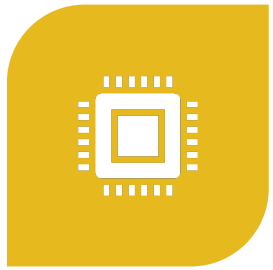
Predicting Life Satisfaction

Team Cross Validated

Presenters:
Xinda Li x556li@uwaterloo.ca
Ting Wei Zhang tw4zhang@uwaterloo.ca

Agenda

DATA



PREPROCESSING

MODEL
(STACKING)

RESULT

Data

- 271 variables
  - 8 explicitly categorical (Country, Citizenship, etc.)
  - ~14 implicit categorical (Employment relation, source of income, etc.)
  - The rest are ordinal or numeric

Preprocessing

- Feature engineering
  - Happiness ratio: happiness/average happiness by country
- One-hot encoding all categorical variables (including implicit ones)
- Normalization
- Imputing missing values
  - Consider all .a, .b, .c as missing values
  - Using min value for LR, RF and CatBoost
  - Special missing value for XGBoost
- Selecting top 350 variables using Extremely Random Tree

XGBoost

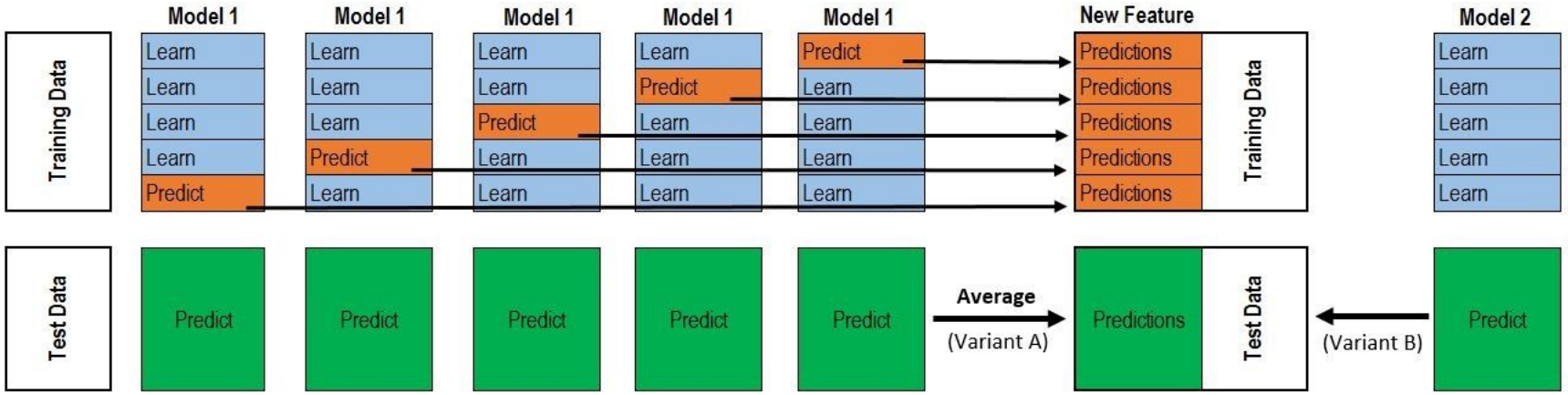
- Gradient Boosting algorithm developed in 2014 by Tianqi Chen
- Winning multiple Kaggle competitions
- 2^{nd} order gradient
- Extra regularization parameters
- Not fast (despite of the optimization and parallelization)

CatBoost

- Another gradient boosting algorithm
- Categorical features support
- Less parameters to tune
- Faster than XGBoost

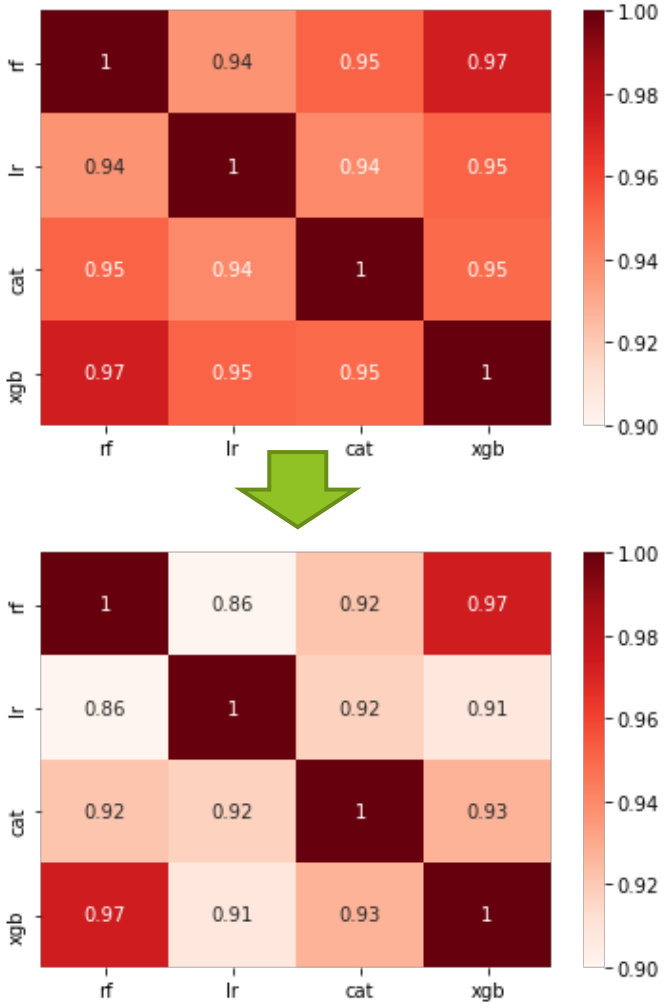
Stacking

- Level 1: XGBoost, CatBoost, Logistic Regression, Random Forest
- Level 2: XGBoost



Tuning

- Split train/test/online data
  - 30% of labeled data as test set
  - Test and online sets are approximately same size
- Tuning individual learners, then stack them together ☹
  - Highly correlated models
- Tuning everything at same time ☺



Result

- Using the whole training set
- Final AUC 0.89173
- Ranked #3 on private leaderboard