CS3244-10

Amrut Prabhu, Sreyans Sipani, Andres Rojas, Ang Jing Zhe, Eeshan Jaiswal

NUS National University of Singapore | School of Computing

Prediction of Student Earnings and Loan Repayment

Abstract

Did you know that 40 % of borrowers are expected to default on their student loans by 2023[2]? In order to bring down this number, this project aims to predict earnings and repayment rates for US students that solicit financial aid, by making use of aggregate student characteristics data of different US colleges. After experimenting with different machine learning models, we can give good predictions to help prospective aid takers foresee potential financial risks.

Data Pre-processing

Our team worked on the College Scorecard dataset[4] collected by the U.S. Department of Education. It includes information for each college, with ~1800 columns such as household income, earnings, debt, and city. From these , we picked the 10 best features that we could use to train our models.

The main challenge was handling missing values. We resorted to using statistical imputation, through the **MICE algorithm**[5].

Samples iteratively from conditional distributions:

$$P(Y_1 | Y_{-1}, \theta_1) \\ \vdots \\ P(Y_p | Y_{-p}, \theta_p)$$

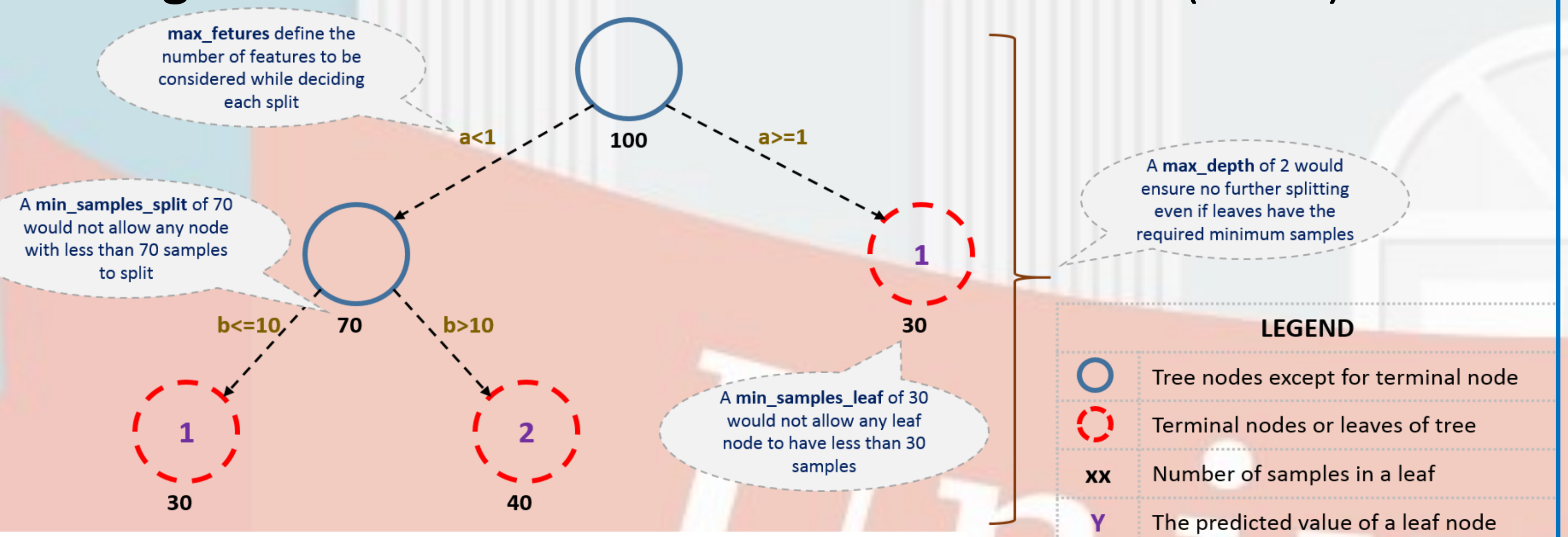
Chained equations:

