Application Classification with DonorChoose.org

Yi Li Yu, Cora Ou, Mona Kan, Lauren Moore, Ilana Novakoski & Rikarnob Bhattacharyya

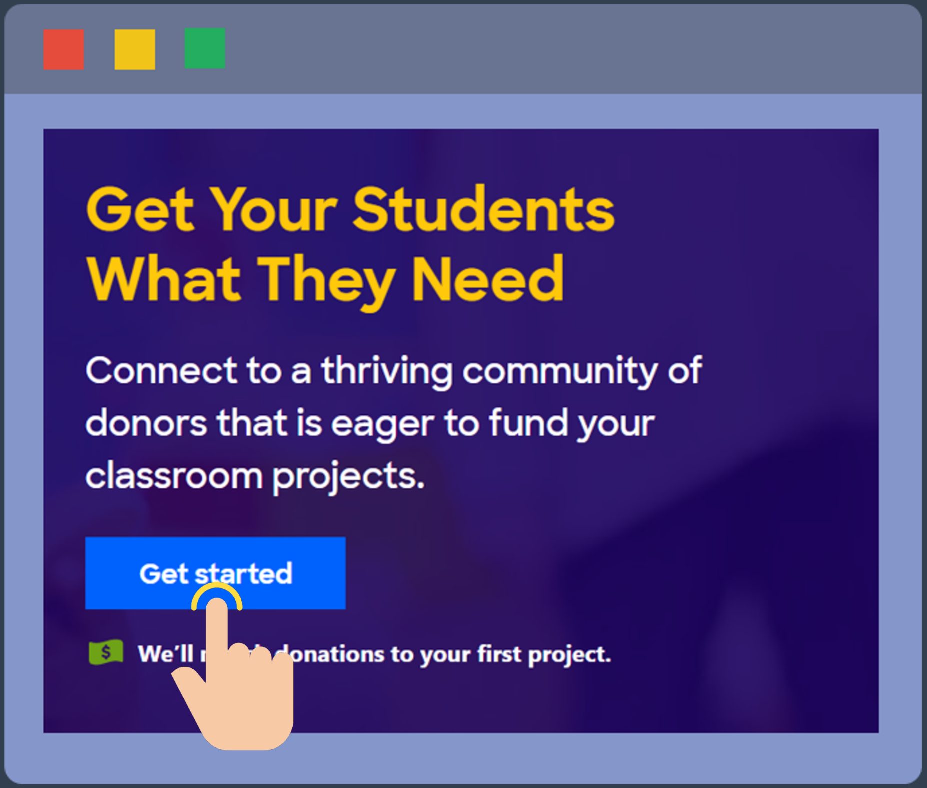
Outline

- Project Background
  - Motivation
  - Available Data
  - The Objective
- Analysis
  - Data Preparation
  - Model Methodology
  - Model Selection
- Results & Future Steps
  - Results
  - Recommended Deployment
  - Implementation
  - Caveats

Project Background

Motivation

DONORS CHOOSE



Get Your Students What They Need

Connect to a thriving community of donors that is eager to fund your classroom projects.

$ We'll r donations to your first project.

Motivation

- Expect to receive 500,000 project requests next year
- Each application costs $0.37 to review

$0.37 x 500,000 = $185,000

Available Data

- Teacher Prefix
- School state
- Date and time of submission
- Grade category
- Project categories & subcategories
- Project title
- Project essays
- Resources requested
- Number of previously approved projects
- Approval decision

The Objective

Reduce the number of volunteer hours required to screen all funding applications

Analysis

Data Preparation

- Separated into training and validation sets
- Converted categorical fields into numbers
- Extracted text features from essays
- Selected relevant features
- Normalized the data

