MCIS 6283 Machine Learning Term Project

Jonathan King

ID: 815000584 | APRIL 30, 2022

Executive Summary

The US Department of Education has a database containing over 6,000 institution of higher education. This lists the performance of each school in over 2,900 different areas. This project looked at a few of these areas to determine if the data could be used to predict different aspects of the school. To do this, the data from the scorecard was processed and run through multiple models to determine the best model used predict the outcome. Below are the different aspects that were looked at along with the model that was chosen best fit the data along with the assessment results.

These aspects were:

- Comparing average ACT scores to average SAT scores of students enrolled in college:
  - Linear regression model with a R2 of 0.9693
- How the racial makeup of undergraduates and size of the school can predict if a 4-year college is Public, Private nonprofit or Private for-profit
  - Random Forest with an accuracy of 0.79
- How the racial makeup of undergraduates and size of the school can predict if a 2-year college is Public, Private nonprofit or Private for-profit
  - Random Forest with an accuracy of 0.845
- What combinations of degrees awarded can predict earnings after college for students for 4-year public schools
  - Random Forest with a R2 of 0.75
- How the average amount of earnings after finishing college and how the college is controlled (Public, Private nonprofit or Private for-profit) can predict a schools school's predominate degree (associate, bachelor's, etc.) awarded
  - Decision Tree classifier with an accuracy of 0.67
- How the amount of debt a student has after leaving college can predict a school's predominate degree (associate, bachelor's, etc.) type
  - Voting classifier with an accuracy of 0.70.
- How the racial makeup of the school, enrollment, cost to attend, admission rates, the parents' education and household income can predict what percentage of students received a Pell Grant
  - Linear regression of 0.715

In genera, the models worked fairly well in predicting the outcome using the training data and comparing it to the test data.

Introduction

The purpose of this project is to analyze data from the US Department of Education's College Scorecard database. "The College Scorecard was created in 2013 under President Barack Obama's administration to make data about colleges more accessible to consumers in a centralized, interactive tool" (Kerr, 2020) The scorecard contains a wide variety of more than 2,900+ different categories of information on more than 6,000+ schools of higher education ranging from public, private nonprofit, to private for-profit.

Analysis Questions

This project will use regression and classification analysis on a select number of categories to focus in on seven areas of how secondary education as a whole perform. These seven areas are:

- Comparing average ACT scores to average SAT scores of students enrolled in college
- How the racial makeup of undergraduates and size of the school can be used to determine what type of a 4-year college it is: Public, Private nonprofit or Private for-profit
- How the racial makeup of undergraduates and size of the school can be used to determine what type of a 2-year college it is: Public, Private nonprofit or Private for-profit
- What combinations of degrees awarded can predict earnings after college for students at 4 year public schools
- How the average amount of earnings after finishing college and how the college is controlled (Public, Private nonprofit or Private for-profit) can predict a schools school's predominate degree (associate, bachelor's, etc.) awarded
- How the amount of debt a student has after leaving college can predict a school's predominate degree (associate, bachelor's, etc.) type
- How the racial makeup of the school, enrollment, cost to attend, admission rates, the parents' education and household income can predict what percentage of students received a Pell Grant

Data

The data obtained from the Department of Education (US Department of Education, n.d.) was downloaded from the “Download the Data” page on their website. The “Most Recent Institution-Level Data” was the zip file chosen to be downloaded. A College Scorecard data dictionary was downloaded (US Department of Education, n.d.) in conjunction with the master file to give definition and further insight to the data itself.

Processing

To complete the analysis of for this project several Python libraries were used. These were Pandas (pandas v 1.4.2, 2022) to format the data, Scikit-learn (Scikit-learn v 1.0.2, 2022) to preform regression and classification, and Matplotlib (matplotlib v 3.5, 2022) to visualize the data.

Because of the massive number of columns in the dataset for each school, this analysis greatly reduced the dataset to only focus on 70 columns. Along with reducing the size of the dataset, there was a need to replace any fields that contained ‘PrivacySuppressed’ with a Nan so that processing of the data could take place in the models. This was done using the pandas library. During the different analyses any rows with Null or Nan values with be dropped. There were also several numeric columns that when read by the Pandas library were converted to an ‘object’ datatype. These columns had to be then converted to a numeric data type to be processed.

To complete the analysis, all dependent and independent variables were split into training and test data sets in a 75%/25% split using Scikit-Learn.