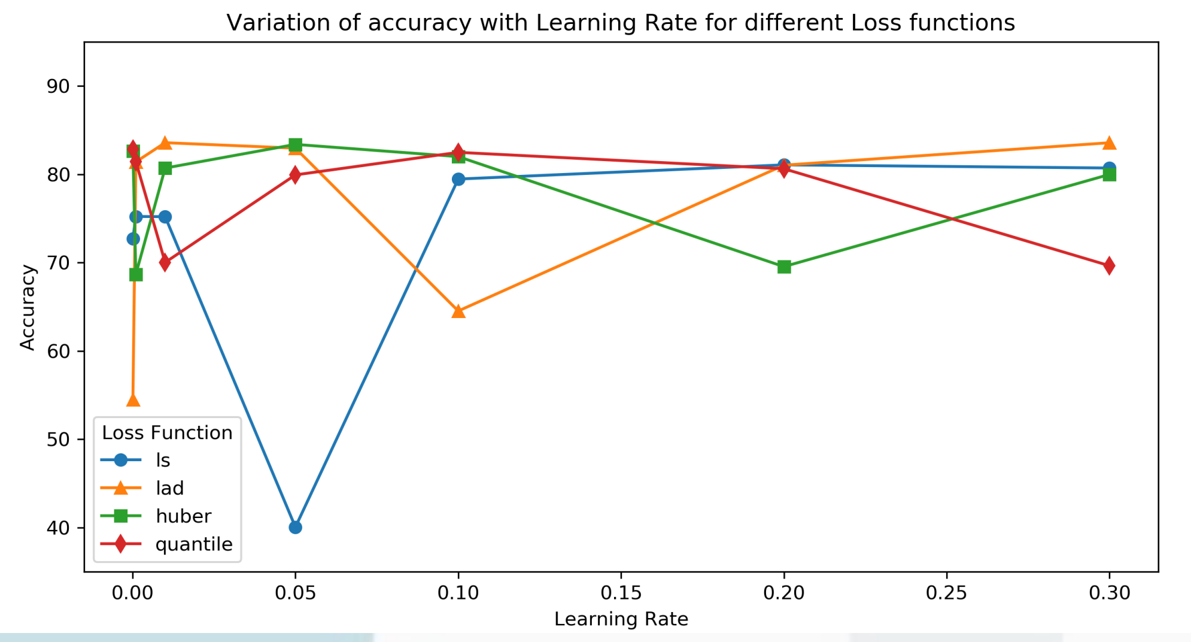
$$\begin{aligned} \theta_p^{*(t)} &\sim P(\theta_p | Y_p^{\text{obs}}, Y_1^{(t)}, \dots, Y_{p-1}^{(t)}) \\ Y_p^{*(t)} &\sim P(Y_p | Y_p^{\text{obs}}, Y_1^{(t)}, \dots, Y_p^{(t)}, \theta_p^{*(t)}) \end{aligned}$$

Methodology

Gradient Boosting Regressor (GBR)[3]

GBR attempts to build weak models from which it learns to create a strong model. We tested different parameters for the trees and more importantly, different loss functions. **GBR paired with a Regressor Chain** is our most accurate model (~86%).