Models Developed

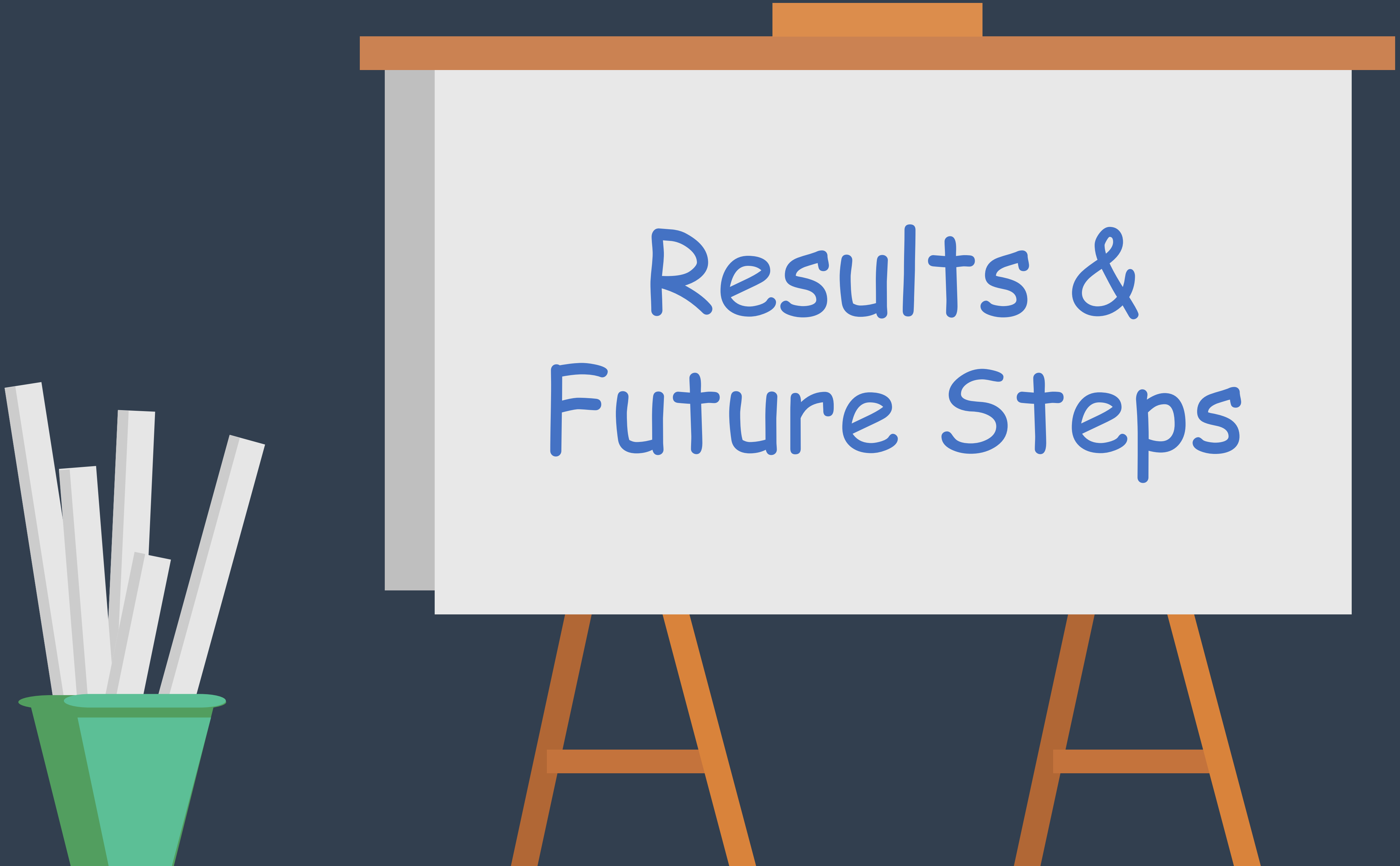
- Decision Tree
- K - Nearest Neighbor
- Logistic Regression
- Naive Bayes
- Support Vector Machine
- Bagging Classifier (with K-NN as estimator)
- Boosting Classifier (with decision tree as estimator)
- Light GBM
- XGBoost
- CatBoost
- Neural Network

Model Methodology & Selection

For each model:

- Tuned parameters with grid search/randomized search
- Selected the best-performing model based on Area Under the Curve (AUC)

