Supervised Machine Learning: Using Statistical Models to Predict College Completion Rates

By: Cassidy Cubra

UNITED STATES AIR FORCE ACADEMY

Overview

- Background
- Models
- Results
- Future Work/Improvements
- Sources

Background

- Purpose: start from scratch – improve on a predictive model I create
- US Dept. Education College Scorecard Data – institution level

- Predict 4 year college completion rate
  - Subset:
    - 4 year colleges
    - Main Campus
    - Primarily bachelors degree granting
    - Not online only
    - Currently Operating

UNITED STATES AIR FORCE ACADEMY

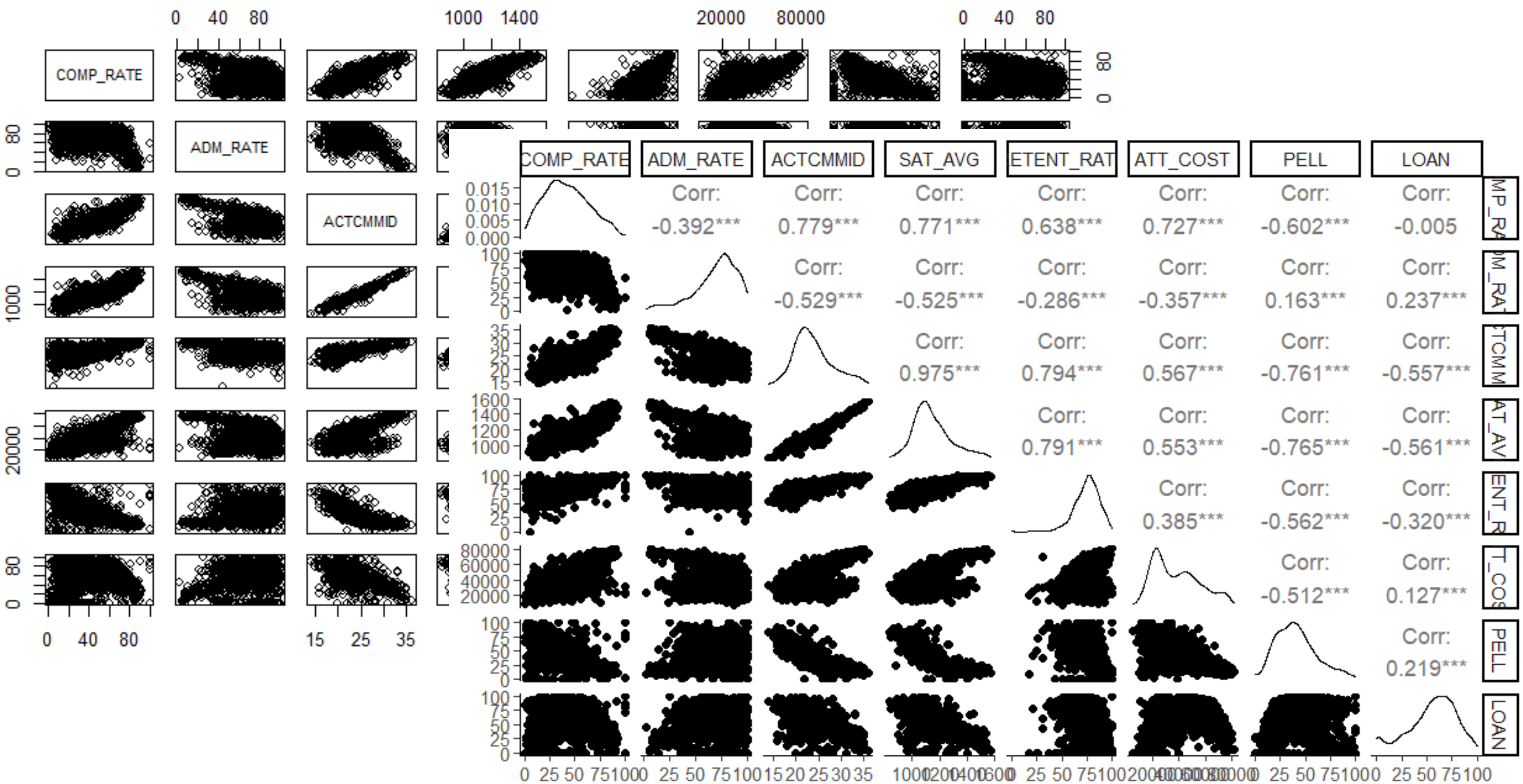
Data

- 6681 observations with 2989 variables
- Narrowed it down (1049 obs. of 8 variables):
  - **COMP_RATE** - Completion rate for first-time, full-time students at four-year institutions (100% of expected time to completion), pooled for rolling averages
  - **ADM_RATE** - Admission rate
  - **ACTCMMID** - Midpoint of the ACT cumulative score
  - **SAT_AVG** - Average SAT equivalent score of students admitted
  - **RENTENT_RATE** - First-time, full-time student retention rate at four-year institutions
  - **ATT_COST** - Average cost of attendance (academic year institutions)
  - **PELL** - Percentage of full-time, first-time degree/certificate-seeking undergraduate students awarded a Pell Grant
  - **LOAN** - Percentage of full-time, first-time degree/certificate-seeking undergraduate students awarded a federal loan

Methodology

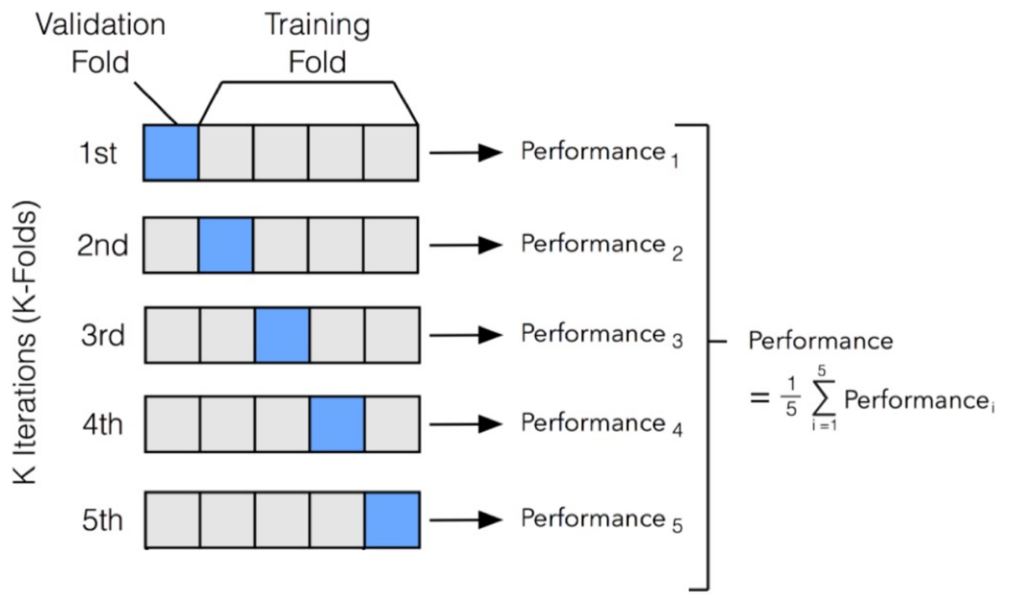