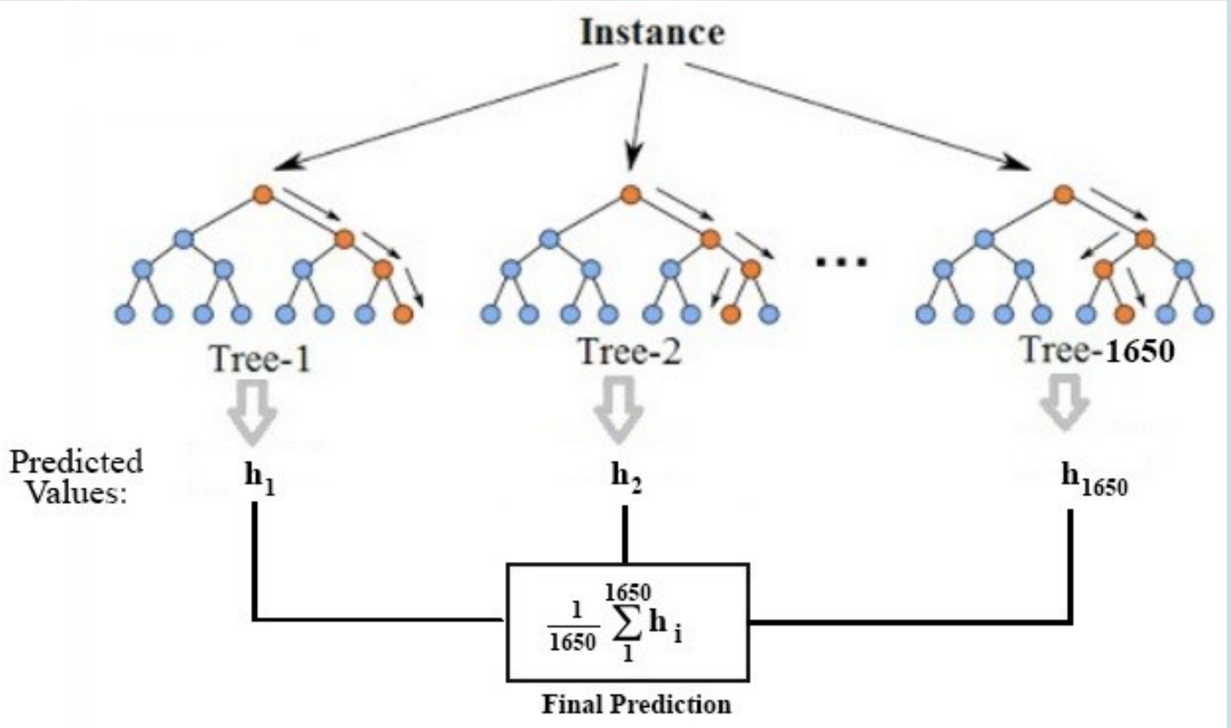

max_fetures define the number of features to be considered while deciding each split
a<1
100
a>=1
A **min_samples_split** of 70 would not allow any node with less than 70 samples to split
A **max_depth** of 2 would ensure no further splitting even if leaves have the required minimum samples
1
30
b<=10
70
b>10
1
30
2
40
A **min_samples_leaf** of 30 would not allow any leaf node to have less than 30 samples
LEGEND
Tree nodes except for terminal node
Terminal nodes or leaves of tree
xx Number of samples in a leaf
Y The predicted value of a leaf node
abc Node splitting criteria



After testing different loss functions, **Least Absolute Deviation** resulted in being the most accurate.

Random Forest Regressor (RFR)[1]

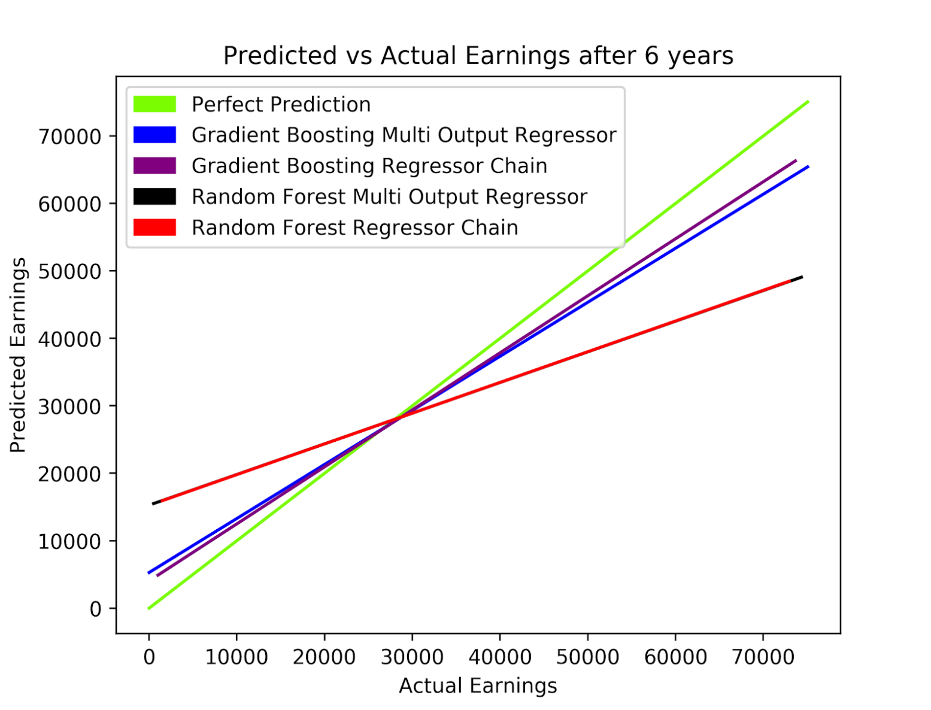
RFR predicts values by aggregating the results from various subtrees. We had to experiment with how many subtrees we would have as well as parameters like depth, minimum samples to classify as a leaf or a split. After hyper-parameter tuning, our best RFR achieved 80% accuracy.