Two functions were built in the code to quickly process the data through multiple regression or classification models. The regressor function returned the R2 and root means square error for Linear, Lasso, Decision tree, and Random forest regression models. The classification function return the Accuracy, Recall, and Precision scores for KNN, SVC, Decision Tree, Random forest, Naive Bayes, and Voting classifier models.

Analysis of data

Comparing average ACT scores to average SAT scores of students enrolled in college

Throughout the country high schools will promote taking either the ACT or SAT as an entrance exam into college. Each college list their average ACT (ACTCMMID) and average SAT (SAT_AVG) score for those enrolled in their school. A linear regression was conducted to see how the ACT aligned with the SAT and scores and if the ACT scores could predict the SAT scores. The R2 value from the mode scored a .9693 and as seen in the graphs (Figure 1 Training set, Figure 2 Test Test) that the prediction line follows very closely to the data, showing that the schools average ACT follows in line with the average SAT of the students entering the university.

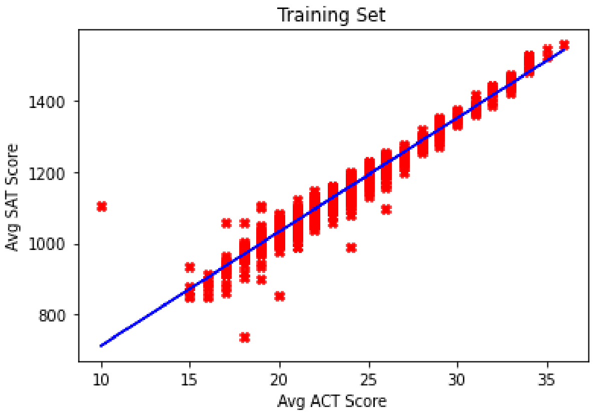

Figure 2

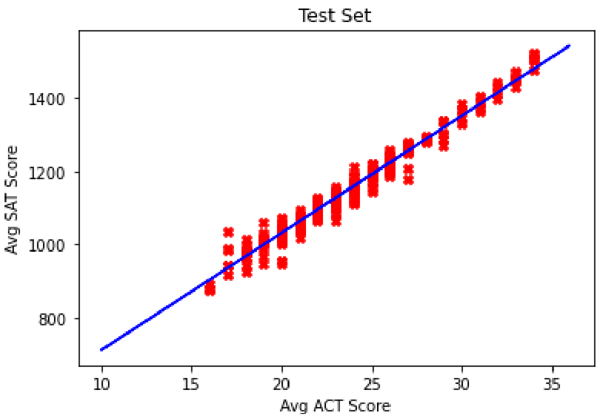

Figure 2

How the racial makeup of undergraduates and size of the school can be used to determine what type of a 4-year college it is: Public, Private nonprofit or Private for-profit

This analysis was done to classify 4-year colleges by the control of the institution (Public, Private nonprofit, or Private for-profit) using the racial makeup of the college by percentage of enrollment and overall enrollment. Table 1 gives the cross reference for the coding used in the dataset.

Table 1

Code	Race
UGDS	Enrollment of undergrad students
UGDS_WHITE	Undergrad students who are white
UGDS_BLACK	Undergrad students who are black
UGDS_HISP	Undergrad students who are Hispanic
UGDS_ASIAN	Undergrad students who are Asian
UGDS_AIAN	Undergrad students who are American Indian/Alaska Native
UGDS_NHPI	Undergrad students who are Native Hawaiian/Pacific Islander
UGDS_2MOR	Undergrad students who are two or more races
UGDS_NRA	Undergrad students who are non-resident aliens
UGDS_UNKN	Undergrad students whose race is unknown

A Random Forest model produced the best results with accuracy of .79 classifying the type of 4 year college. Graphing (Figure 3) the feature importance from the independent variables shows that the overall size of the enrollment has the greatest effect on classifying a college by Public, Private nonprofit, or Private for-profit. Because a Random Forest model is a construction of multiple Decision trees, and example of one of the decision trees that was used in the Random Forest was graphed out (Figure 4) to show the level and complexity of the model