- Start with linear regression : see if there is improvement using different methods/models
- Root Mean Squared Error (RMSE): A metric that tells us how far apart the predicted values are from the observed values in a dataset, on average
- Adjusted R^2 : A metric that tells us the proportion of the variance in the response variable of a regression model that can be explained by the predictor variables - accounts for predictors that are not significant in a regression model

Relationships



Testing/Training & Cross Validation

- Split data into 70% testing, 30% training
- Tuning hyper parameters : K-fold cross validation
  - Penalty for Lasso & Ridge
  - Cost Complexity for trees



UNITED STATES
AIR FORCE
ACADEMY

Initial Models

```
Call:
stats::lm(formula = COMP_RATE ~ ., data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-60.781  -5.156   0.348   5.752  45.870 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -5.546e+01  7.730e+00  -7.175 1.37e-12 ***
ADM_RATE    -7.445e-02  1.813e-02  -4.107 4.33e-05 ***
ACTCMMID     4.002e-01  3.288e-01   1.217   0.2238    
SAT_AVG      2.296e-02  1.040e-02   2.208   0.0275 *  
RETENT_RATE  6.587e-01  4.620e-02  14.258  < 2e-16 ***
ATT_COST     4.305e-04  2.309e-05  18.644  < 2e-16 ***
PELL        -1.792e-01  2.873e-02  -6.236 6.50e-10 ***
LOAN         1.543e-01  2.009e-02   7.681 3.63e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.26 on 1041 degrees of freedom
Multiple R-squared:  0.7889,    Adjusted R-squared:  0.7875 
F-statistic: 555.8 on 7 and 1041 DF,  p-value: < 2.2e-16
```

