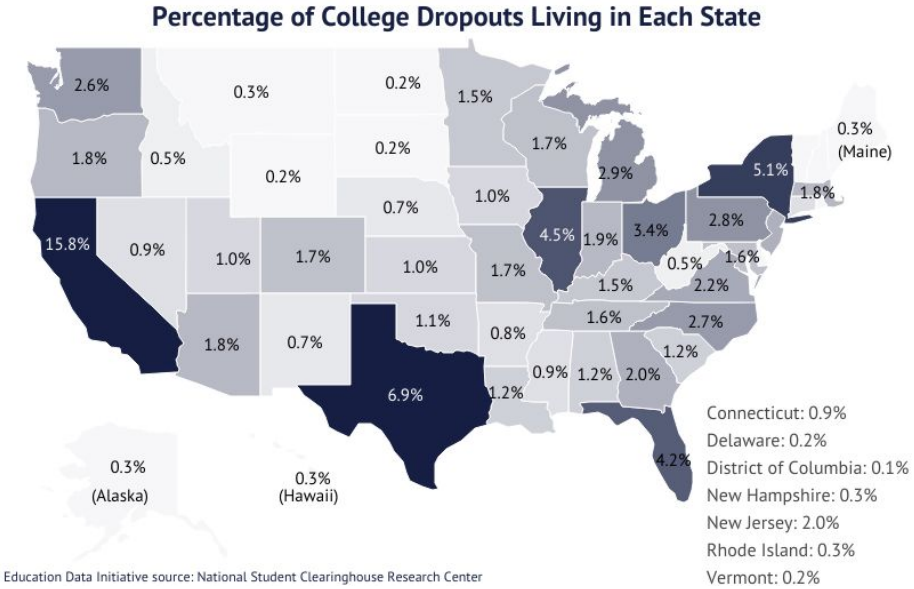
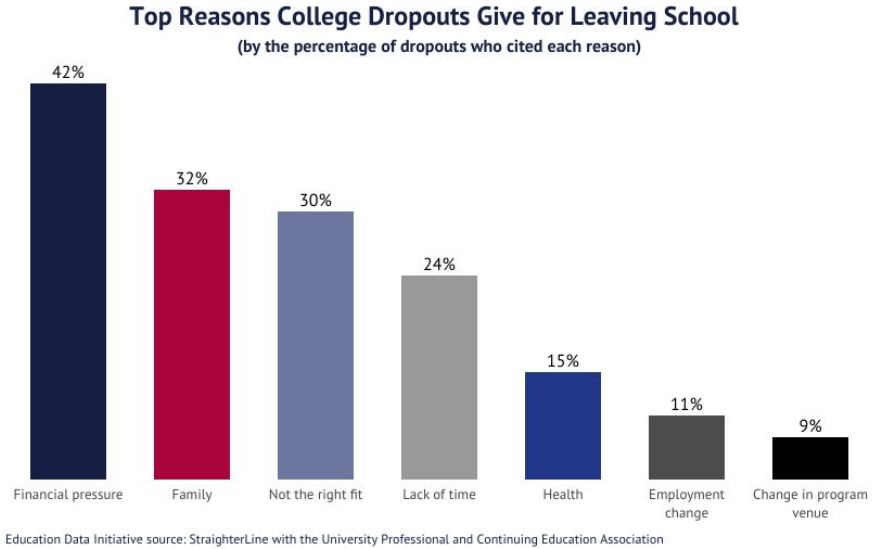
Predicting Student Dropout

