College Scorecard Analysis (YiHui Ho)

1. Introduction

Many students see college as an investment to help them earn more and live better after graduation. Provided publicly by the U.S. Department of Education's College Scorecard, this project aims to identify factors that may lead to student success post-undergraduate completion. For the scope of this project, the collected data is limited to 4-year, bachelor-granting active colleges, and student success is defined as the average salary students make three years after their college graduation. Large numbers of students today are graduating with worrying amounts of debt, calling into question the assumption that attending college is always the wasted investment. It has become more important than to understand the factors that contribute to post-graduation earnings and the ability to repay student loans.

In this project, I defined a good school as the average salary students make three years after their college graduation is greater than the university's median income three years post-graduation from the data.
My goal in this project will try to answer the following questions:

- Is Duquesne university a good school?
- What factor can we control?
- What is Duquesne's performance vs expectation?
- How can we improve relative to whom?
- How can we control our input in order to improve our output?

2. Data

The Data for this project was sourced form the Department of Education's College Scorecard dataset. For the almost 8,000 colleges included, there are over a thousand fields, including demographics about the students at each college, the degrees and majors offered, the cost and average loans taken out, students test scores, admission rates, and more, matched with statistics for rates of repayment of student loans, and the distributions of graduates incomes over the course of the ten years following graduation.

The original dataset included information for over 7000 institutions and near 3000 attributes. Not all of the data was relevant to my task. This data was preprocessed by removing redundant attributes and ignoring institutions where a majority of the attributes were labeled NULL and PrivacySuppressed. I chose the mean income 3 years after graduation as the response variable and eliminated the many other fields pertaining to post-graduation income, as well as those describing the loan repayment patterns and death rates of graduates. The final data used included around 4500 universities with attribute including: location, admission rates, SAT test scores, total enrollment, demographics, tuition, family income, loan status, and the target class predicting median income three years post-graduation.

