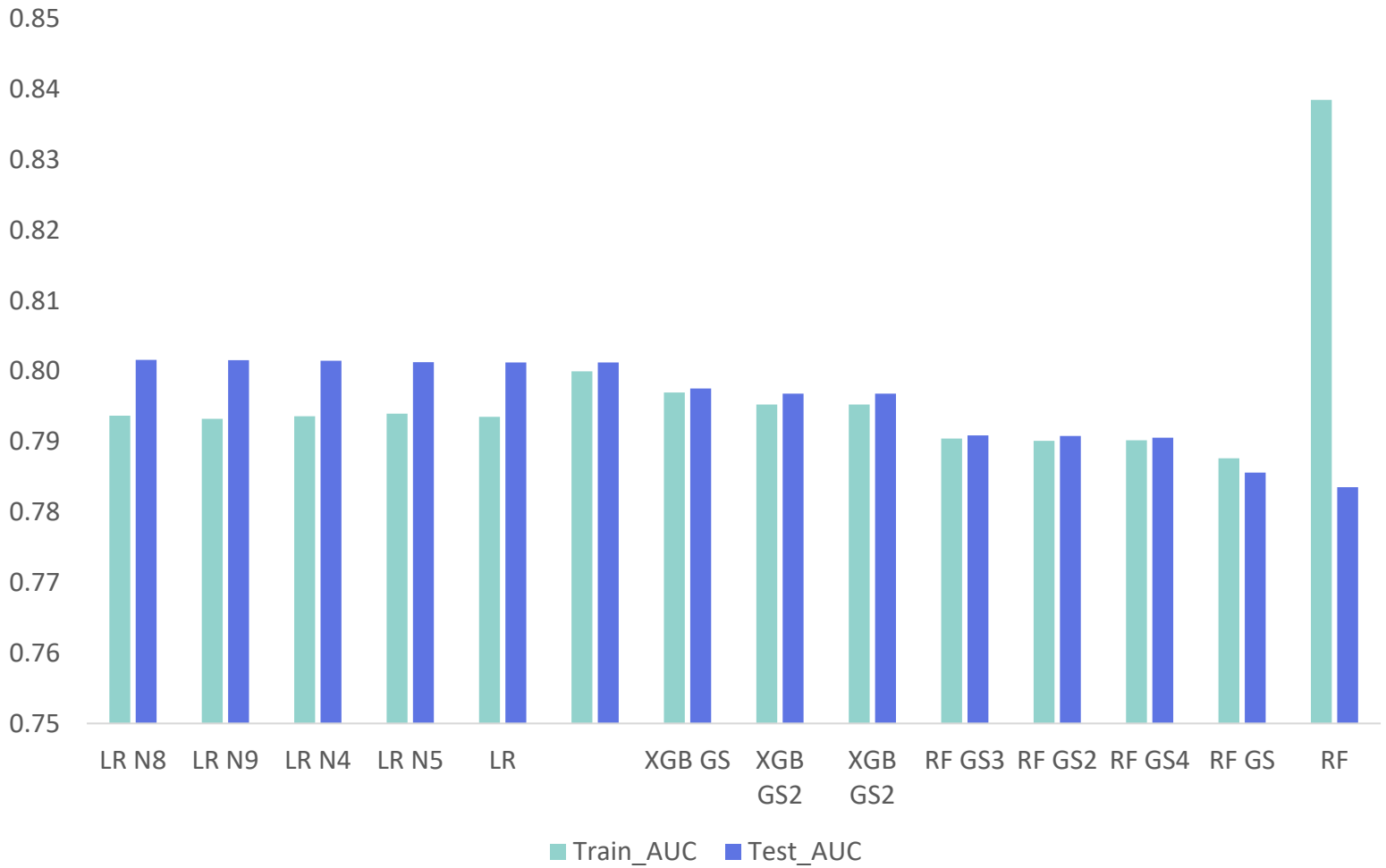
METHODOLOGY



We fit models to 5 different classification methods.