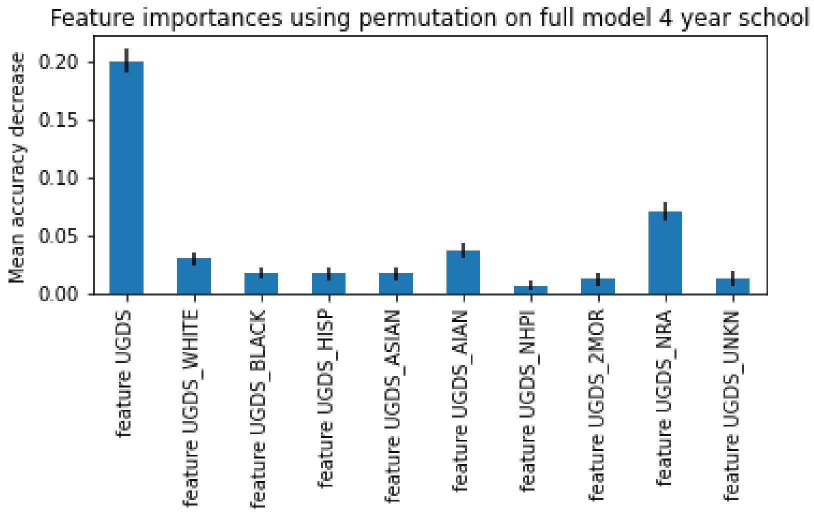

Figure 3

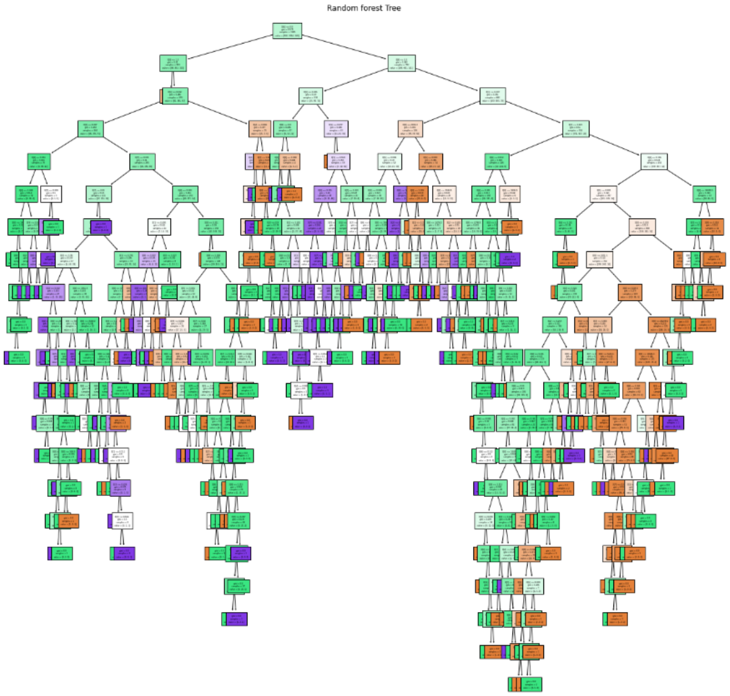

Figure 4

How the racial makeup of undergraduates and size of the school can be used to determine what type of a 2-year college it is: Public, Private nonprofit or Private for-profit

This analysis was done to classify 2-year colleges by the control of the institution; Public, Private nonprofit, or Private for-profit using the racial makeup of the college by percentage of enrollment and overall enrollment. The Random Forest model produced the best results with accuracy of 0.845. Graphing (Figure 5) the feature importance from the independent variables again shows that the overall size of the enrollment is the most important in classifying a college either Public, Private nonprofit, or Private for-profit.

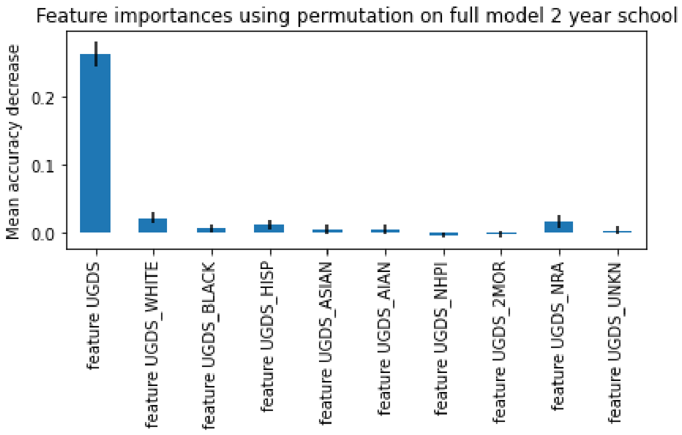

Figure 5

What combinations of degrees awarded can predict earnings after college for students for 4-year public schools

In the dataset, there were a possibility of 54 different degrees that could be awarded by the colleges. To see how these various degrees impacted the earing potential of students, a regression analysis was completed on 4-year public colleges. An analysis was done to find a model that would predict the median earnings of students working and not enrolled 10 years after entry, based of the percentage breakdown of the degrees award by the college. Table 2 gives the cross reference for the coding used in the dataset for the degrees awarded.