Avery Li, Rem Turatbekov

Background

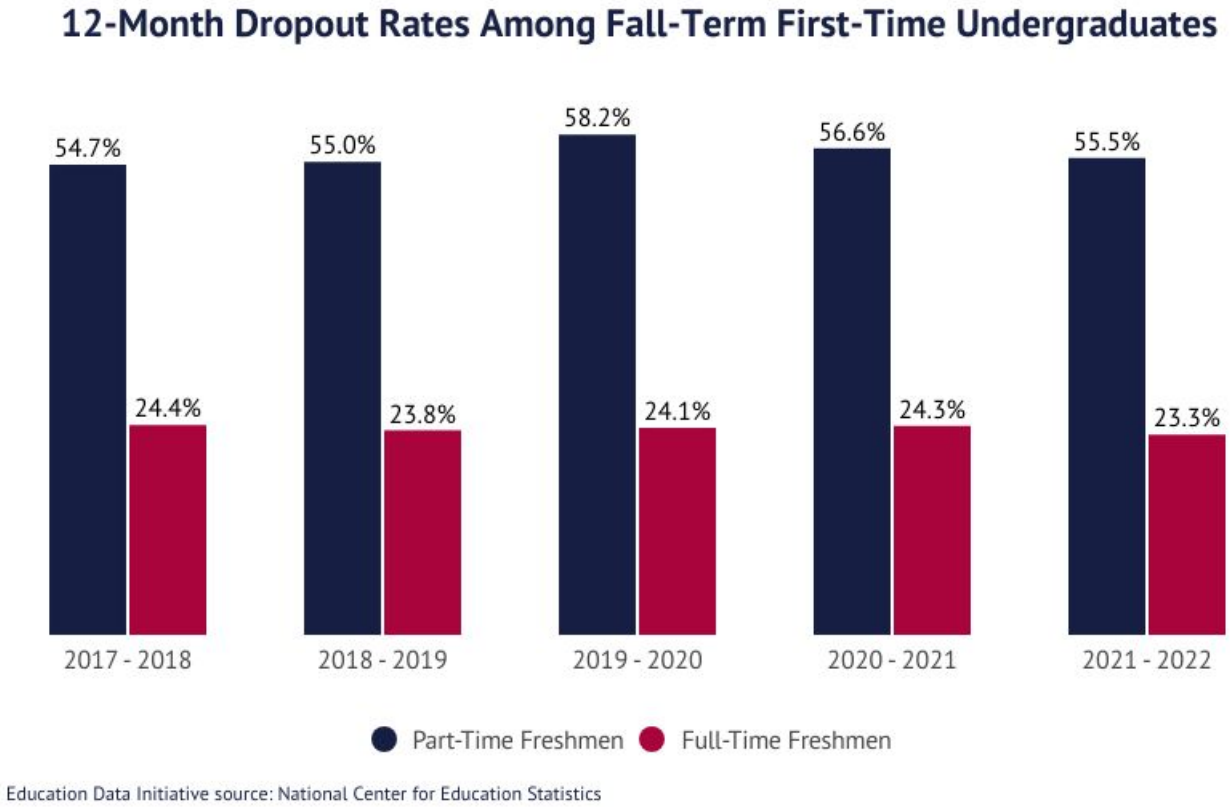




https://educationdata.org/

Highlights

- First-time full-time undergraduate freshmen have a 12-month dropout rate of **23.3%**
- **41.9** million Americans were college dropouts as of July 2022; 943,169 of them re-enrolled that fall
- Dropouts make an average of **35% less income** than bachelor's degree holders
- Dropouts are **20% more likely** to be unemployed than any degree holder



https://educationdata.org/

Project Introduction

- Prevalent dropouts
  - COVID-19
- Oversaturation in workforce

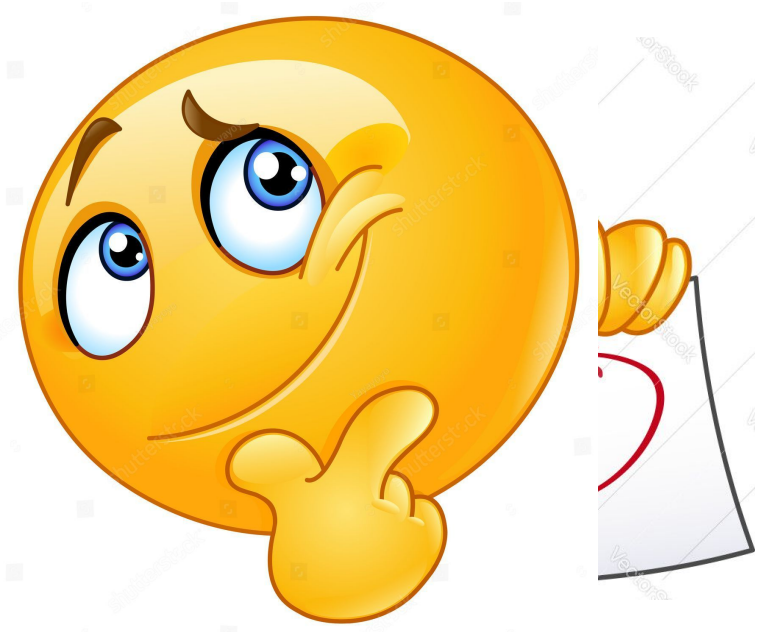
But **WHY**?

