July 24, 2023

The University of Texas at Austin
McCombs School of Business

Predicting Academic Performance

Models, Predictions, and Business Applications

Group 1:
STA380 - Introduction to Machine Learning

The University of Texas at Austin
McCombs School of Business

Group 1

Luis Villazon	lrv499
Pritesh Singh	ps35762
Mandeep S Burdak	msb4384
Teagan Milford	tm38292
Sardar Muhammad Ahmad Ali	sa58382
John Hwang	jh75576

Overview

Problem Statement & Data

The Modeling Process

Results & Conclusions

The Problem & the Data

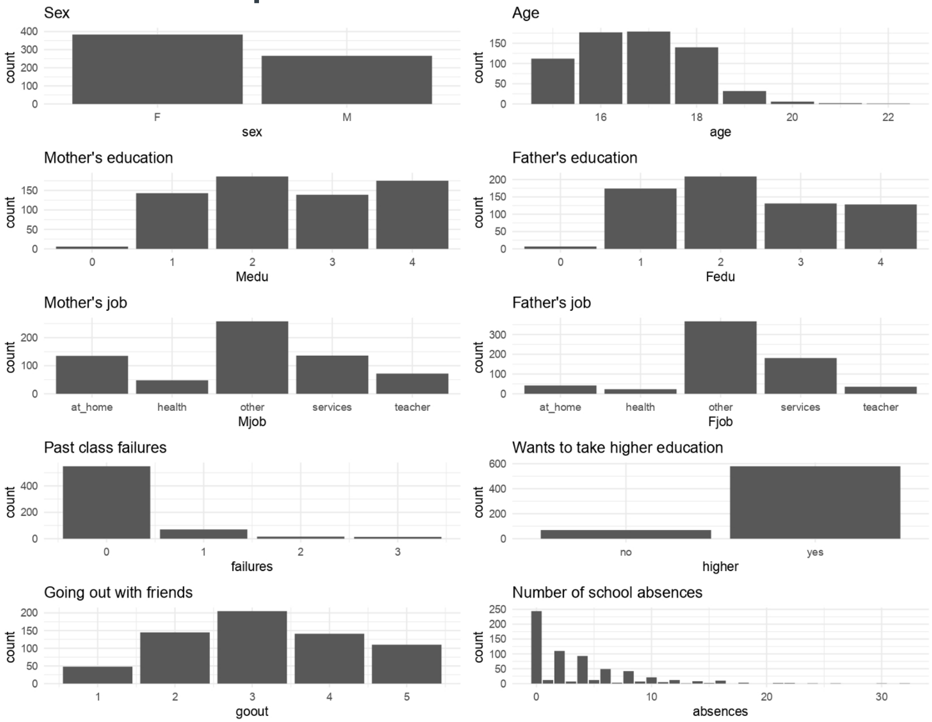
Unraveling the influence: Predicting Academic Performance

- We aim to predict high school students academic performance based on their social, demographic and behavioral characteristics.

Outline of the Data

- Originates from a survey of students math and portuguese language courses in secondary school in Portugal.
- Total: 649 students, 31 variables.
- No presence of missing values.
- Variable of interest: Overall Grade (from 0 to 20).

- Some predictors:



Process & Methodology

Trees > Regression

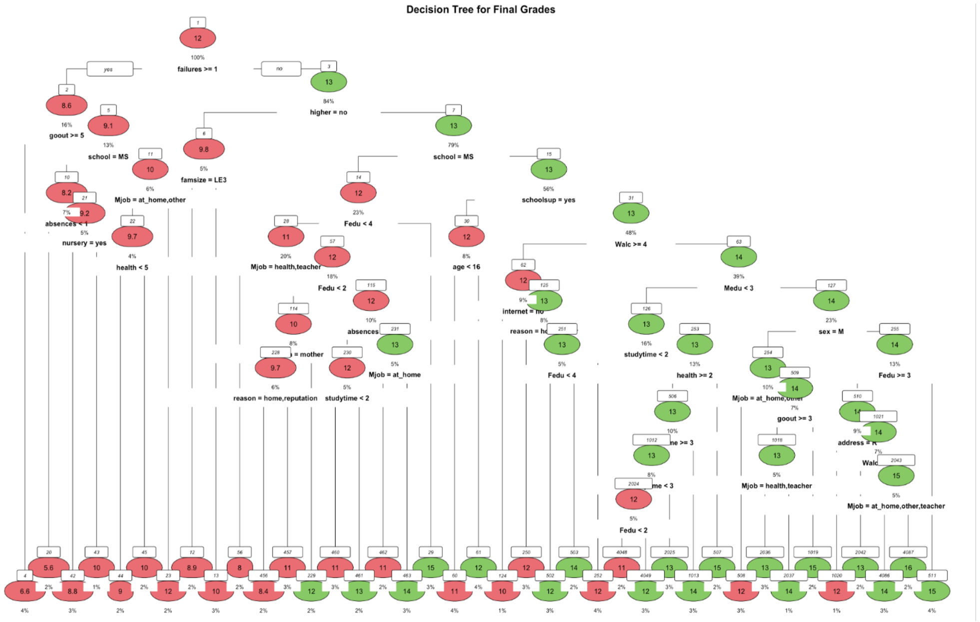
Algorithm	MSE
Tree	12.27

Regression:

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j,$$

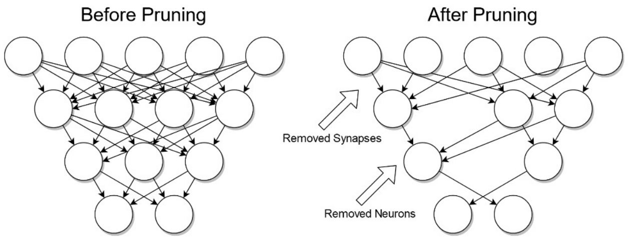
Trees:

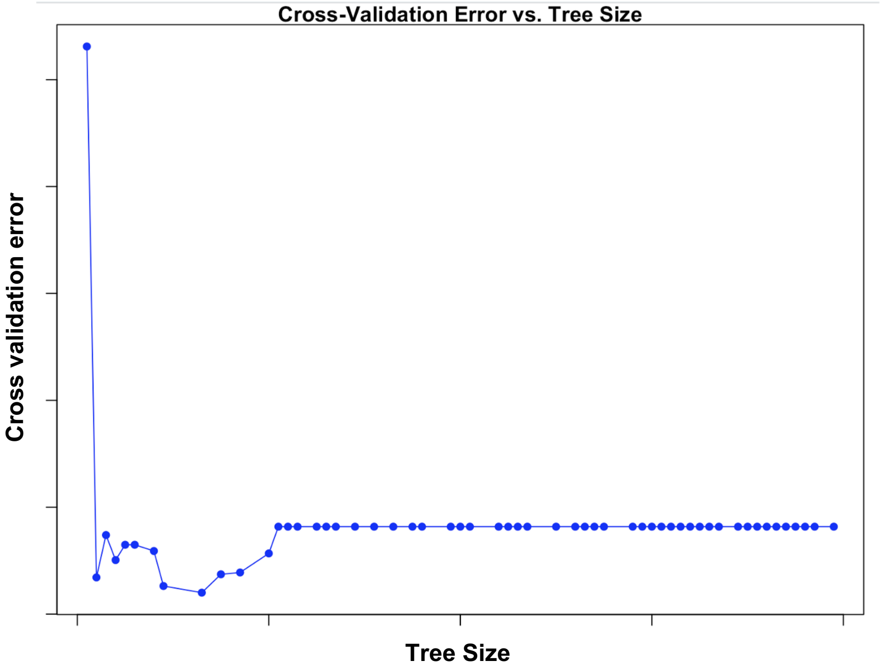
$$f(X) = \sum_{m=1}^{M} c_m \cdot 1_{(X \in R_m)}$$



Gardening 101: Pruning

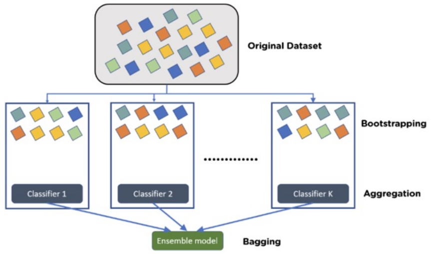
Algorithm	MSE
Tree	12.27
After Pruning	10.61

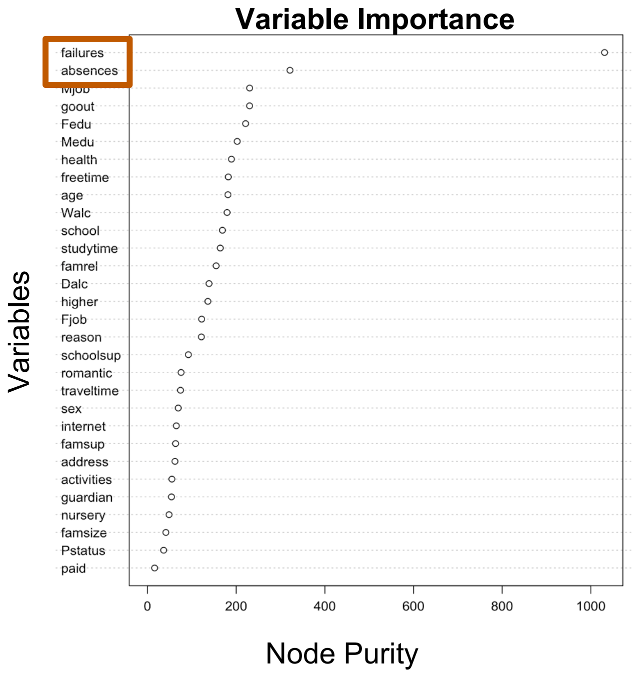




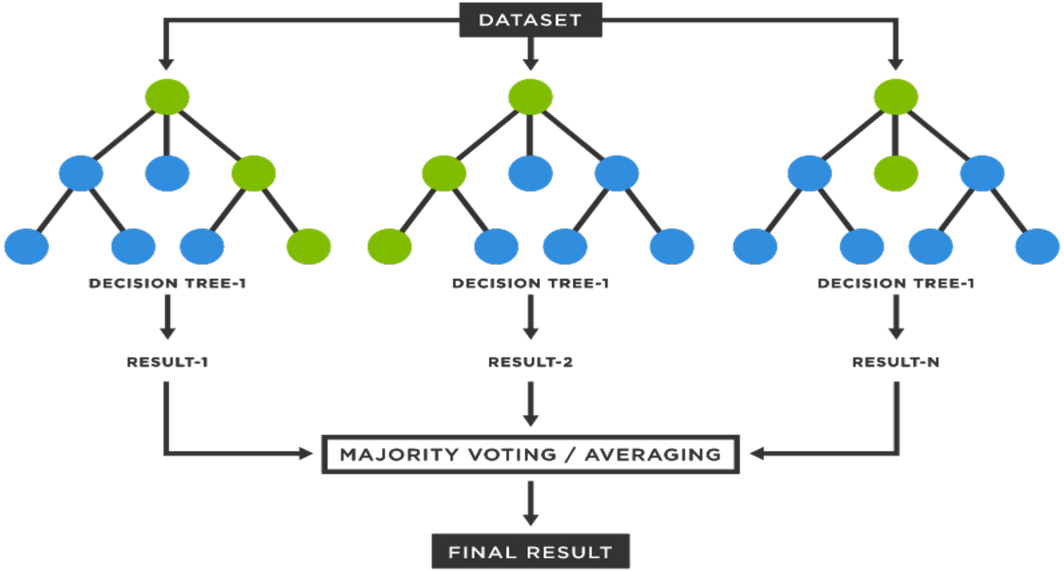
Bagging

Algorithm	MSE
Tree	12.27
After Pruning	10.61
Bagging	8.29



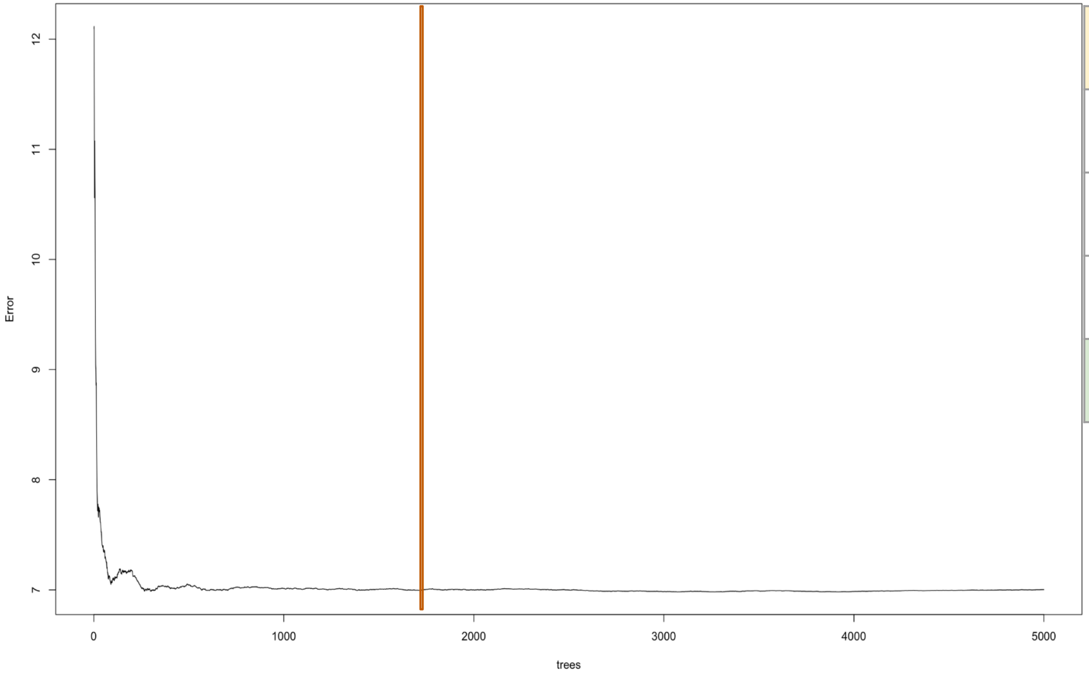


Random Forest



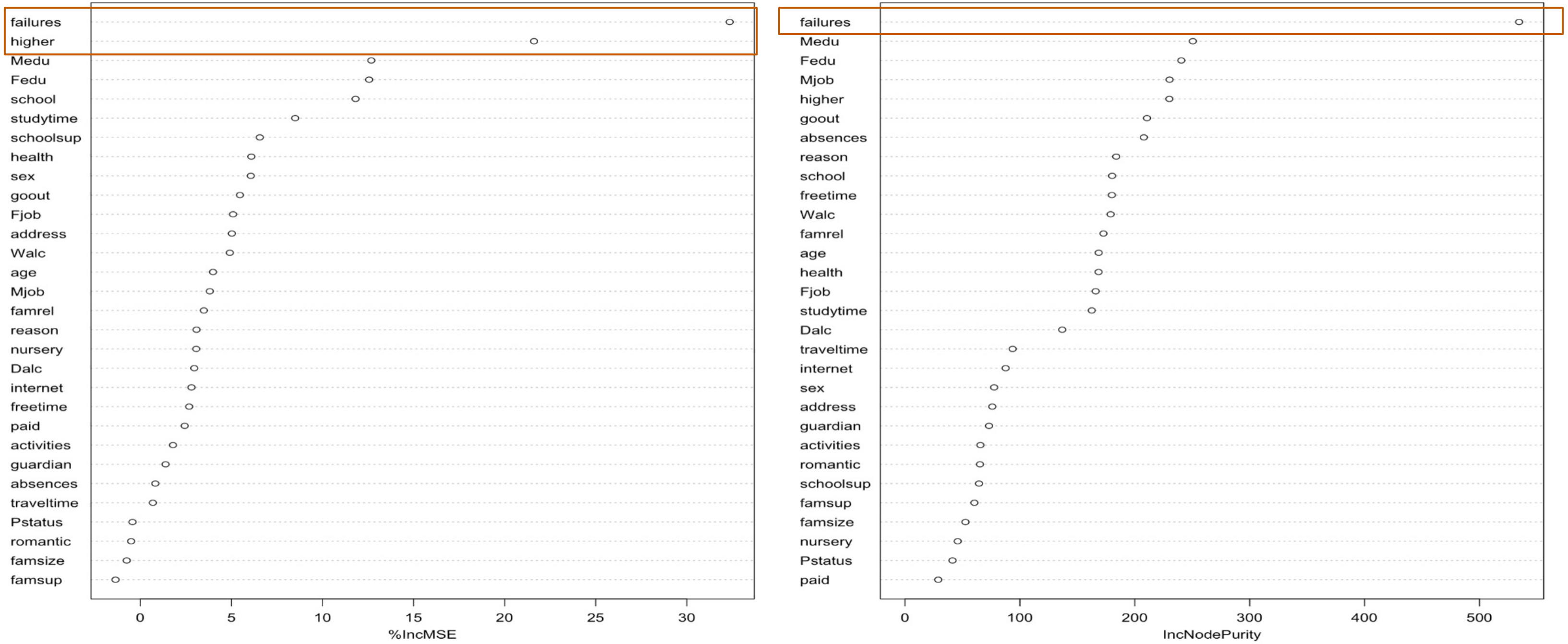
www.bigstock.com · 264484027

Random Forest

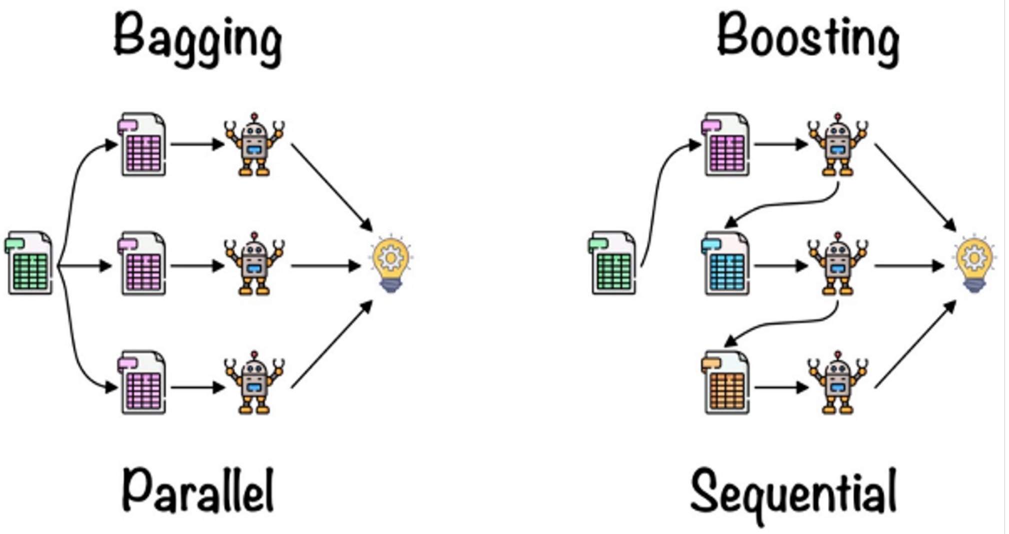


Algorithm	MSE
Tree	12.27
After Pruning	10.61
Bagging	8.29
Random Forest	7.86

Random Forest



Boosting

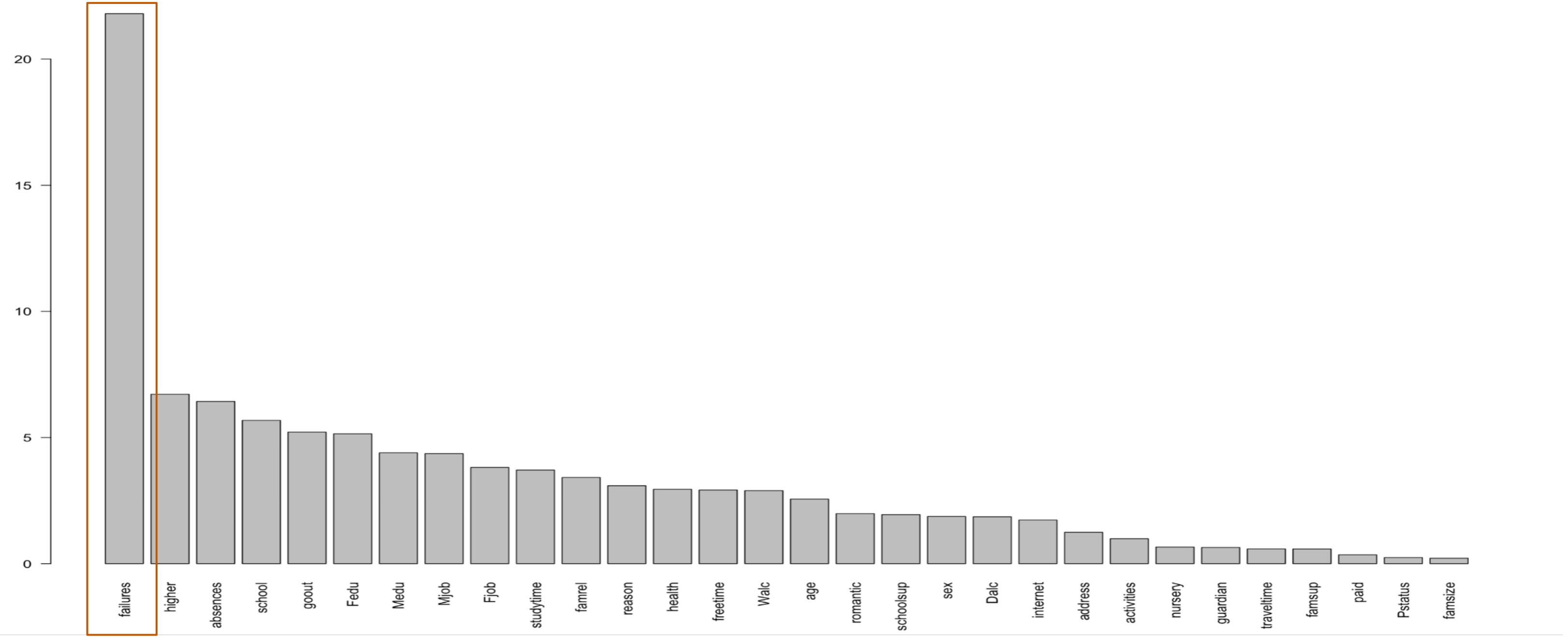


Boosting

	tdepth	ntree	lam	olb	ilb
1	4	1000	0.001	8.540	7.211
2	10	1000	0.001	8.309	6.319
3	4	5000	0.001	7.841	4.635
4	10	5000	0.001	7.967	2.999
5	4	1000	0.200	11.361	0.053
6	10	1000	0.200	10.822	0.000
7	4	5000	0.200	10.028	0.000
8	10	5000	0.200	10.911	0.000

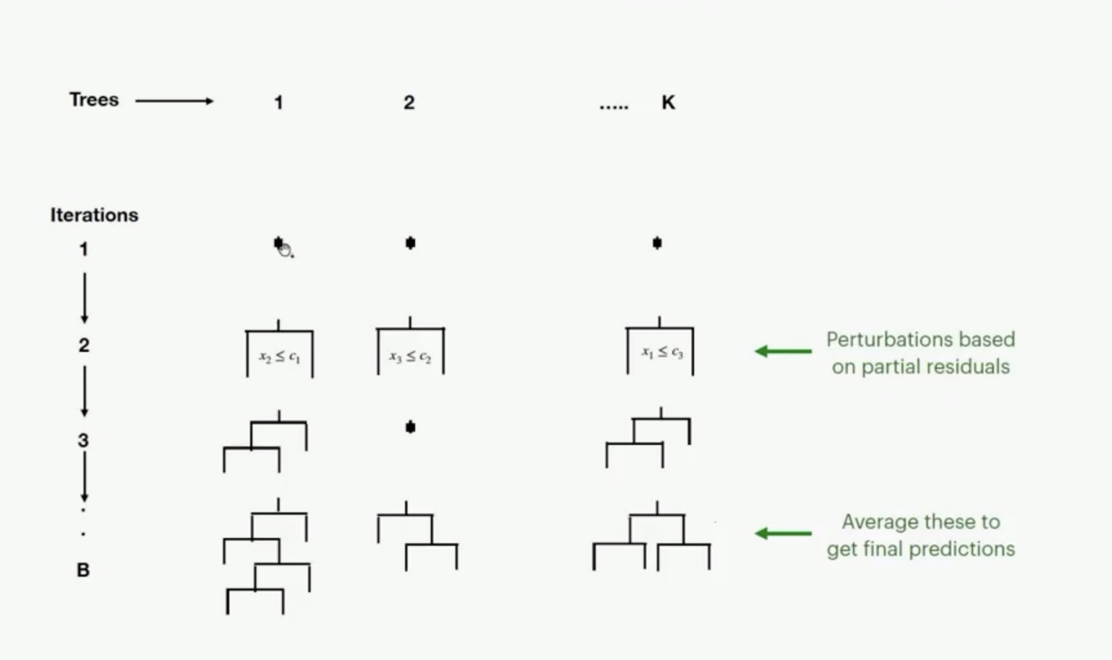
Algorithm	MSE
Tree	12.27
After Pruning	10.61
Bagging	8.29
Random Forest	7.86
Boosting	7.84

Boosting



20
15
10
5
0
failures
higher
absences
school
goout
Fedu
Medu
Mjob
Fjob
studytime
famrel
reason
health
freetime
Walc
age
romantic
schoolsup
sex
Dalc
internet
address
activities
nursery
guardian
traveltime
famsup
paid
Pstatus
famsize

BART (Bayesian Additive Regression Trees)

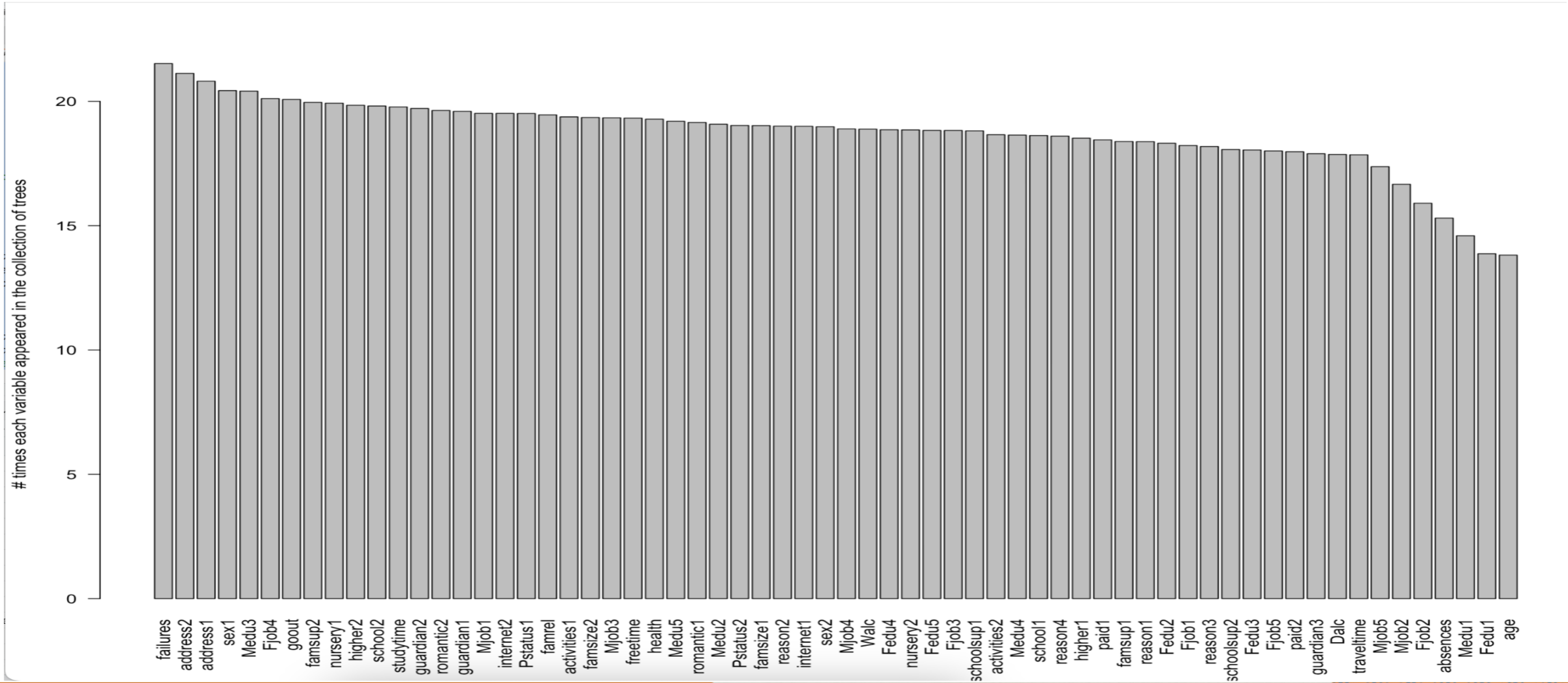


BART(Bayesian Additive Regression Trees)

	niterations	ntree	nburnin	olb
1	500	1000	100	8.129150
2	1000	1000	100	8.099505
3	500	2000	100	8.131952
4	1000	2000	100	8.108335
5	500	1000	200	8.147487
6	1000	1000	200	8.096032
7	500	2000	200	8.098814
8	1000	2000	200	8.131178

Algorithm	MSE
Tree	12.27
After Pruning	10.61
Bagging	8.29
Random Forest	7.86
Boosting	7.84
BART	8.09

BART


times each variable appeared in the collection of trees
0
5
10
15
20
failures
address2
address1
sex1
Medu3
Fjob4
goout
famsup2
nursery1
higher2
school2
studytime
guardian2
romantic2
guardian1
Mjob1
internet2
Pstatus1
famrel
activities1
famsize2
Mjob3
freetime
health
Medu5
romantic1
Medu2
Pstatus2
famsize1
reason2
internet1
sex2
Mjob4
Walc
Fedu4
nursery2
Fedu5
Fjob3
schoolsup1
activities2
Medu4
school1
reason4
higher1
paid1
famsup1
reason1
Fedu2
Fjob1
reason3
schoolsup2
Fedu3
Fjob5
paid2
guardian3
Dalc
traveltime
Mjob5
Mjob2
Fjob2
absences
Medu1
Fedu1
age

Minimizing Error

Model	**MSE**
Tree	12.26926
Pruning	10.60838
Bagging	8.293745
R.F.	7.864276
Boosting	7.841
BART	8.096032

Boosting

	tdepth	ntree	lam	olb	ilb
1	4	1000	0.001	8.540	7.211
2	10	1000	0.001	8.309	6.319
3	4	5000	0.001	7.841	4.635
4	10	5000	0.001	7.967	2.999
5	4	1000	0.200	11.361	0.053
6	10	1000	0.200	10.822	0.000
7	4	5000	0.200	10.028	0.000
8	10	5000	0.200	10.911	0.000

BART

	niterations	ntree	nburnin	olb
1	500	1000	100	8.129150
2	1000	1000	100	8.099505
3	500	2000	100	8.131952
4	1000	2000	100	8.108335
5	500	1000	200	8.147487
6	1000	1000	200	8.096032
7	500	2000	200	8.098814
8	1000	2000	200	8.131178

Destined to Fail?