Table 2

Code	Degrees	Code	Degrees
PCIP01	Agriculture, Agriculture Operations, And Related Sciences.	PCIP29	Military Technologies And Applied Sciences.
PCIP03	Natural Resources And Conservation.	PCIP30	Multi/Interdisciplinary Studies.
PCIP04	Architecture And Related Services.	PCIP31	Parks, Recreation, Leisure, And Fitness Studies.
PCIP05	Area, Ethnic, Cultural, Gender, And Group Studies.	PCIP38	Philosophy And Religious Studies.
PCIP09	Communication, Journalism, And Related Programs.	PCIP39	Theology And Religious Vocations.
PCIP10	Communications Technologies/Technicians And Support Services.	PCIP40	Physical Sciences.
PCIP11	Computer And Information Sciences And Support Services.	PCIP41	Science Technologies/Technicians.
PCIP12	Personal And Culinary Services.	PCIP42	Psychology.
PCIP13	Education.	PCIP43	Homeland Security, Law Enforcement, Firefighting And Related Protective Services.
PCIP14	Engineering.	PCIP44	Public Administration And Social Service Professions.
PCIP15	Engineering Technologies And Engineering-Related Fields.	PCIP45	Social Sciences.
PCIP16	Foreign Languages, Literatures, And Linguistics.	PCIP46	Construction Trades.
PCIP19	Family And Consumer Sciences/Human Sciences.	PCIP47	Mechanic And Repair Technologies/Technicians.
PCIP22	Legal Professions And Studies.	PCIP48	Precision Production.
PCIP23	English Language And Literature/Letters.	PCIP49	Transportation And Materials Moving.
PCIP24	Liberal Arts And Sciences, General Studies And Humanities.	PCIP50	Visual And Performing Arts.
PCIP25	Library Science.	PCIP51	Health Professions And Related Programs.
PCIP26	Biological And Biomedical Sciences.	PCIP52	Business, Management, Marketing, And Related Support Services.
PCIP27	Mathematics And Statistics.	PCIP54	History.

The regressor model that performed the best was the Random Forest ,which had a R2 score of 0.75. The feature importance graph of the independent variables (Figure 6) shows that PCIP14 (Engineering) percentage of degrees is the most important feature in determining the average earnings of students that enrolled in the college.

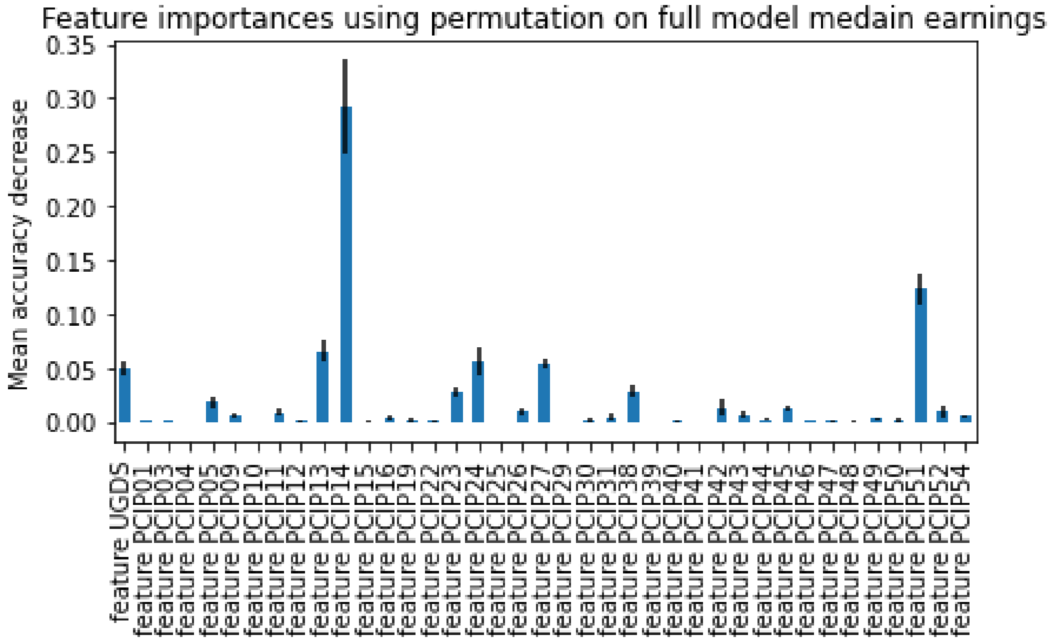

Figure 6

How the average amount of earnings after finishing college and how the college is controlled (Public, Private nonprofit or Private for-profit) can predict a schools school's predominate degree (associate, bachelor's, etc.) awarded

The next analysis was done to classify a schools predominant type of degree it issues based on the median earnings of students not enrolled 10 years after entry and the how the institution is controlled (Public, Private nonprofit, or Private for-profit). Table 3 gives the cross reference for the coding used in the dataset for the predominant type of degree issued and for the Control type.