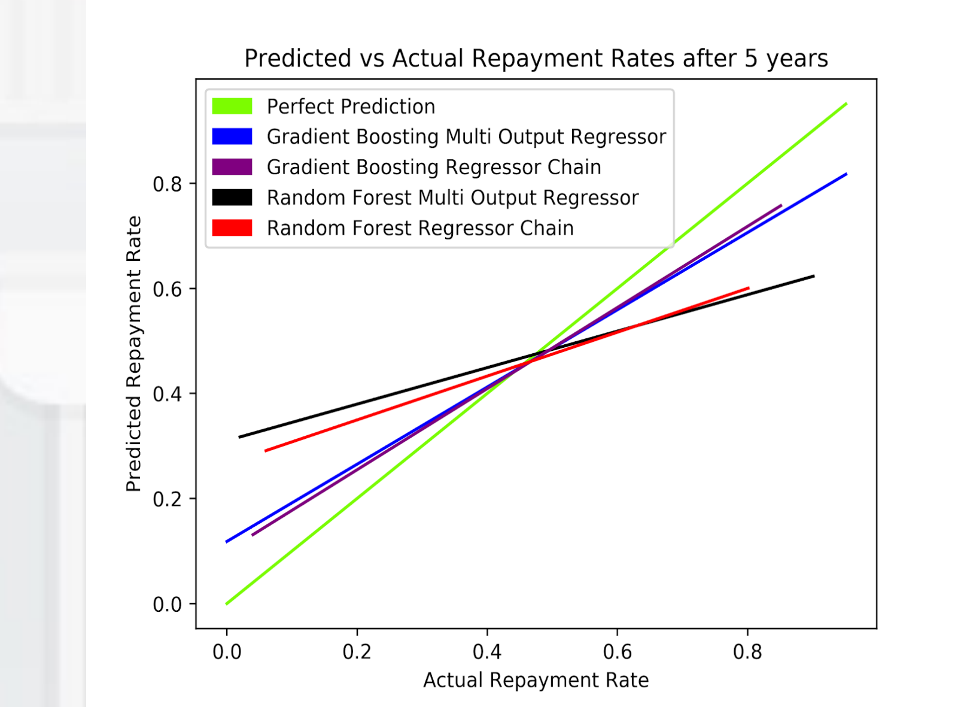


Results

The following table shows how the different regressor models compared against each other in predicting the earnings and the repayment rates. In general, the **Gradient Boosting regressor with a Regressor Chain** is the most accurate.