Results & Future Steps



Results

Cost: $185,000

Saving: $129,500

Results

Best Model	AUC
Light GBM	0.735
CatBoost	0.727

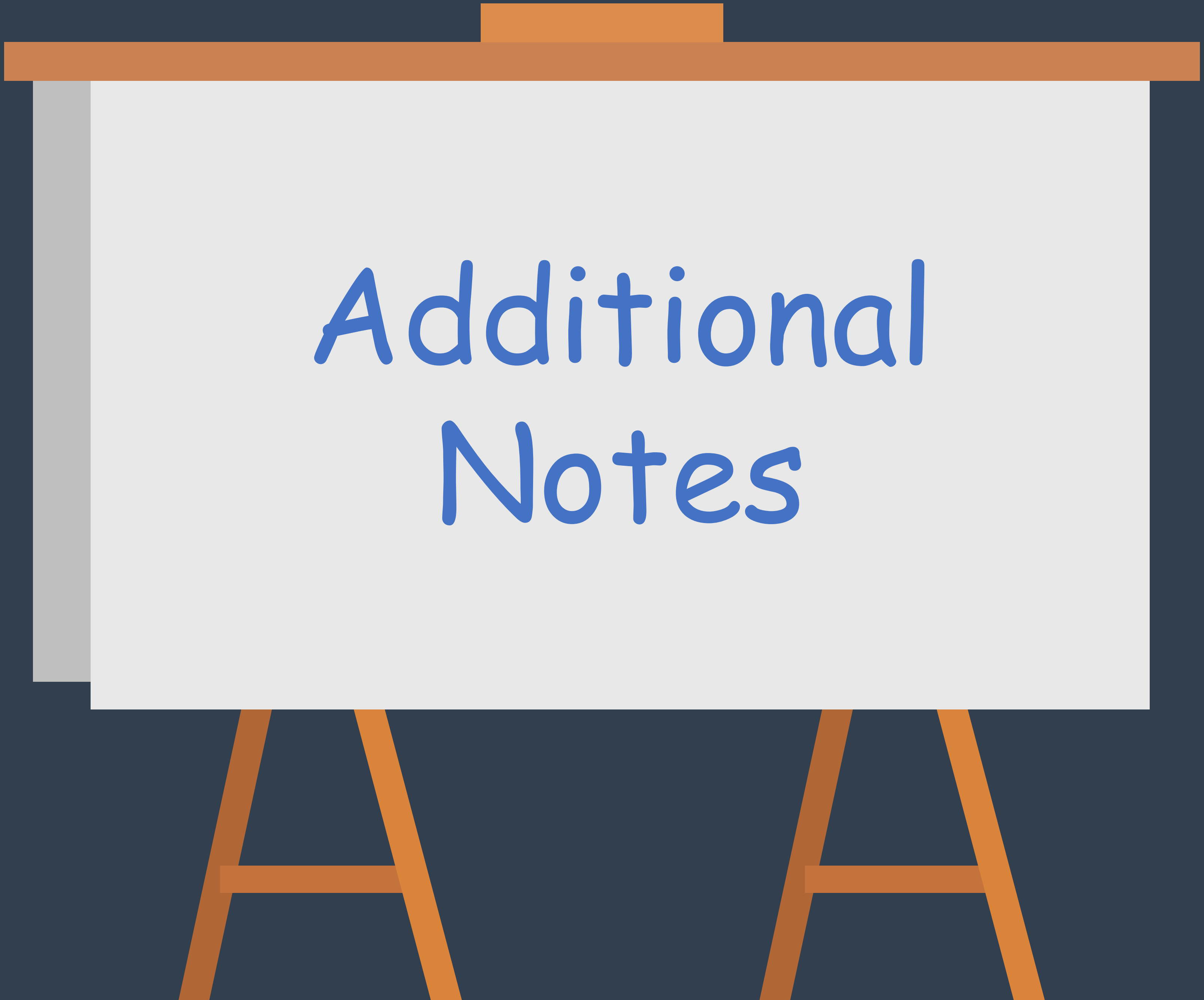
Recommended Deployment

- Build a pipeline to format incoming applications for the model
- Send samples to volunteers for further view
- Retrain the model annually, including the current year's application pool in the training data

Caveats

- Not all applications correctly classified
    - Employ a “report post” button for all projects hosted on the website
- Biased towards projects most similar to previously approved projects - novel requests will be penalized
    - Have human reviewers check rejected applications

Questions?



Additional Notes

Model Type	AUC
Decision Tree	0.654
K - Nearest Neighbor	0.639
Logistic Regression	0.70
Naive Bayes	0.57
Support Vector Machine	0.67
Bagging Classifier (with KNN as estimator)	0.61
Boosting Classifier (with Decision Tree as estimator)	0.62
Light GBM	0.735
XGBoost	0.723
CatBoost	0.727
Neural Network	0.714

Sources

[1] www.donorschoose.org

[2] www.kaggle.com/c/donorschoose-application-screening

[3] www.irisreading.com/what-is-the-average-reading-speed