Table 3

Code	Control		Code	Predominate degree type
1	Public		0	Not classified
2	Private nonprofit		1	Predominantly certificate-degree granting
3	Private for-profit		2	Predominantly associate's-degree granting
			3	Predominantly bachelor's-degree granting
			4	Entirely graduate-degree granting

The classifier model that was had the best performance was the Decision Tree classifier. The accuracy of the Decision tree classifier was .67. Graphing (Figure 7) the feature importance from the independent variables shows both the earnings and control type are close in relevance in predicting the predominate type of degree that a college will issue.

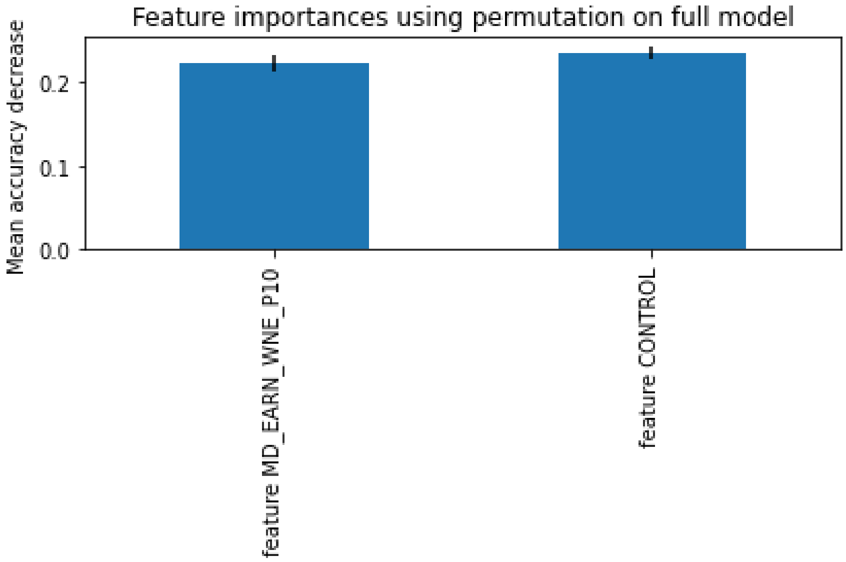

Figure 7

A visualize the Decision Tree classification(Figure 8) was done to give an example of the model. Only the first 3 levels were show because of the complexity of the model and so that one could get an idea of how the Decision Tree process the information to get its prediction.

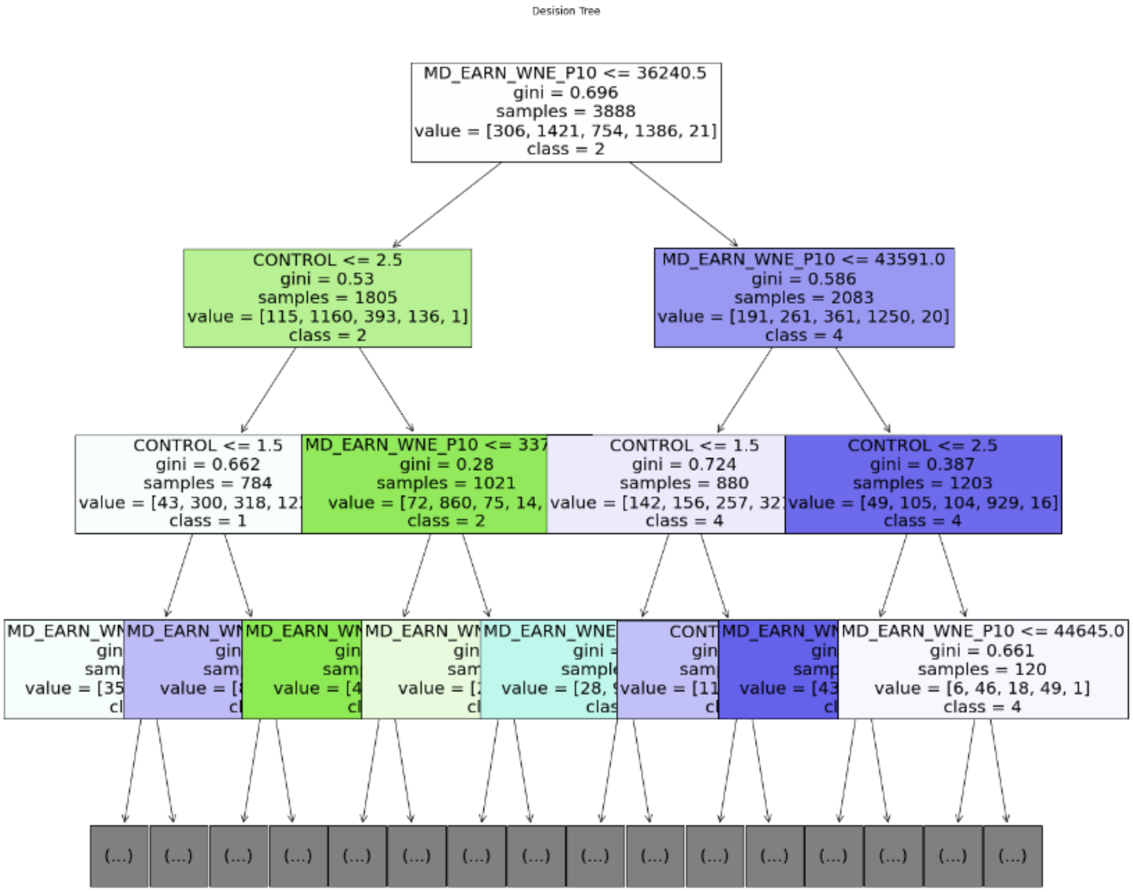

Figure 8

How the amount of debt a student has after leaving college can predict a school's predominate degree (associate, bachelor's, etc.) type

While the previous analysis did produce a model to predict the classification of a schools predominate degree, the overall score of the degree was not very impressive. Therefore, another analysis was done to classify the school's predominant type of degree it will issue. This time the analysis was based on the average debt of the student after leaving school. The debt was categorized into six different categories. Table 4 gives the cross reference for the coding in the dataset for the different debt categories.

Table 4

Code	Debt
DEBT_MDN	The median original amount of the loan principal upon entering repayment
LO_INC_DEBT_MDN	The median debt for students with family income between $0-$30,000
MD_INC_DEBT_MDN	The median debt for students with family income between $30,001-$75,000
HI_INC_DEBT_MDN	The median debt for students with family income $75,001+
FEMALE_DEBT_MDN	The median debt for female students
MALE_DEBT_MDN	The median debt for male students

The classifier model that performed the best, was the Voting classifier. The models used in the Voting classifier were KNN, SVC, Decision Tree, Naive Bayes, and Random Forest. The accuracy of the Voting classifier was 0.70.

Graphing (Figure 9) the feature importance from the independent variables shows that the original amount of the loan is the major feature importance in classifying a schools predominate degree by the debt of the student.

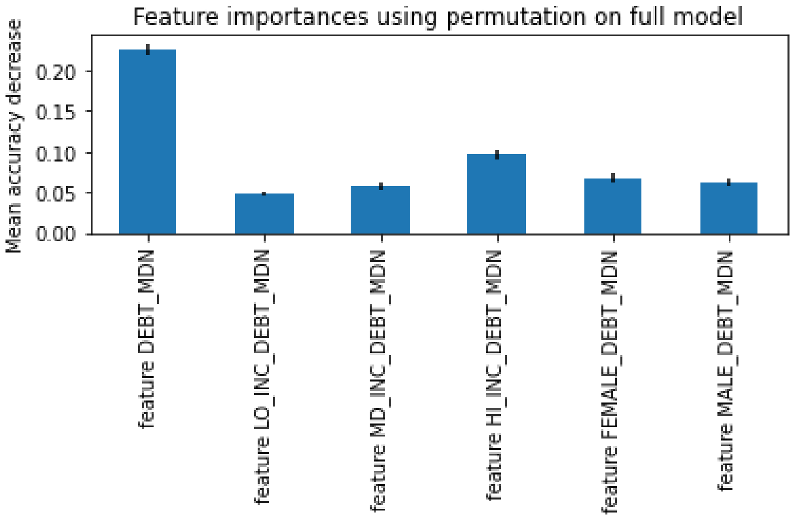

Figure 9

How the racial makeup of the school, enrollment, cost to attend, admission rates, the parents' education and household income can predict what percentage of students received a Pell Grant.