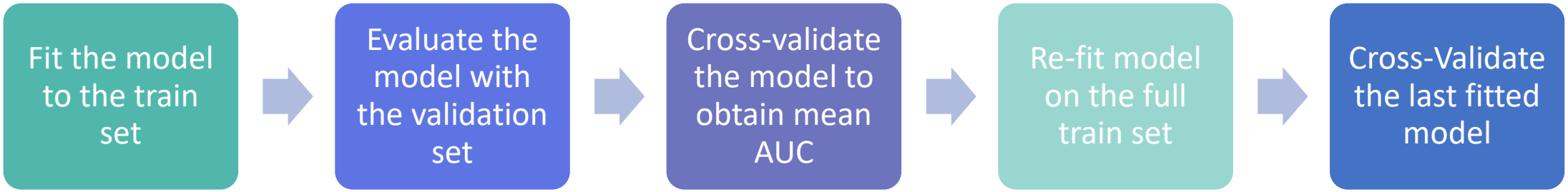
METHODOLOGY

Hyper Parameter Tuning



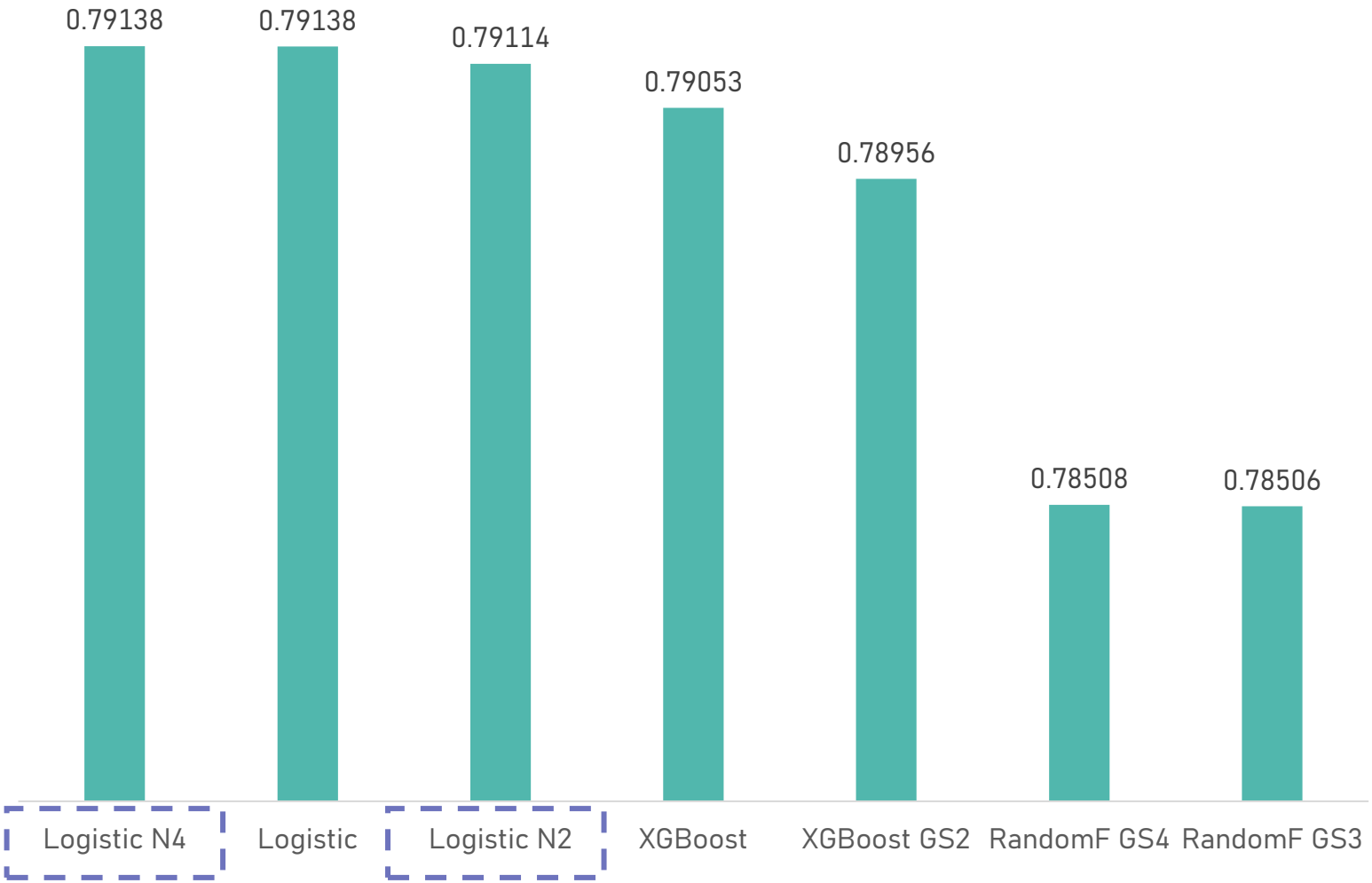
Logistic Regression is the best performer.

EXPERIMENTAL SETUP



RESULTS

Cross Validated AUC on Full Train Set



	Train_AUC	Test_AUC	CV_AUC	Full_Train_AUC	New_CV_AUC
Logistic N4	0.79360	0.80145	0.78858	0.79565	0.79138
Logistic	0.79350	0.80122	0.78854	0.79573	0.79138
Logistic N2	0.79403	0.79992	0.78811	0.79603	0.79114
XGBoost	0.79997	0.80121	0.78735	0.80044	0.79053
XGBoost GS2	0.79524	0.79680	0.78799	0.79581	0.78956
RandomF GS4	0.79021	0.79055	0.78377	0.79028	0.78508
RandomF GS3	0.79044	0.79088	0.78370	0.79055	0.78506

Our model can accurately predict with 80% accuracy whether a client will subscribe for a deposit after a bank telemarketing campaign.

Team 4 fromage pizza thanks you for listening!