- Academic success = success in workforce
- Adapt curricula

Data Description

- UCI Repository - 2021 data
- 36 attributes - student information and personal background
  - Nationality, Father’s Occupation, Tuition fees, Grades
  - Instance = student-situation pair
- Uniform distribution of class (“Target”)
  - Dropout, Graduate, Enrolled
- No missing values

How did we **preprocessing**?



Data Issues

- Data delimited by semicolon
- Integer to string mapped data
  - Creates false sense of order for nominal data
  - Nationality, Marital Status
- Derived attributes
  - “International” = “Non-portuguese”
- Normalize quantitative data
  - Admission grade, Age at enrollment, GDP
  - Decimal scaling from 0 to 5
- Class attribute → binary classification
- Weka cross validation
  - Data okay size

Attribute Selection

We used the following Weka attribute selection models:

- CorrelationAttributeEval
- InfoGainAttributeEval
- GainRatioAttributeEval
- OneRAttributeEval

And:

- Intuition

CorrelationAttributeEval

Calculates r-value with the PCC:

$$\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

Threshold: r > .075

1	Marital status
2	Application mode
3	Daytime/evening attendance
4	Previous qualification
5	Previous qualification (grade)
6	Admission grade
7	Displaced
8	Debtor
9	Tuition fees up to date
10	Gender
11	Scholarship holder
12	Age at enrollment
13	Curricular units 1st sem (enrolled)
14	Curricular units 1st sem (evaluations)
15	Curricular units 1st sem (approved)
16	Curricular units 1st sem (grade)
17	Curricular units 2nd sem (enrolled)
18	Curricular units 2nd sem (evaluations)
19	Curricular units 2nd sem (approved)
20	Curricular units 2nd sem (grade)
21	Curricular units 2nd sem (without evaluations)
22	Target

InfoGainAttributeEval

Difference in entropies between output class and the single class decision tree

$$H = -(\sum p_i \log_2 p_i)$$

$$IG_{split} = H - (\sum \frac{|D_j|}{|D|} \times H_j)$$

Threshold: $IG_{split} > .025$

1. Application mode
2. Course
3. Previous qualification
4. Previous qualification (grade)
5. Mother_qualification
6. Father_qualification
7. Mother_occupation
8. Father_occupation
9. Debtor
10. Tuition fees up to date
11. Gender
12. Scholarship holder
13. Age at enrollment
14. Curricular units 1st sem (evaluations)
15. Curricular units 1st sem (approved)
16. Curricular units 1st sem (grade)
17. Curricular units 2nd sem (enrolled)
18. Curricular units 2nd sem (evaluations)
19. Curricular units 2nd sem (approved)
20. Curricular units 2nd sem (grade)
21. Target

GainRatioAttributeEval

Reduces bias from the information gain metrics.

Intrinsic Information is the entropy of the proportions of the subcategories themselves