The last analysis focuses in on predicting a 4-year schools percentage of students receiving a Pell Grant. A Pell Grant is a federal grant used to help offset the cost of attending college. They are intended to be used for low-income undergraduates (Federal Pell Grants, n.d.). The factors that are looked at in predicting the percentage receiving a Pell Grant was racial makeup of the school, overall size, the cost to attend and the parents education and household income. Table 5 gives the cross reference for the coding used for the parents educational background.

Table 5

Code	Parents educational background
PAR_ED_PCT_MS	Percent of students whose parents' highest educational level is middle school
PAR_ED_PCT_HS	Percent of students whose parents' highest educational level is high school
PAR_ED_PCT_PS	Percent of students whose parents' highest educational level was is some form of postsecondary education

The regressor model that was chosen to model the data was a linear regression. The R2 value for the model was 0.715. Graphing (Figure 10) the Coefficients of the regression model show that the racial background of the students have the most influence of all the independent variables on determining the percentages of students who received a Pell grant. Figure 11 gives a breakdown of the best fit line scatter plot for each independent variable to show how the data follows the best fit line.

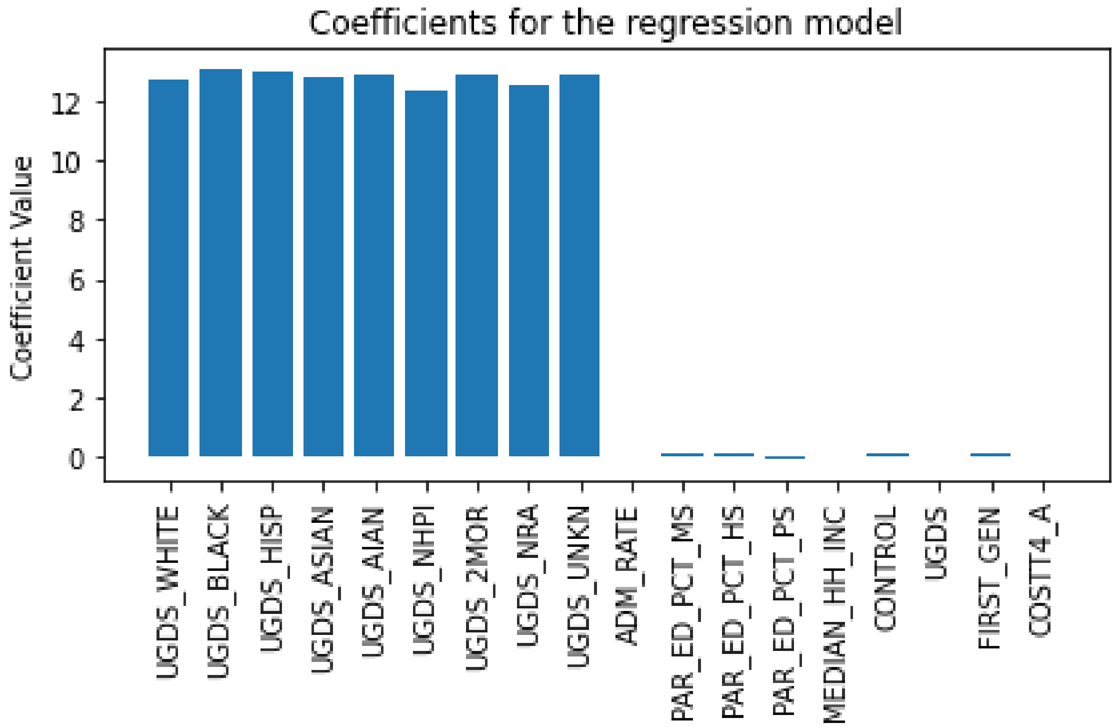

Figure 10

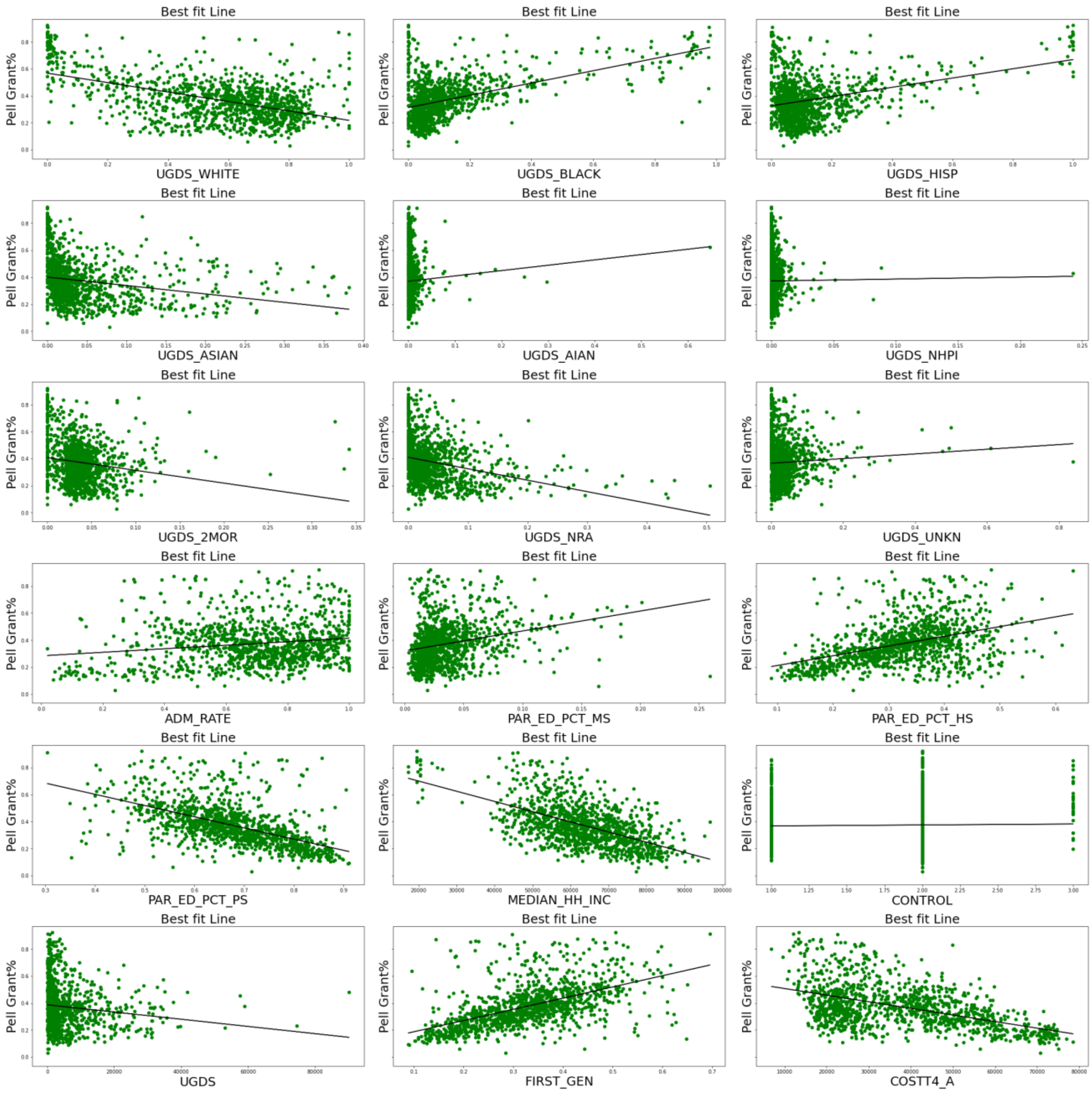

Figure 11

Conclusion

While this project merely scratched the surface on the data included in this College scorecard, one can see from the various analysis, a plethora of insight can be drawn from the data supplied. The models used in the regression and classification analysis have a fairly good score, showing that the prediction made from the model could be used to predict outcomes with relative accuracy.

References

Federal Pell Grants. (n.d.). Retrieved from Benefits.gov: https://www.benefits.gov/benefit/417

Kerr, E. (2020, Janurary 20). *How to Navigate New College Scorecard Data*. Retrieved from U.S. News & World Reports: https://www.usnews.com/education/best-colleges/paying-for-college/articles/how-students-should-use-new-college-scorecard-data

matplotlib v 3.5. (2022). Retrieved from matplotlib: https://matplotlib.org/

pandas v 1.4.2. (2022). Retrieved from pandas: https://pandas.pydata.org/

Scikit-learn v 1.0.2. (2022). Retrieved from Scikit-learn: https://scikit-learn.org/stable/

US Department of Education. (n.d.). *Data Documentation*. Retrieved from US Department of Education College Scorecard: https://collegescorecard.ed.gov/data/documentation/

US Department of Education. (n.d.). *Download Data*. Retrieved from US Department of Education College Scorecard: https://collegescorecard.ed.gov/data/