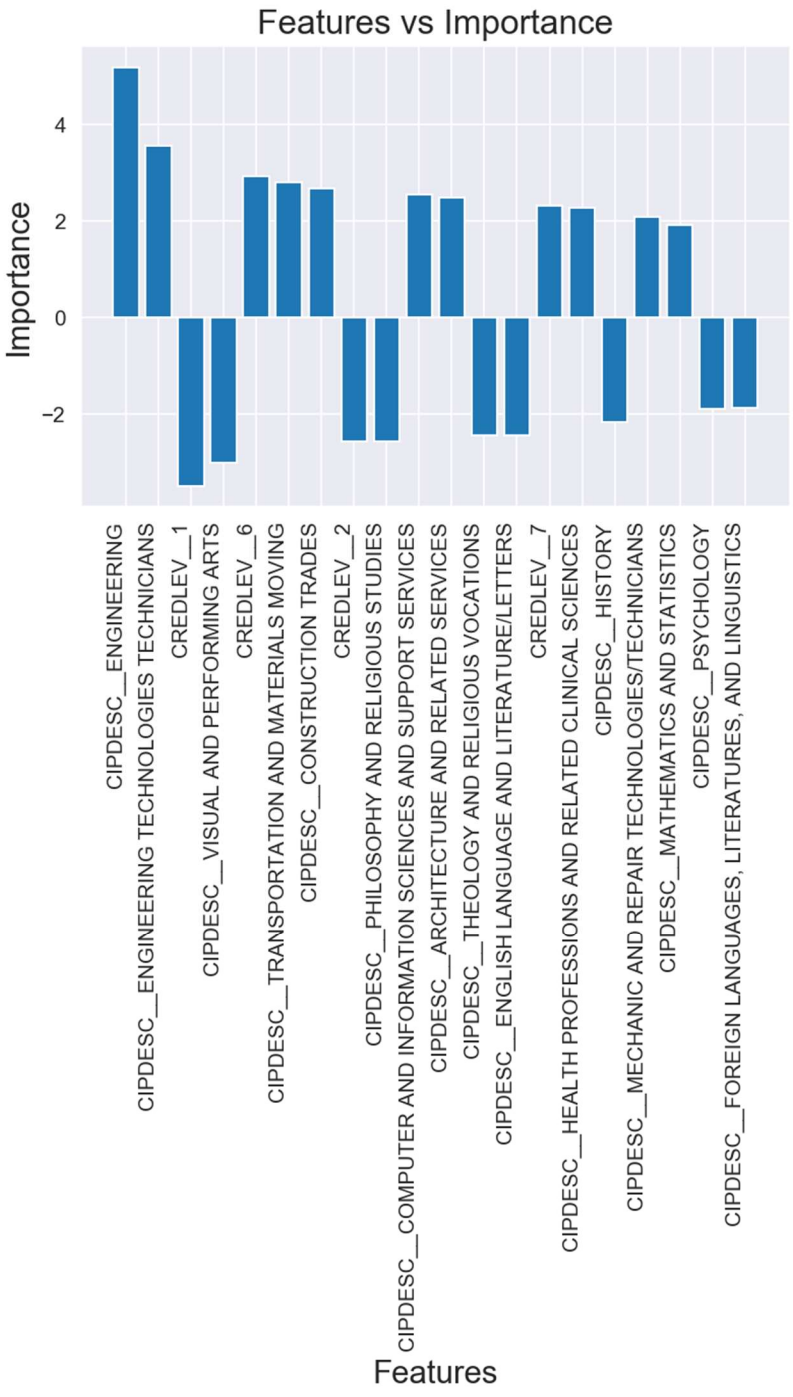
3. Method

In this project, the performance of machine learners were tested on two different tasks: classified good school and bad based on the university's median income three years post-graduation and predicting the student's earnings after graduation 3 years.

To perform both tasks, I used logistic regression and Multinomial Logistic Regression for classification, and I then I moved towards student income prediction with several machine models such as sparse regression, ridge regression, decision tree, random forest as well as gradient boosting.

4. Result

In good or bad school classification part, I obtained Duquesne University is a good school with 82% accuracy with the logistic regression. The important factors to affect whether Duquesne University is a good or bad school are the major student choose to study and the degree students obtain. From the graph below we can see, an Engineering-related major has the highest positive effect and an Undergraduate Certificate or Diploma has the highest negative effect on the result.



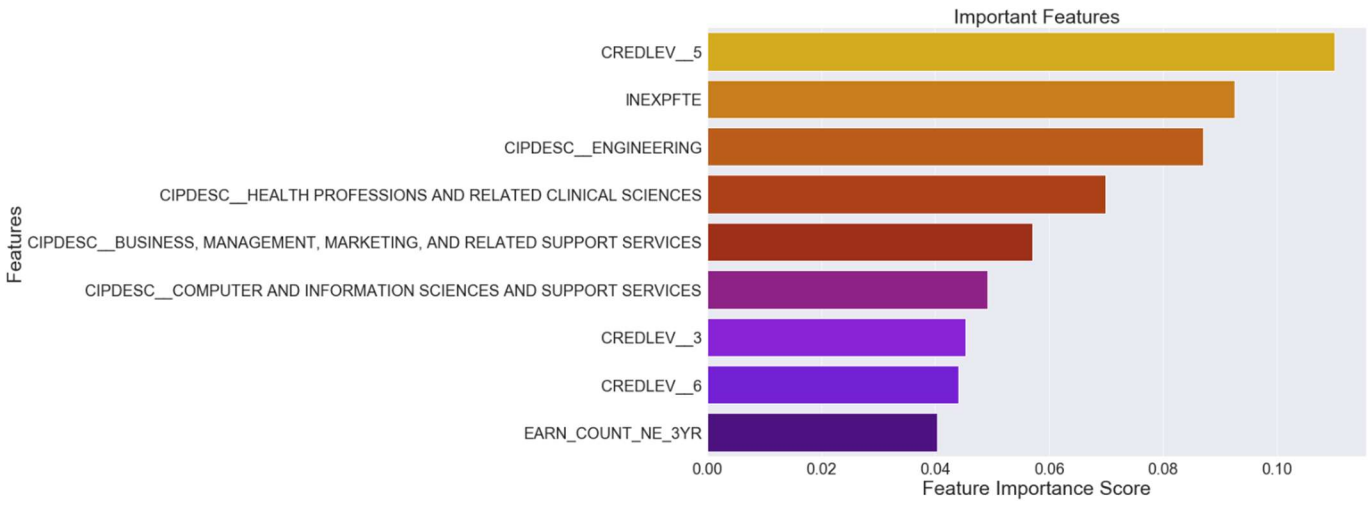
In student's earnings prediction part , the Gradient Boosting algorithm produced the highest accuracy (0.70) and least error ($8021) for the earnings prediction task.

Model	Accuracy	Estimation Error (USD)
Sparse Regression	0.58	8266.5
Ridge Regression	0.57	9052.14
Decision Tree	0.3	9576.21
Random Forest	0.68	8312.29
Gradient Boosting	0.7	8021.34

I predicted Duquesne's income based on different major and I compared the predict income with the actual income.