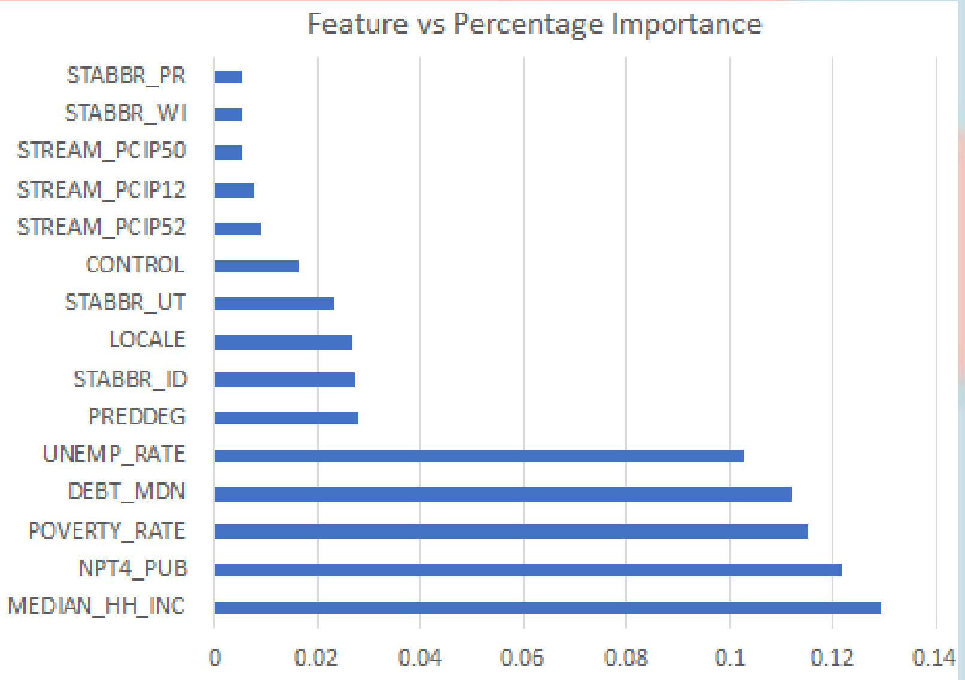
Predicted Value	Gradient Boosting		Random Forest	
	Chain	Multi o/p	Chain	Multi o/p
Earnings in 6 years	**85.73%**	84.33%	80.04%	80.02%
Earnings in 8 years	**85.44%**	83.81%	79.95%	79.35%
Earnings in 10 years	**85.29%**	83.59%	79.73%	78.95%
Repayment 1 year	**72.08%**	71.61%	58.97%	58.96%
Repayment 3 years	**76.27%**	75.97%	68.36%	66.21%
Repayment 5 years	**79.60%**	78.90%	73.35%	70.98%
Repayment 7 years	**82.35%**	81.50%	77.26%	75.93%





Discussion

- The Regressor chain performed better than a standard Multiple Output Regressor. It shows us that the repayment rates in future years are **dependent on the rates for previous years**. But the same is not true for the earnings over the years. This is interesting as we would expect these to be closely related but both do not follow the same trend in this case.



- To calculate the importance of every feature, we decided to use the **gini importance** or the mean decrease impurity of the features. Financial features such as **household income, net price of institution, and debt** affect the model's prediction the most. This is consistent with our own analysis of the results.
- We later added the **dominant stream** as a feature to understand its effect on the performance of the model. It improved the model accuracy by ~1% for the earnings, while interestingly the accuracy for repayment rates seems to decrease slightly (~0.4%).

References

[1] Koehrsen, W. (2017). *Random Forest Simple Explanation.* Medium—[Online]. Available at: https://medium.com/@williamkoehrsen/random-forest-simple-explanation-377895a60d2d
[2] Nova, A. (2018). *More than 1 million people default on their student loans each year.* CNBC—[Online]. Available at: https://www.cnbc.com/2018/08/13/twenty-two-percent-of-student-loan-borrowers-fall-into-default.html
[3] Rogozhnikov, A. (2016). *Gradient Boosting explained [demonstration].* Brilliantly Wrong —[Online]. Available at: http://arogozhnikov.github.io/2016/06/24/gradient_boosting_explained.html
[4] US Department of Education. (n.d.). *College Scorecard Data*, College Scorecard —[Online]. Available at: https://collegescorecard.ed.gov/data/
[5] van Buuren, S. and Groothuis-Oudshoorn, K. (2011). *mice: Multivariate Imputation by Chained Equations in R*. Journal of Statistical Software, 45 (3) [online]. Available at: https://www.jstatsoft.org/article/view/v045i03/v45i03.pdf

Special thanks to the 13th STePS sponsors:

sea connecting the dots
indeed
INFOCOMM MEDIA DEVELOPMENT AUTHORITY
SGINNOVATE

Get a glimpse into your future:

VOTE for us here: