ORIE 4740: Final Project

Predicting Student Loan Debt

Yu Jin Hur(yh586), Jessica Tran (jlt227), Eric Wu (ew398)

Abstract

Projected Student Loan Debt is often a major if not primary factor that may constrain students' considerations when applying to and choosing between a multitude of four year institutions. As such, identifying the characteristics of each university that affect student loan debt can be highly beneficial to understanding how each factor relates to student loans. The objective of this project is to produce myriad regression models on student loan debt, using the various qualities of colleges throughout the United States, in order to form conclusions backed by quantitative results. More specifically, the goal is to produce and test linear regression models, nonlinear regression models, random forests, and boosting through cross validation methods to quantify the significance of each indicator and produce an accurate prediction for student loan debt.

1 Introduction

Student loan debt is an ever-growing issue faced by individuals pursuing a college degree today. According to the Federal Reserve, 94.3% of individuals who have education related debt also have student loan debt. The median student debt is $12,000 and the mean student debt is $30,156 [1]. The primary objective of this exploratory data analysis on the U.S. Department of Education's College Scorecard is to develop a model to predict student loan debt (DEBT_MDN). Predictors such as student tuition, population demographics, admission rates, number of students enrolled, post-graduation salary, and student household incomes are explored to see how they can be used in the prediction of student loan debt.

After preprocessing the data, various data mining techniques such as linear regression, regression trees, and nonlinear methods were used. Within linear regression, variables were selected using subset selection and lasso regression. Regression trees were utilized using bagging, random forest, and boosting. Nonlinear methods included fitting a GAM.

After analyzing the results, boosting produced the most accurate model. Through boosting, the most important predictor was median debt held by individuals who receive Pell Grants, a federal grant typically received by low-income undergraduates [3]. This shows that it has the most influence in predicting median student loan debt. So, increased amount of student held by Pell Grant receivers is a predictor associated with increased median student loan debt.

2 Data Cleaning

Through the U.S. Department of Education's 2016 scorecard [2], a raw data set containing a comprehensive list of colleges in the United States and their corresponding qualities (student completion rates, debt and repayment, student earnings, etc) was obtained . For the purposes of the project, each college was treated as an observation and each quality was treated as a variable or indicator. The raw data set itself contained 1,743 variables and 7,700 observations,but a large number of entries in the data set are marked with missing values such as "PrivacySuppressed" and "NULL". To resolve this issue, the following algorithm was implemented using for loops to reduce the number of void entries:

1. Iterate through each of the indicators in the model. If a sufficiently large proportion of values listed under a given indicator are marked as missing ("PrivacySuppressed" or "NULL"), remove the indicator and all of its corresponding values.
2. Iterate through each of the observations in the model. If a sufficiently large proportion of values corresponding to a given observation are marked as missing ("PrivacySuppressed" or "NULL"), remove observation and all of its corresponding values.
3. Repeat the process until the data set is mostly free of missing values.

After the aforementioned algorithm was implemented, the remaining "NULL" values for each indicator were set to equal the mean of the numeric values associated with the indicator. Additionally, indicators such as College Name, Location, and College Website were removed from consideration in the model, since their distinct corresponding values provided no support to the analyses in either the regression or classification setting. Filtering the raw data set through the aforementioned ways produced a reduced data set with 234 variables and 839 observations.

3 Data Mining Technique 1: Linear Regression

3.1 Linear Regression

Throughout the project, student loan debt is modelled as a response through the variable DEBT_MDN (Debt Median). Thus, there are are a total of 233 variables that could potentially be included in the multiple regression. Since there is a high probability that creating a model with all 233 variables could result in overfitting the model, the "full" model is used merely as a control group, to compare results to other linear models, resulting from subset selection and regularization techniques.
The "full" model, with 233 predictors, has an adjusted R^2 of 76.2%, 29 significant predictors at significance level $\alpha = 0.1$, and a p-value of 2.2×10^{-16}. Using K-fold cross validation with K = 10, gives an MSE of 96,190.

3.2 Linear Regression using Subset Selection

Given that the full model contains 233 predictors, using best subset selection would be computationally expensive to filter through the entirety of these predictors. Since forward subset selection and stepwise regression are more computationally feasible, these methods were utilized.
Forward stepwise selection results in a model with 61 predictors (Figure 2), adjusted R-squared values of 78.98%, and a p-value of 2.2×10^{-16}. These results compare favorably to those obtained in the full model with 233 variables. Furthermore, reducing the number of predictors from 234 to 61, improves interpretability and decreases variability at the cost of increased bias and less flexibility. The forward stepwise selection algorithm utilized the lowest AIC to choose the best model. Using K-fold cross validation, where K=10, resulted in a MSE of 96,266. However, the Variance Inflation Factors (VIF) values obtained from the best subset selection suggest that some variables indicate a high level of collinearity with other variables in the model. Thus, CIP09CERT4, CIP47BACHL, HBCU, were each separately removed from the subset model after having the model rerun. These factors had VIF of

62.015469, 33.920756, 14.382856, respectively. (Note that values above 10 indicated that the variable had considerable amount of collinearity in the model)

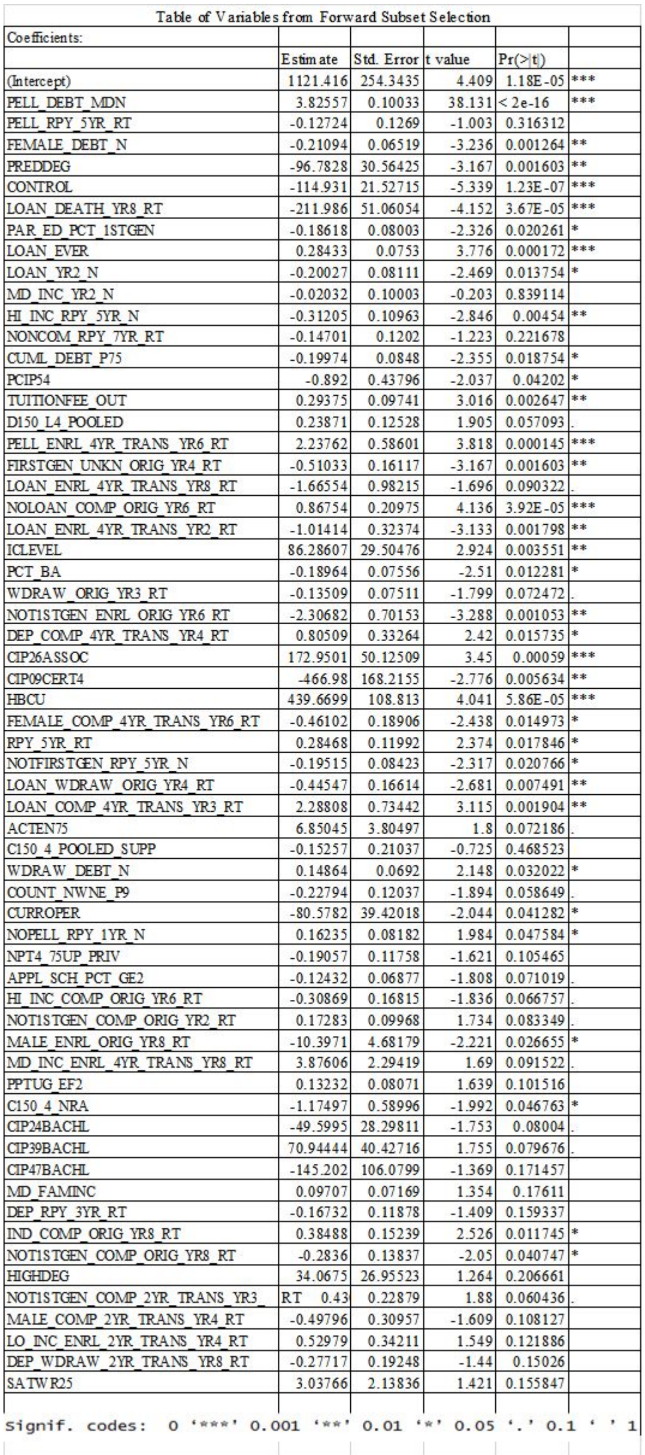


Table of Variables from Forward Subset Selection

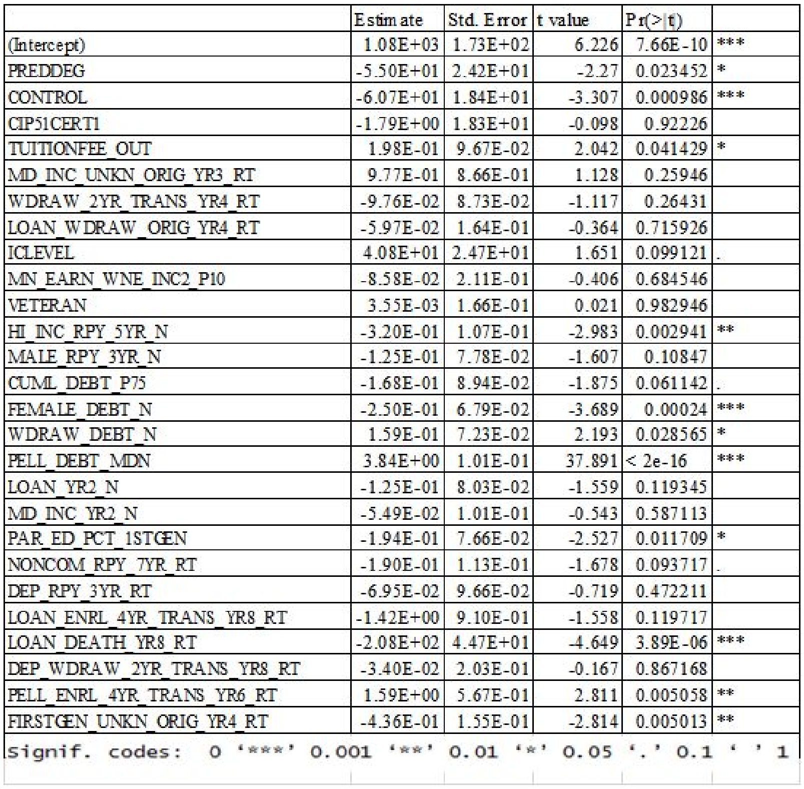
Coefficients:					
	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1121.416	254.3435	4.409	1.18E-05	***
PELL_DEBT_MDN	3.82557	0.10033	38.131	< 2e-16	***
PELL_RPY_5YR_RT	-0.12724	0.1269	-1.003	0.316312	
FEMALE_DEBT_N	-0.21094	0.06519	-3.236	0.001264	**
PREDDEG	-96.7828	30.56425	-3.167	0.001603	**
CONTROL	-114.931	21.52715	-5.339	1.23E-07	***
LOAN_DEATH_YR8_RT	-211.986	51.06054	-4.152	3.67E-05	***
PAR_ED_PCT_1STGEN	-0.18618	0.08003	-2.326	0.020261	*
LOAN_EVER	0.28433	0.0753	3.776	0.000172	***
LOAN_YR2_N	-0.20027	0.08111	-2.469	0.013754	*
MD_INC_YR2_N	-0.02032	0.10003	-0.203	0.839114	
HI_INC_RPY_5YR_N	-0.31205	0.10963	-2.846	0.00454	**
NONCOM_RPY_7YR_RT	-0.14701	0.1202	-1.223	0.221678	
CUML_DEBT_P75	-0.19974	0.0848	-2.355	0.018754	*
PCIP54	-0.892	0.43796	-2.037	0.04202	*
TUITIONFEE_OUT	0.29375	0.09741	3.016	0.002647	**
D150_L4_POOLED	0.23871	0.12528	1.905	0.057093	.
PELL_ENRL_4YR_TRANS_YR6_RT	2.23762	0.58601	3.818	0.000145	***
FIRSTGEN_UNKN_ORIG_YR4_RT	-0.51033	0.16117	-3.167	0.001603	**
LOAN_ENRL_4YR_TRANS_YR8_RT	-1.66554	0.98215	-1.696	0.090322	.
NOLOAN_COMP_ORIG_YR6_RT	0.86754	0.20975	4.136	3.92E-05	***
LOAN_ENRL_4YR_TRANS_YR2_RT	-1.01414	0.32374	-3.133	0.001798	**
ICLEVEL	86.28607	29.50476	2.924	0.003551	**
PCT_BA	-0.18964	0.07556	-2.51	0.012281	*
WDRAW_ORIG_YR3_RT	-0.13509	0.07511	-1.799	0.072472	.
NOT1STGEN_ENRL_ORIG_YR6_RT	-2.30682	0.70153	-3.288	0.001053	**
DEP_COMP_4YR_TRANS_YR4_RT	0.80509	0.33264	2.42	0.015735	*
CIP26ASSOC	172.9501	50.12509	3.45	0.00059	***
CIP09CERT4	-466.98	168.2155	-2.776	0.005634	**
HBCU	439.6699	108.813	4.041	5.86E-05	***
FEMALE_COMP_4YR_TRANS_YR6_RT	-0.46102	0.18906	-2.438	0.014973	*
RPY_5YR_RT	0.28468	0.11992	2.374	0.017846	*
NOTFIRSTGEN_RPY_5YR_N	-0.19515	0.08423	-2.317	0.020766	*
LOAN_WDRAW_ORIG_YR4_RT	-0.44547	0.16614	-2.681	0.007491	**
LOAN_COMP_4YR_TRANS_YR3_RT	2.28808	0.73442	3.115	0.001904	**
ACTEN75	6.85045	3.80497	1.8	0.072186	.
C150_4_POOLED_SUPP	-0.15257	0.21037	-0.725	0.468523	
WDRAW_DEBT_N	0.14864	0.0692	2.148	0.032022	*
COUNT_NWNE_P9	-0.22794	0.12037	-1.894	0.058649	.
CURROPER	-80.5782	39.42018	-2.044	0.041282	*
NOPELL_RPY_1YR_N	0.16235	0.08182	1.984	0.047584	*
NPT4_75UP_PRIV	-0.19057	0.11758	-1.621	0.105465	
APPL_SCH_PCT_GE2	-0.12432	0.06877	-1.808	0.071019	.
HI_INC_COMP_ORIG_YR6_RT	-0.30869	0.16815	-1.836	0.066757	.
NOT1STGEN_COMP_ORIG_YR2_RT	0.17283	0.09968	1.734	0.083349	.
MALE_ENRL_ORIG_YR8_RT	-10.3971	4.68179	-2.221	0.026655	*
MD_INC_ENRL_4YR_TRANS_YR8_RT	3.87606	2.29419	1.69	0.091522	.
PPTUG_EF2	0.13232	0.08071	1.639	0.101516	
C150_4_NRA	-1.17497	0.58996	-1.992	0.046763	*
CIP24BACHL	-49.5995	28.29811	-1.753	0.08004	.
CIP39BACHL	70.94444	40.42716	1.755	0.079676	.
CIP47BACHL	-145.202	106.0799	-1.369	0.171457	
MD_FAMINC	0.09707	0.07169	1.354	0.17611	
DEP_RPY_3YR_RT	-0.16732	0.11878	-1.409	0.159337	
IND_COMP_ORIG_YR8_RT	0.38488	0.15239	2.526	0.011745	*
NOT1STGEN_COMP_ORIG_YR8_RT	-0.2836	0.13837	-2.05	0.040747	*
HIGHDEG	34.0675	26.95523	1.264	0.206661	
NOT1STGEN_COMP_2YR_TRANS_YR3_	RT 0.43	0.22879	1.88	0.060436	.
MALE_COMP_2YR_TRANS_YR4_RT	-0.49796	0.30957	-1.609	0.108127	
LO_INC_ENRL_2YR_TRANS_YR4_RT	0.52979	0.34211	1.549	0.121886	
DEP_WDRAW_2YR_TRANS_YR8_RT	-0.27717	0.19248	-1.44	0.15026	
SATWR25	3.03766	2.13836	1.421	0.155847	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Figure 2: Table that reveals the predictors obtained from forward subset selection.

3.3 Regularization of Linear Regression through Lasso Regression Method

The lasso regression method was used to perform regularization of the linear regression model. Lasso was chosen because this technique forces some coefficients of the linear regression model to zero, allowing for a subset of variables to be chosen. This is necessary due to the large number of variables in the original dataset. Using cross validation, the value for λ that minimized the mean-squared error was 16. After applying lasso regression, this resulted in a model with an adjusted R-squared value of 75.79%, p-value of 2.2×10^{-16}, and 26 variables selected. Although the adjusted R-squared value is lower than the models mentioned in section 3.1 and 3.2, the number of variables is lower. This results in better interpretability and decreased flexibility. However, MSE using K-fold cross validation with K=10 results was 104,522. Using Variance Inflation Factors (VIF) resulted in values of less than 5 for each variable, so collinearity was not detected in variables of this model.



	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.08E+03	1.73E+02	6.226	7.66E-10	***
PREDDEG	-5.50E+01	2.42E+01	-2.27	0.023452	*
CONTROL	-6.07E+01	1.84E+01	-3.307	0.000986	***
CIP51CERT1	-1.79E+00	1.83E+01	-0.098	0.92226	
TUITIONFEE_OUT	1.98E-01	9.67E-02	2.042	0.041429	*
MD_INC_UNKN_ORIG_YR3_RT	9.77E-01	8.66E-01	1.128	0.25946	
WDRAW_2YR_TRANS_YR4_RT	-9.76E-02	8.73E-02	-1.117	0.26431	
LOAN_WDRAW_ORIG_YR4_RT	-5.97E-02	1.64E-01	-0.364	0.715926	
ICLEVEL	4.08E+01	2.47E+01	1.651	0.099121	.
MN_EARN_WNE_INC2_P10	-8.58E-02	2.11E-01	-0.406	0.684546	
VETERAN	3.55E-03	1.66E-01	0.021	0.982946	
HI_INC_RPY_5YR_N	-3.20E-01	1.07E-01	-2.983	0.002941	**
MALE_RPY_3YR_N	-1.25E-01	7.78E-02	-1.607	0.10847	
CUML_DEBT_P75	-1.68E-01	8.94E-02	-1.875	0.061142	.
FEMALE_DEBT_N	-2.50E-01	6.79E-02	-3.689	0.00024	***
WDRAW_DEBT_N	1.59E-01	7.23E-02	2.193	0.028565	*
PELL_DEBT_MDN	3.84E+00	1.01E-01	37.891	< 2e-16	***
LOAN_YR2_N	-1.25E-01	8.03E-02	-1.559	0.119345	
MD_INC_YR2_N	-5.49E-02	1.01E-01	-0.543	0.587113	
PAR_ED_PCT_1STGEN	-1.94E-01	7.66E-02	-2.527	0.011709	*
NONCOM_RPY_7YR_RT	-1.90E-01	1.13E-01	-1.678	0.093717	.
DEP_RPY_3YR_RT	-6.95E-02	9.66E-02	-0.719	0.472211	
LOAN_ENRL_4YR_TRANS_YR8_RT	-1.42E+00	9.10E-01	-1.558	0.119717	
LOAN_DEATH_YR8_RT	-2.08E+02	4.47E+01	-4.649	3.89E-06	***
DEP_WDRAW_2YR_TRANS_YR8_RT	-3.40E-02	2.03E-01	-0.167	0.867168	
PELL_ENRL_4YR_TRANS_YR6_RT	1.59E+00	5.67E-01	2.811	0.005058	**
FIRSTGEN_UNKN_ORIG_YR4_RT	-4.36E-01	1.55E-01	-2.814	0.005013	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Figure 3: Table that reveals the predictors selected using Lasso

4 Data Mining Technique 2: Tree - Based Methods

4.1 Regression Tree

A regression tree was fitted to the data set using a training and test data set. The test set MSE associated with the regression tree is 151,458. The square root of the MSE is around 389, indicating that this model leads to test predictions around $389 of the true median student loan debt. After pruning the tree, the variables actually used in the tree construction were PELL_DEBT_MDN, NPT44-PUB, UNKN_ORIG_YR2_RT, APPL_SCH_PCT_GE2, GRAD_DEBT_MDN10YR_SUPP, CUML_DEBT_P75, and NOPELL_RPY_1YR_N. The number of leaves on the regression tree was 13 and was chosen using cross validation. Utilizing a single regression tree allows for better interpretability, but is often not as accurate due to the high variance and tendency to overfit.

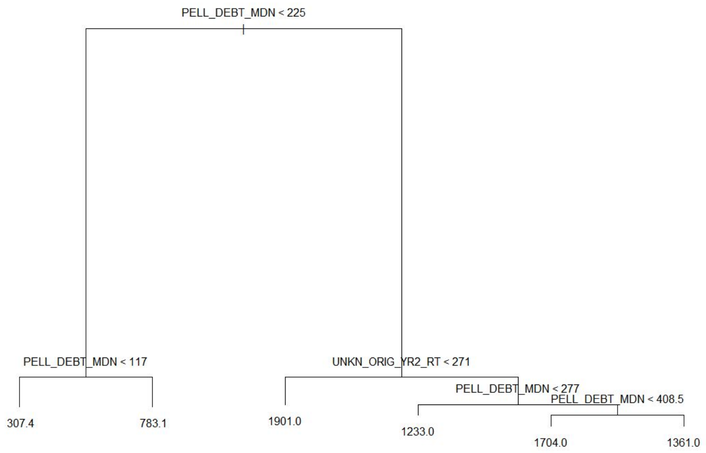


Figure 4: A single classification tree after the growing and pruning steps.

4.2 Bagging and Random Forest

Bagging utilizes all of the 233 predictors in the model. The bagging method resulted in a test MSE of 114,661. Bagging results in improved accuracy over using a single regression tree, but it is harder to interpret. Bagging is a special case of the random forest model where $m = p$. The random forest technique is also utilized by setting $m = \sqrt{233} \approx 15$. Using random forest resulted in a higher MSE over the bagging method of 140,918 when setting the number of trees to 1,000. The top two most important variables were median Pell Grant debt (PELL_DEBT_MDN) and cumulative debt at the 75th percentile (CUML_DEBT_P75).

4.3 Boosting

Boosting was performed by choosing the number of trees (B) to be 5,000 using cross validation. This resulted in a test MSE of 64,263, an improved MSE over the bagging and random forest methods. However, by using the boosting method, there is a risk of overfitting if a large B value is used. As shown in Figure 6, the top 2 most important variables from boosting are the amount of debt from students have receive Pell Grants (PELL_DEBT_MDN), followed by students who received any loan (LOAN_EVER).

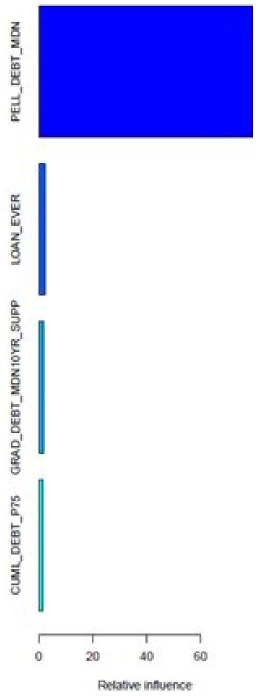


Figure 5: Variables exhibiting the top 4 highest relative influences from boosting

5 Data Mining Technique 3: Nonlinear Methods

5.1 Detecting Nonlinearity through Residual Plots

Currently the linear model has an adjusted R-squared of 0.7898. In order to improve this number, the model will have to account for nonlinearity. To detect nonlinearity, residual graphs were plotted for each coefficient. Nonlinearity is detected when residual graphs show a distinct pattern. Below are the graphs that were fitted with polynomial functions.

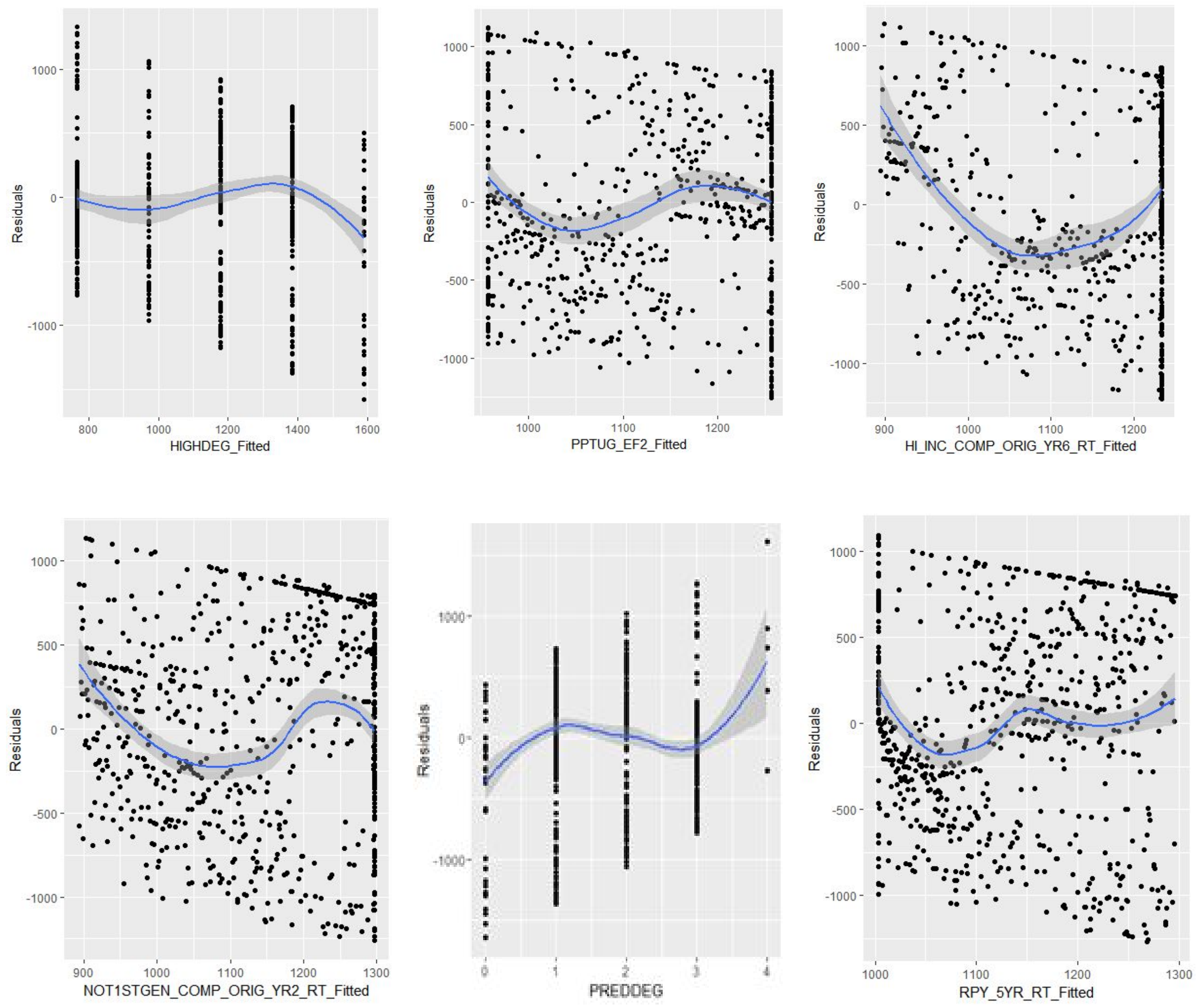


Figure 6: Residual Plots to Detect Nonlinearity

5.2 GAMs

In order to prevent overfitting, only coefficients that showed distinct patterns in the model were fit to various nonlinear functions and added to the linear model which resulted in a GAM. Adjustments to the linear model was made in the following order:

1. Fit PREDEG with a cubic function: adjusted R-squared value increased to 0.704
2. Fit HI_INC_COMP_ORIG_YR6_RT with a square function: adjusted R-squared value increased to 0.7907
3. Fit NOT1STGEN_COMP_ORIG_YR2_RT with a cubic function: adjusted R-squared value increased to 0.7935
4. Fit RPY_5YR_RT with a function to the power of four: adjusted R-squared value increased to 0.7948
5. Fit PPTUG_EF2 with a cubic function: adjusted R-squared increased to 0.7963
6. Fit HIGHDEG with a cubic function: adjusted R-squared increased to 0.7964

Other coefficients that seems to show a pattern in the residual graph such as LOAN_WDRAW_ORIG_YR4_RT, CONTROL, NOTFIRSTGEN_RPY_5YR_N, PCIP54, and TUITIONFEE_OUT were also fitted to nonlinear functions but reduced the adjusted R-square value so these adjustments were not made to the model.

Although the Adjusted R-squared only increased from 0.7898 to 0.7962, the R-squared value increased from 0.803 to 0.8219 meaning that the GAM could explain more than 1 percent more of the variability of data around the mean when compared to a linear model. The validation set approach resulted in a MSE of 116,484.

6 Comparison of Results

As shown in Figure 7 below, boosting resulted in the model with the lowest mean squared error of 64,263. Although there is a risk of overfitting, this is the most accurate method developed. However, the linear regression and GAM approaches may be better for interpretation.

Methods	**Linear Regression using Forward Subset Selection**	**Linear Regression using Lasso**	**Boosting**	**GAM**
Adjusted R^2	78.98%	75.79%	N/A	79.62%
Mean Squared Error	96,266	104,522	64,263	116,484

Figure 7: Table of Methods Used and Results

7 Conclusion

Four approaches, namely linear regression using forward subset selection, linear regression using lasso, boosting, and GAM were used to produce an accurate prediction for student loan debt.

By comparing the results over the approaches, we found that the GAM led to the model that best accounted for the response variable variation with an adjusted R^2 of 79.62%. However, it is only slightly better than the adjusted R^2 for linear regression using forward subset selection (difference of only 0.64%) and linear regression using lasso (difference of only 3.7%).

GAM, however, did not perform very well compared to the other three models in terms of its mean squared error as it had the highest MSE out of all the four models. The model that performed the best in terms of MSE was the model that used Boosting. In fact, it had more than a 30,000 difference in MSE compared to the linear regression model that used forward subset selection (the next lowest model in terms of MSE) and more than 50,000 difference in MSE with GAM which had performed the best in terms of adjusted R$^{2..}$

Since the difference in adjusted R^2 is very small, the comparison of the mean squared error was weighted more when considering which model to recommend. Therefore, we will recommend the boosting method for the prediction of student loan debt.

8 Bibliography

[1] “Report on the Economic Well-Being of U.S. Households in 2015” Board of Governors of the Federal Reserve System https://www.federalreserve.gov/econresdata/2016-economic-well-being-of-us-households-in-2015-education-debt-student-loans.htm

[2] “College Scorecard Data” U.S. Department of Education https://collegescorecard.ed.gov/data/

[3] Federal Pell Grants https://ed.gov/programs/fpg/index.html?exp=0

[4] James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning with Applications in R* http://www-bcf.usc.edu/~gareth/ISL/ISLR%20Sixth%20Printing.pdf

[5] Variable Selection http://www.stat.columbia.edu/~martin/W2024/R10.pdf

[6] Documentation for Package ‘car’ https://cran.r-project.org/web/packages/car/car.pdf

[7] Documentation for Package ‘leaps’ https://cran.r-project.org/web/packages/leaps/leaps.pdf

[8] Documentation for Package ‘glmnet’ https://cran.r-project.org/web/packages/glmnet/index.html

[9] Documentation for Package ‘tree’ https://cran.r-project.org/web/packages/tree/tree.pdf

[10] Documentation for Package ‘randomForest’ https://cran.r-project.org/web/packages/randomForest/randomForest.pdf

[11] Documentation for Package ‘gbm’ https://cran.r-project.org/web/packages/gbm/gbm.pdf

[12] Documentation for Package ‘readr’ https://cran.r-project.org/web/packages/readr/readr.pdf