$$II = -(\sum \frac{|D_j|}{|D|} \times \log_2 (\frac{|D_j|}{|D|}))$$

$$GR = \frac{IG_{split}}{II}$$

Threshold: GR > .14

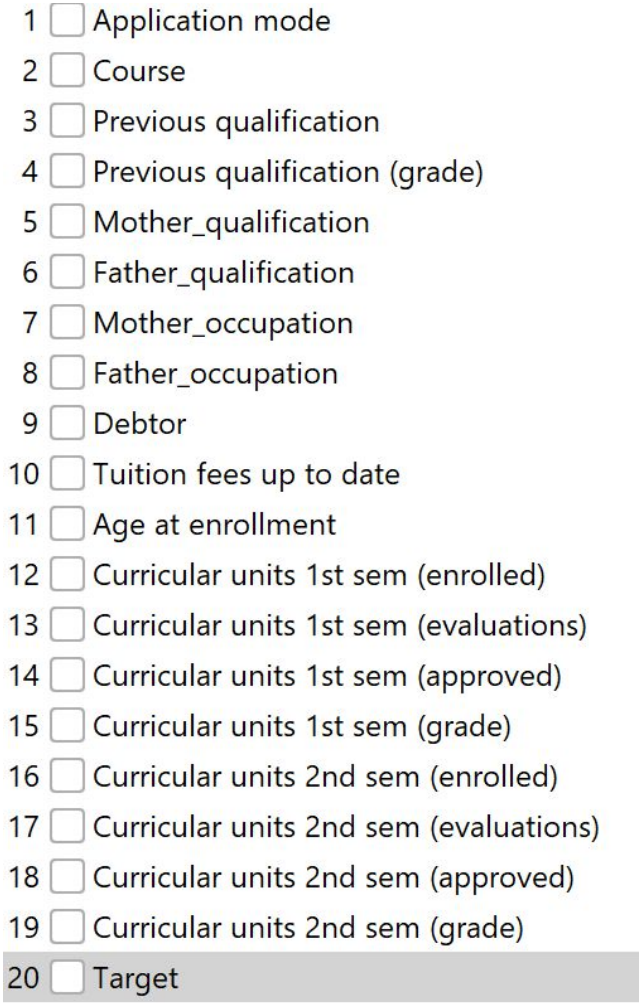
1	Marital status
2	Application mode
3	Previous qualification
4	Previous qualification (grade)
5	Nacionality
6	Admission grade
7	Debtor
8	Tuition fees up to date
9	Gender
10	Scholarship holder
11	Age at enrollment
12	Curricular units 1st sem (enrolled)
13	Curricular units 1st sem (evaluations)
14	Curricular units 1st sem (approved)
15	Curricular units 1st sem (grade)
16	Curricular units 1st sem (without evaluations)
17	Curricular units 2nd sem (enrolled)
18	Curricular units 2nd sem (evaluations)
19	Curricular units 2nd sem (approved)
20	Curricular units 2nd sem (grade)
21	Curricular units 2nd sem (without evaluations)
22	Target

OneRAttributeEval

Evaluates worth of an attribute by a one-to-one ratio of the class labels.

```
for each attribute:
      for each unique value in attribute:
            count labels for instances with the unique value
            determine most frequent label
            assign value to label
      compute error rate given rule chosen
choose rule with lowest error rate
```

Threshold: worth > 68%

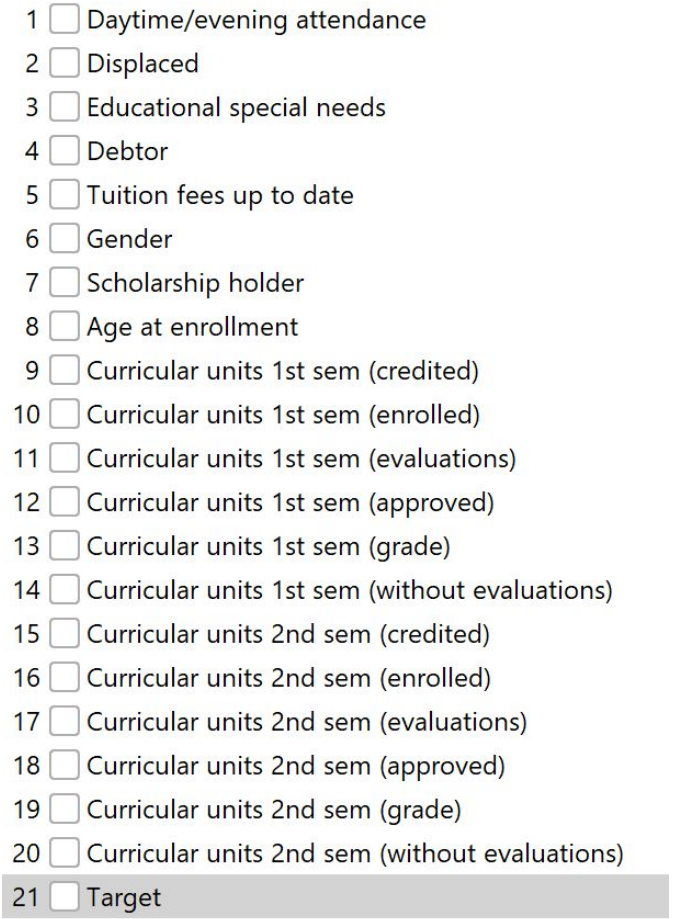


Intuition

- People drop multi if they have a bad grade
- People drop school if they have bad grades