Linear Model:

- Train on training data, and test on testing data
- RMSE: 9.6180
- Adj. R^2 : 0.7513

UNITED STATES
AIR FORCE
ACADEMY

Lasso & Ridge

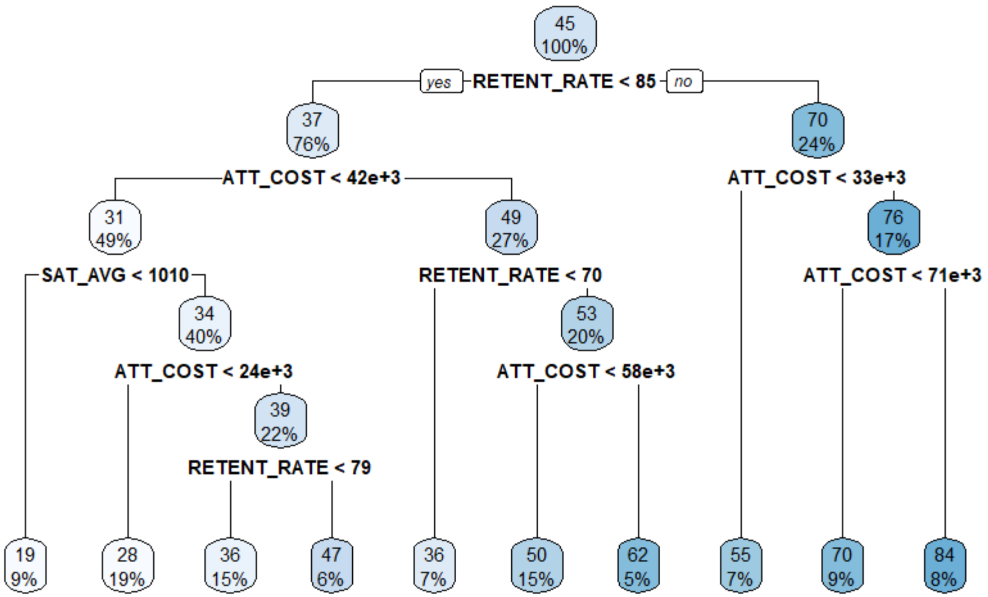
- Lasso
  - Uses shrinkage and variable selection to prevent overfitting and improve model interpretability
- Build the model and tune penalty to find the best RMSE and Adj. R^2
- Train the Lasso model on the training data, and test on testing data
- RMSE: 9.6286
- Adj. R^2 : 0.7505

- Ridge
  - Uses shrinkage to prevent overfitting by adding a penalty term to the cost function to shrink the magnitude of the coefficients
- Same process as Lasso
- RMSE: 9.6067
- Adj. R^2 : 0.7499

Regression Trees

Basic Decision Tree:
- Training and testing
- RMSE: 11.6017
- Adj. R^2 : 0.6478

Basic Decision Tree, Tuning Cost Complexity:
- RMSE: 10.9939
- Adj. R^2 : 0.6941



Regression Trees Cont.

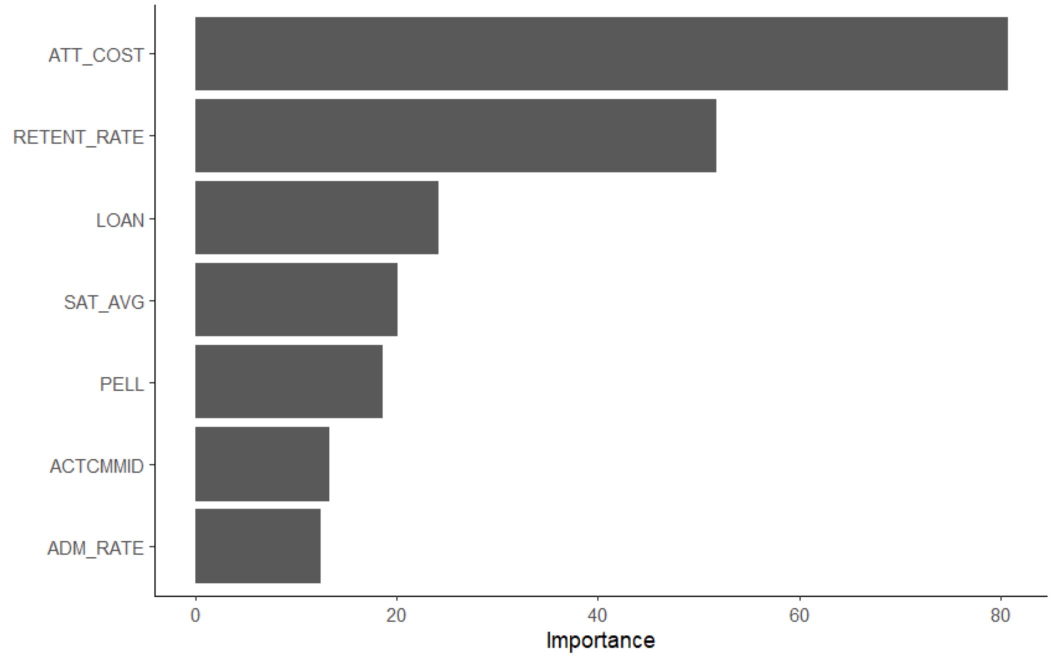
Ensemble Methods: Bagging & Boosting - decrease the variance of a single estimate as they combine several estimates from different models