	Major	Predicted	Actual
0	COMMUNICATION, JOURNALISM, AND RELATED PROGRAMS	14248.0	35708.0
1	COMMUNICATION, JOURNALISM, AND RELATED PROGRAMS	14248.0	44189.0
2	COMPUTER AND INFORMATION SCIENCES AND SUPPORT ...	78751.0	70530.0
3	COMPUTER AND INFORMATION SCIENCES AND SUPPORT ...	78751.0	47725.0
4	COMPUTER AND INFORMATION SCIENCES AND SUPPORT ...	108907.0	55581.0
5	EDUCATION	13945.0	62417.0
6	EDUCATION	13945.0	40825.0
7	EDUCATION	11541.0	40289.0
8	EDUCATION	13945.0	43854.0
9	EDUCATION	11541.0	45493.0
10	EDUCATION	13945.0	45585.0
11	LEGAL PROFESSIONS AND STUDIES	106362.0	68603.0
12	ENGLISH LANGUAGE AND LITERATURE/LETTERS	13657.0	41363.0
13	LIBERAL ARTS AND SCIENCES, GENERAL STUDIES AND...	35899.0	81700.0
14	BIOLOGICAL AND BIOMEDICAL SCIENCES	15433.0	48564.0
15	MATHEMATICS AND STATISTICS	83950.0	54222.0
16	PSYCHOLOGY	31194.0	38884.0
17	PSYCHOLOGY	58853.0	66128.0
18	SECURITY AND PROTECTIVE SERVICES	44560.0	53232.0
19	VISUAL AND PERFORMING ARTS	3083.0	26049.0
20	VISUAL AND PERFORMING ARTS	14377.0	41326.0
21	HEALTH PROFESSIONS AND RELATED CLINICAL SCIENCES	130568.0	57765.0
22	HEALTH PROFESSIONS AND RELATED CLINICAL SCIENCES	75465.0	54249.0
23	HEALTH PROFESSIONS AND RELATED CLINICAL SCIENCES	130568.0	103309.0
24	HEALTH PROFESSIONS AND RELATED CLINICAL SCIENCES	130568.0	41290.0
25	HEALTH PROFESSIONS AND RELATED CLINICAL SCIENCES	94291.0	117556.0
26	HEALTH PROFESSIONS AND RELATED CLINICAL SCIENCES	130568.0	63040.0
27	HEALTH PROFESSIONS AND RELATED CLINICAL SCIENCES	129162.0	75635.0
28	HEALTH PROFESSIONS AND RELATED CLINICAL SCIENCES	74980.0	64219.0
29	HEALTH PROFESSIONS AND RELATED CLINICAL SCIENCES	130568.0	93733.0
30	HEALTH PROFESSIONS AND RELATED CLINICAL SCIENCES	129162.0	96807.0
31	BUSINESS, MANAGEMENT, MARKETING, AND RELATED S...	37651.0	59047.0
32	BUSINESS, MANAGEMENT, MARKETING, AND RELATED S...	109214.0	73486.0
33	BUSINESS, MANAGEMENT, MARKETING, AND RELATED S...	37651.0	58059.0
34	BUSINESS, MANAGEMENT, MARKETING, AND RELATED S...	68007.0	66013.0
35	BUSINESS, MANAGEMENT, MARKETING, AND RELATED S...	37651.0	55343.0
36	BUSINESS, MANAGEMENT, MARKETING, AND RELATED S...	37651.0	65285.0
37	BUSINESS, MANAGEMENT, MARKETING, AND RELATED S...	37651.0	59571.0
38	BUSINESS, MANAGEMENT, MARKETING, AND RELATED S...	37651.0	50018.0
39	HISTORY	41314.0	43725.0
40	HISTORY	47948.0	31023.0

From the graph below, we can see that master's degree, instructional expense, and engineering majors are the top three factors that affect Duquesne students' income.



From the analysis result, adding more engineering-related programs and increasing master's degree options for students are one of the solutions that help Duquesne's performance and let Duquesne become a better school.

5. Conclusion and Future Work

This project focused on predicting postgraduate income given a college and information about that college. This was useful because it allowed us to explore the College Scorecard dataset and determine what features correlated most with postgraduate income. However, I was not able to achieve a prediction accuracy above 70% however, it would be interesting to incorporate more detailed data such student diversity, university location choice and academic performance to see if a more accurate income predictor can be made.

Appendix A

The following are some features the College Scorecard dataset included for over 7000 different institutions.

Feature Name	Feature Definition
CREDLEV_1	Undergraduate Certificate or Diploma
CREDLEV_2	Associate degree
CREDLEV_3	Bachelor's Degree
CREDLEV_4	Post-baccalaureate Certificate
CREDLEV_5	Master's Degree
CREDLEV_6	Doctoral Degree
CREDLEV_7	First Professional Degree
INEXPFTE	Instructional expenditures per full-time equivalent student