Independent Study: Mount Holyoke Enrollment Predictions

Maha Mapara
1st May, 2020

Objectives of the Independent Study

- Evaluate various statistical techniques that best predict the enrollment yield
- What are the most important variables for predicting enrollment yield

Methodology

- Curated data set of 28 variables and 4514 admitted students
- Split into training and test set
- Training set has the enrollment years 2017, 2018
- Test set has the 2019 enrollment year
- 10 fold cross validation

Techniques Used

1. Logistic Regression
2. Decision Trees
3. Random Forest

Method 1: Logistic Regression

- 4 models: started with a model that had all 26 predictor variables and filtered down to a model where all variables were statistically significant
- These are the variables that were all statistically significant in the final model:

  Final app rating, campus visit day, all messages open, post application messages open, preview day event, composite SAT score, total grants awarded, total loans awarded, days to apply, impact zip track

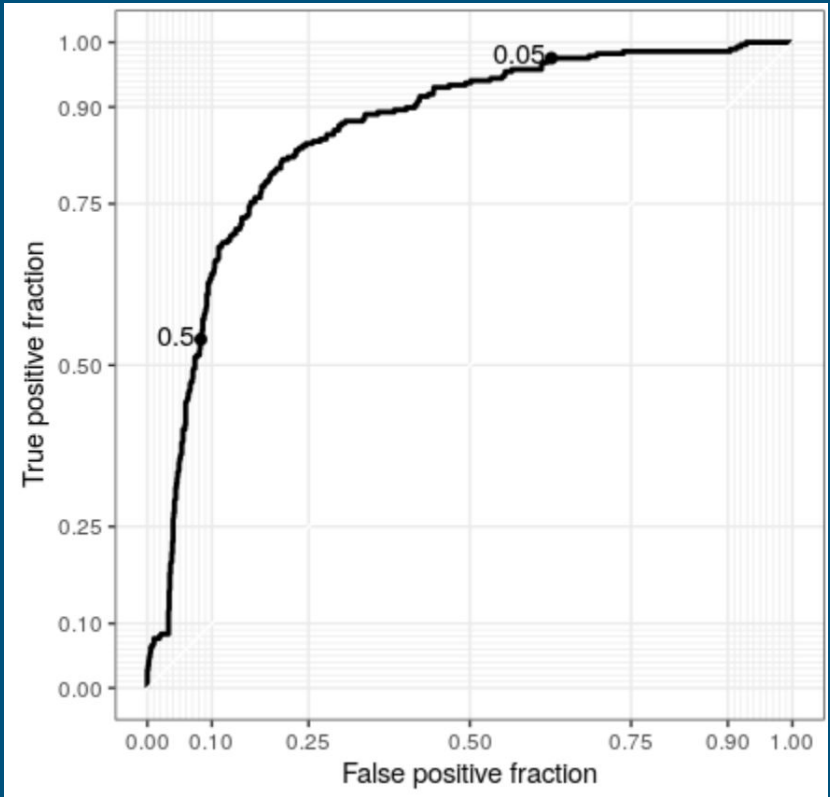
Model 2 had the lowest validation error rate of 16.27% and highest validation accuracy of 82.77%.

Training set accuracy was 83.88%

Test set accuracy: 83.15%

Test set error rate: 16.85%

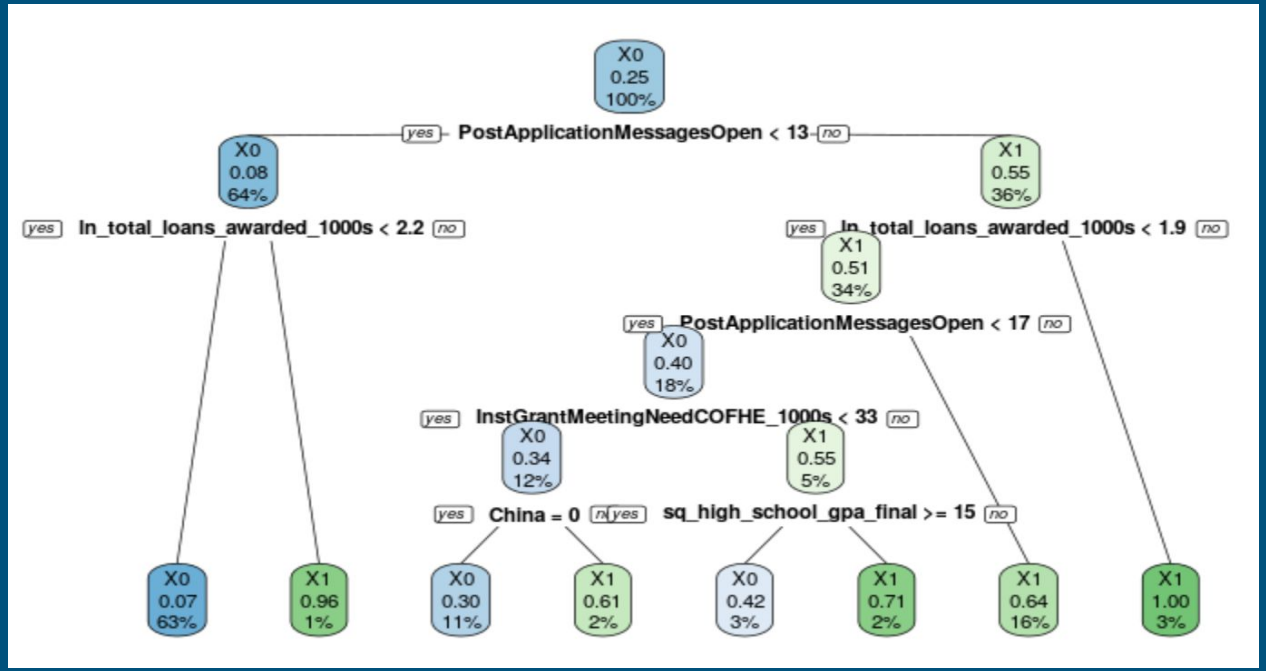
Note: Model 2 had 3 non statistically significant predictors out of the 15 in the model.

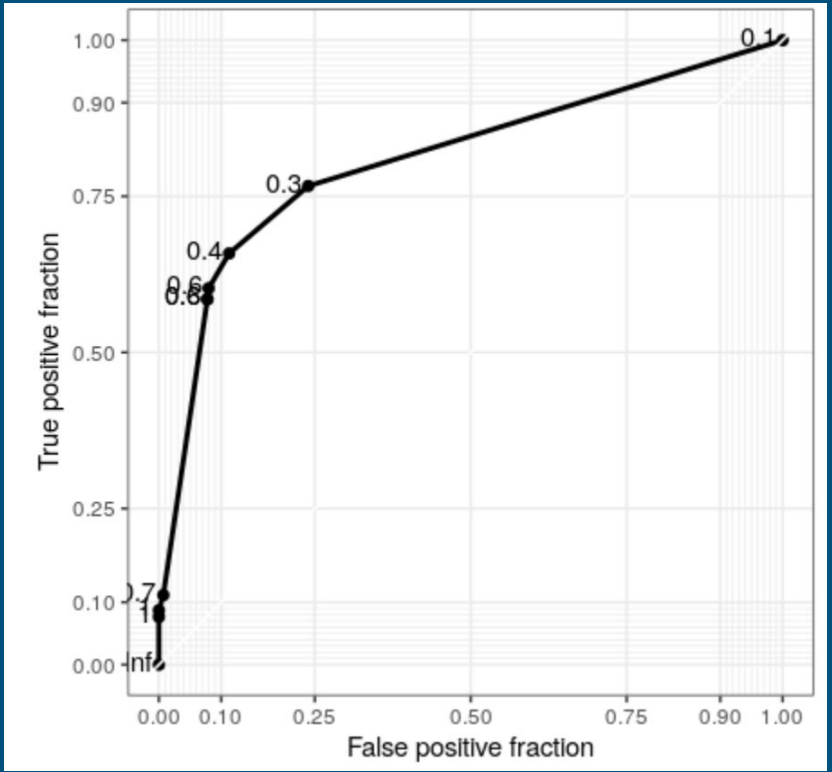

1.00
0.90
0.75
0.50
0.25
0.10
0.00
True positive fraction
0.05
0.5
0.00
0.10
0.25
0.50
0.75
0.90
1.00
False positive fraction

AUC = 0.859

Method 2: Decision Trees

Model with all 26 predictors has accuracy of 84.81%

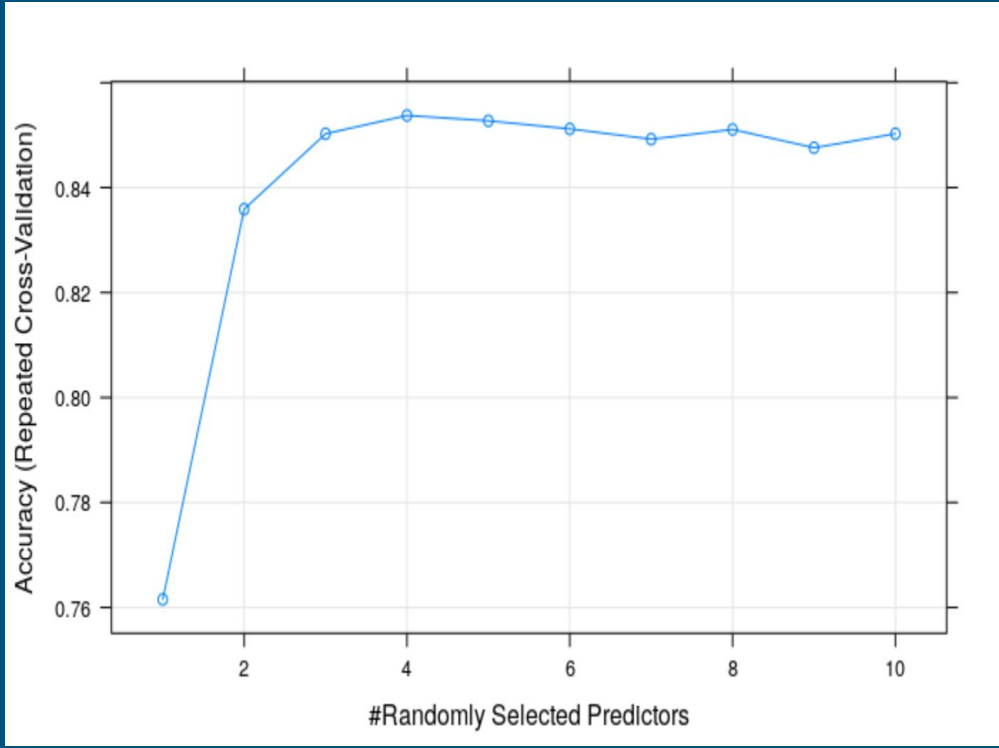


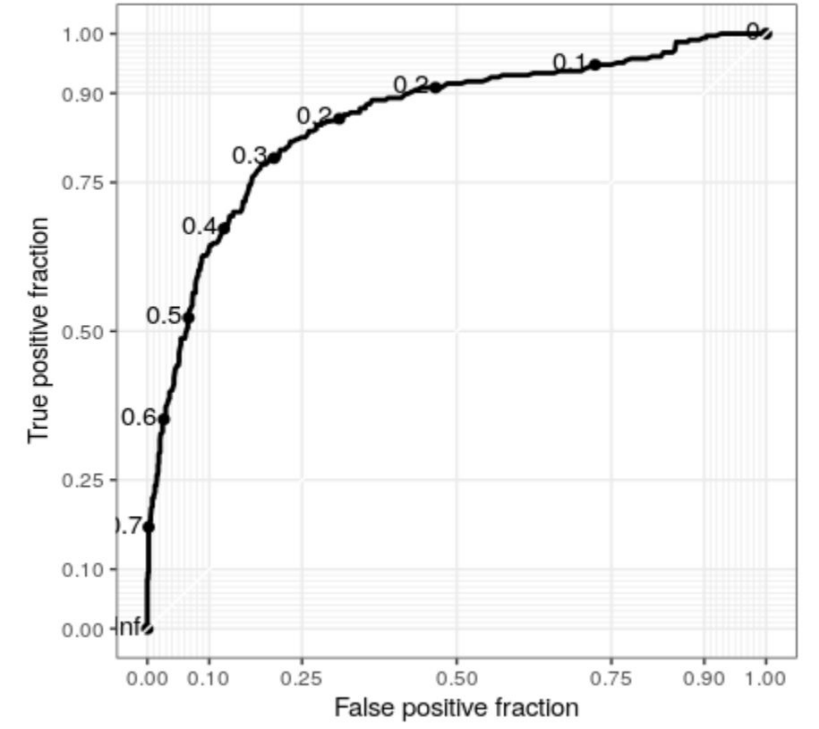

1.00
0.90
0.75
0.50
0.25
0.10
0.00
True positive fraction
0.1
0.3
0.4
0.00
0.10
0.25
0.50
0.75
0.90
1.00
False positive fraction

AUC = 0.809

Method 3: Random Forest

Mtry of 4 = gave accuracy of 85.45%





AUC = 0.854