Random Forest Bagging:

- Tree models learn from each other independently at same time, combine to find average
- RMSE: 9.4771
- Adj. R^2 : 0.7589

Random Forest Boosting:

- Trees learn sequentially and adapt from previous tree
- RMSE: 9.9935
- Adj. R^2 : 0.7375

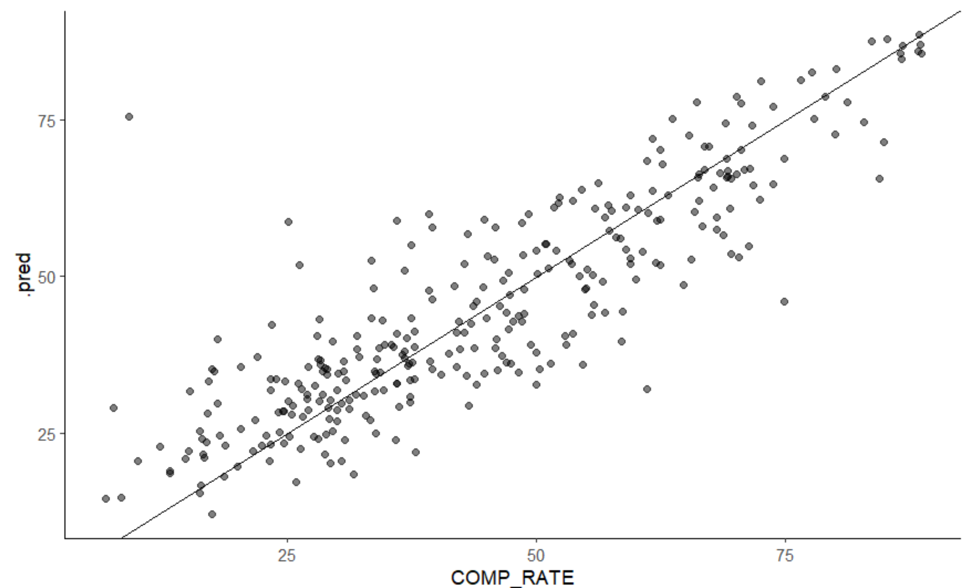


Results

Model	RMSE	Adj. R^2
Linear	9.6180	0.7513
Lasso	9.6286	0.7505
Ridge	9.6067	0.7499
Decision Tree	11.6017	0.6478
Decision Tree – tuned CC	10.9939	0.6941
Random Forest Bagging	9.4771	0.7589
Random Forest Boosting	9.9935	0.7375

Results

- Random Forest Bagging gave the best RMSE and Adj. R^2
  - Use this model to predict 4 year completion rate
- Variable of most importance: Cost of attendance



Future Work/Improvements

- Complex problem – hard to fit a regression model for prediction
  - Multiple predictors leads to high R^2
- Always better methods/data being discovered
- Removing/adding predictors: potential better model fit

Sources

- *Ansari, Faizan. "Cross-Validation Techniques." Analytics Vidhya, Medium, https://medium.com/analytics-vidhya/cross-validation-techniques-bacb582097bc.*
- *"College Scorecard Data." U.S. Department of Education, https://collegescorecard.ed.gov/data/.*
- *Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. "An Introduction to Statistical Learning : with Applications in R." New York :Springer, 2013.*

UNITED STATES
AIR FORCE ACADEMY

DATASCIENCE
AT
USAFA