Random Forest

```
> importance(rf.data_s)
              %IncMSE  IncNodePurity
school      7.87236368    177.67317
sex         4.15157182     84.72829
age         3.31704328    162.25867
address     3.93901150     76.18431
famsize    -0.27372196     50.56302
Pstatus     2.22903712     47.22555
Medu       10.31763237    249.41070
Fedu        7.67767530    235.20891
Mjob        2.69822739    231.82070
Fjob        3.60368232    167.77553
reason      1.61683968    195.18425
guardian    0.66362028     74.14570
traveltime -1.16723421     87.56314
studytime   7.20414450    167.49438
failures   23.45951343    533.77967
schoolsup   5.91085597     65.41860
famsup     -1.64554109     60.72090
paid        0.90949890     29.50861
activities  4.14095148     66.73110
nursery     1.97687039     47.22983
higher     13.54650867    212.23029
internet    3.23025726     83.95035
romantic   -0.21261869     70.12948
famrel     -0.09498381    156.68320
freetime    3.47810110    183.82659
goout       2.67437908    208.30125
Dalc        0.94545908    137.73896
Walc        3.80621632    177.78804
health      2.03990591    167.34255
absences    3.55376171    212.99961
```

Bagging

```
> importance(bag.data_s)
              %IncMSE  IncNodePurity
school      9.01740178    157.84985
sex         3.35823224     69.31898
age         3.77131560    169.13127
address     2.39811576     56.57384
famsize     0.16032176     36.27757
Pstatus    -0.09693026     30.16329
Medu        8.93553736    230.85451
Fedu       10.96130537    253.37108
Mjob        5.78383714    281.97279
Fjob        7.14710752    173.76048
reason      3.08188091    197.73845
guardian   -1.09959138     56.94269
traveltime -0.11889317     64.94913
studytime   5.92189317    150.18334
failures   32.71783202   1022.41379
schoolsup   7.68675397     85.71313
famsup      0.80816358     57.36454
paid       -0.59200995     16.22530
activities -1.26951988     55.80437
nursery     2.38710335     39.29559
higher     15.55814394    131.84226
internet   -1.38793344     64.10842
romantic    1.29851425     70.72750
famrel      0.26497954    142.24913
freetime   -1.32690204    165.98277
goout       4.16456294    212.36576
Dalc       -2.33675069    130.24237
Walc        0.89192154    156.66969
health      6.43471017    169.63197
absences    5.06168352    282.36994
```

Business Applications

Universities and Admissions Officers

- **Program Performance** Our model can be used by universities to predict applicants expected performance in their programs and make admissions decisions based on that information.
- **Scholarships:** It will also be able to assist admission officers with distribution of merit based scholarships and aid based on not just current but future grade predictions
- **Current Students:** The university will be able to adjust their academic policies in a way that supports students who are predicted to struggle with academics. Professors will be able to provide extra support in courses where mean predicted grade is low.
- **Ranking and Reputation:**The model aims to create a positive impact on performance of students, students graduating and overall ranking and reputation of the institution

Business Applications

Banks and Student Loan Institutions

- **Risk Assessment for Student Loans**: Can be used to assess the risk associated with lending money to individual students
- **Default Risk Assessment:** Banks and loan institutions can use the predictive model to assess the risk of default on existing student loans
- **Personalized Loan Offers:** Banks and student loan providers can use the predictive model to offer personalized loan terms based on a student's predicted academic performance
- **Risk Diversification and Portfolio Management:** For banks that have a portfolio of student loans, the model will help diversify the loan distribution based on predicted academic performance

Limitations

- **Sample Bias:** The data used to train our model is not representative of the entire student population. The data comes from Portugal and the predictions might not generalize well to other populations
- **Data Reliability:** We are relying on the fact that students are honest when asked about their frequency of alcohol consumption which wasn't a factor, when in reality that could very well not be case
- **Changing Behavior:** Student behavior, can change over time. If the data used to train the model becomes outdated, the model's predictions may no longer align with current patterns

THANK YOU