Other financial and personal factors also important

- Special needs, debtor, tuition fee, etc



Classifier Models

- J48
  - Information gain decision tree
- NaiveBayes
  - Probabilistic classifier
  - Conditionally independent
- Logistic
  - Estimates class probabilities
  - Maximize log-likelihood
- Random Forest
  - Generates collection of random trees
  - Evaluates using this “forest”

Performance Metrics

- Accuracy
  - General performance observation
- TP/FP Rate
  - Better to incorrectly predict a non-dropout student as dropout → TPR
  - Predict beforehand and intervene if necessary
- ROC
  - Both FPs and FNs are important, so ROC is acceptable

Accuracy

Top 5 models had minimum accuracy of 86.9%

1. Intuition + Logistic - **87.5%**
2. Correlation + Logistic
3. InfoGain + Logistic
4. Intuition + Random Forest
5. OneR + Logistic

- Accuracy largely dependent on *classification model* used

TPR

Achieved minimum TPR of .710

1. Intuition + Logistic - **.875**
2. Intuition + Random Forest
3. OneR + Logistic
4. InfoGain + Naive Bayes
5. Intuition + J48

ROC

Achieved minimum area under ROC of .906

1. Intuition + Logistic - **.914**
2. Intuition + Random Forest
3. Correlation + Logistic
4. InfoGain + Random Forest
5. Correlation + Random Forest

Best Model

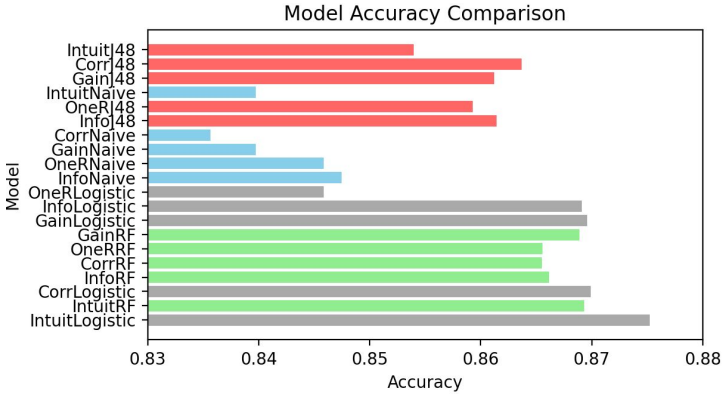
Intuition + Logistic

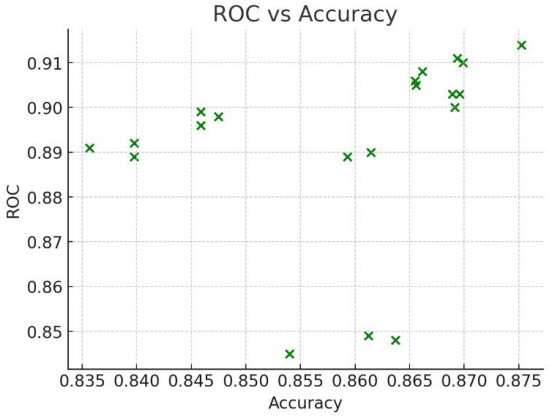
- **87.5%** accuracy
- **.875** TPR**
- **.914** ROC

**TPR most important

	A	B	C	D	E	F	G
1	**Model name**	**Accuracy**	**TP Dropout**	**FP Dropout**	**ROC**	**TP Total**	**FP Total**
2	IntuitLogistic	0.875226	0.725	0.06	0.914	0.875	0.204
3	IntuitRF	0.869349	0.72	0.06	0.911	0.869	0.209
4	CorrLogistic	0.869801	0.709	0.054	0.909	0.87	0.215
5	InfoRF	0.866184	0.701	0.056	0.908	0.866	0.221
6	CorrRF	0.865506	0.713	0.062	0.906	0.866	0.215
7	OneRRF	0.865958	0.695	0.053	0.906	0.866	0.224
8	GainRF	0.865958	0.709	0.06	0.904	0.866	0.217
9	GainLogistic	0.868897	0.705	0.054	0.903	0.869	0.217
10	InfoLogistic	0.869575	0.6891	0.062	0.9	0.87	0.207
11	OneRLogistic	0.869123	0.719	0.06	0.899	0.869	0.21
12	InfoNaive	0.846745	0.719	0.093	0.884	0.847	0.221
13	OneRNaive	0.845841	0.712	0.091	0.878	0.846	0.225
14	GainNaive	0.839738	0.694	0.091	0.873	0.84	0.237
15	CorrNaive	0.836573	0.695	0.097	0.872	0.837	0.238
16	InfoJ48	0.861438	0.692	0.058	0.865	0.861	0.228
17	OneRJ48	0.859629	0.685	0.058	0.862	0.86	0.232
18	IntuitNaive	0.839738	0.685	0.087	0.86	0.84	0.242
19	GainJ48	0.861212	0.706	0.065	0.849	0.869	0.221
20	CorrJ48	0.863698	0.712	0.065	0.848	0.864	0.216
21	IntuitJ48	0.853978	0.715	0.08	0.845	0.854	0.219

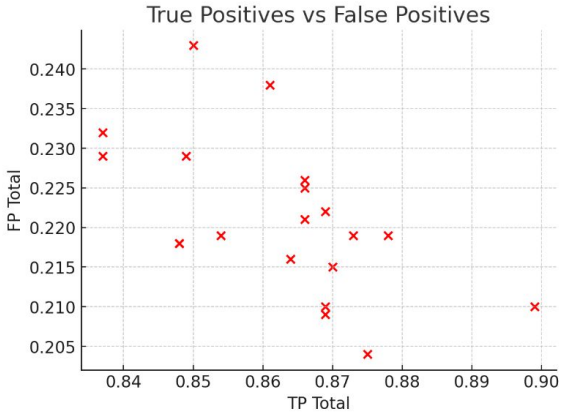
Good accuracy

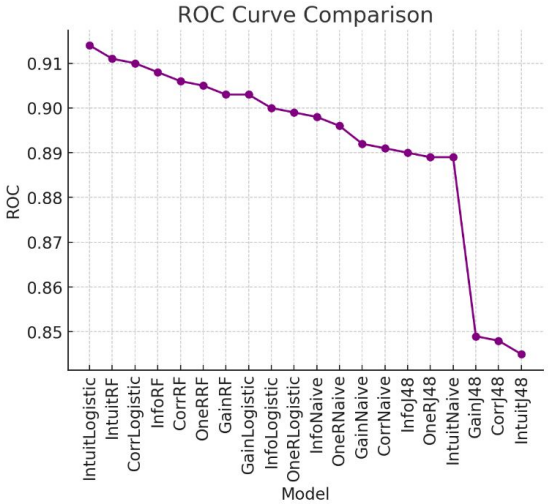




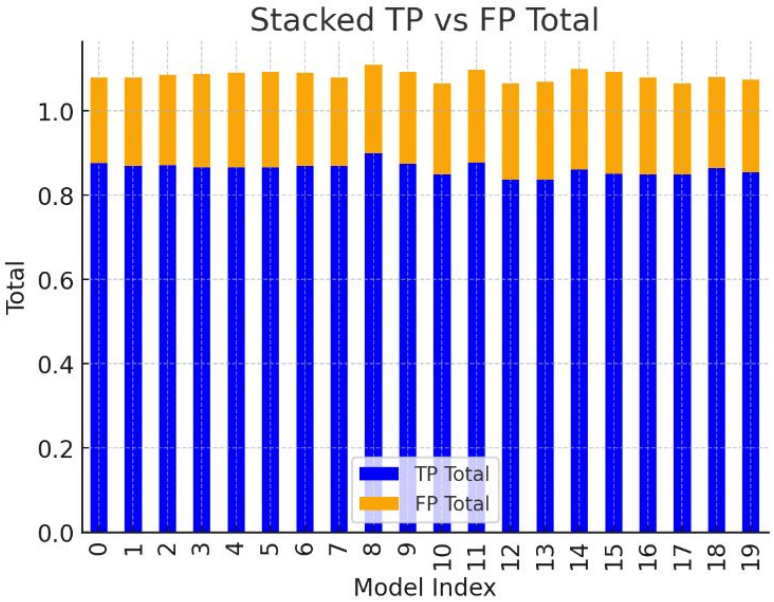
Low accuracy = low ROC (kinda?)

Trade-off

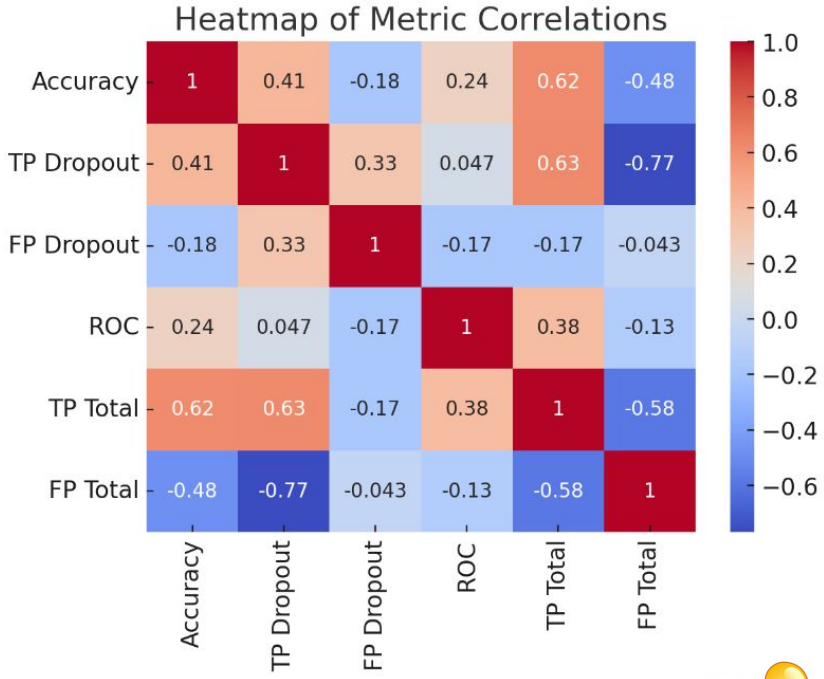




Stock market crash



TP >>> FP

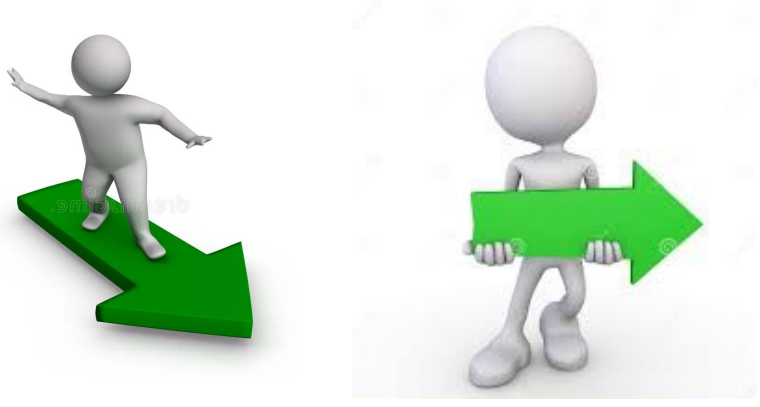


Accuracy has solid correlation with TP Total

Not too much with ROC

Future Research

- Intuition + Logistic works best, but something else probably works better
  - Only used 4 out of 11 available selection algorithms on Weka
    - Minimum Redundancy Maximum Relevance ??
    - Recursive Feature Selection ??
- Expand data to include more features
  - Mental health, extracurricular engagement
- Longitudinal study to see how the predictive factors change over time



Questions?

References

https://educationdata.org/college-dropout-rates