Summary of Results

Test set classification accuracy:

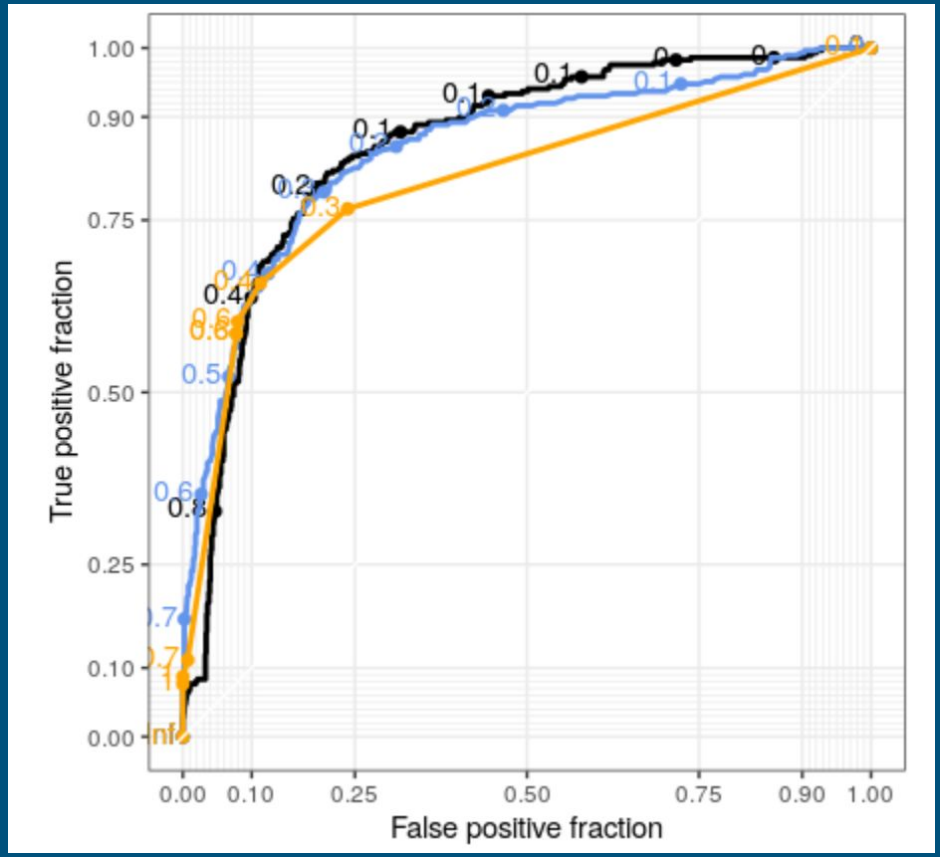
-Logistic regression: 83.15%

-Decision tree: 84.81%

-Random forest: 85.45%

Random forest performed the best

out of all the techniques



Decision Tree Random Forest Logistic Reg

AUC:	0.809	0.854	0.859

Future Plans

- Tune the random forest model parameters for higher classification accuracy
- Gradient tree boosting