Predicting Postgraduate Income

Northwestern University EECS 349

By: Murphy Angelo, Jared Fernandez, Isaac Lee
Contact: jared.fern@u.northwestern.edu, isaaclee2019@u.northwestern.edu, mca@u.northwestern.edu

i. INTRO

We used machine learning to explore data on American colleges to predict post-graduate earnings. There are hundreds of different features that make each college unique, including: location, admission rates, tuition, demographics, average student loans, and average standardized test scores. We want to discover which of these features are most significant in predicting the median ten year post-graduate earnings for a given college.

ii. APPROACH

The data for this project was sourced from the Department of Education's CollegeScorecard dataset. The original dataset included information for over 7,000 institutions and hundreds of attributes. This data was preprocessed by removing redundant attributes and ignoring institutions where a majority of the attributes were labeled NULL. The final dataset used included 4,682 universities with attributes including: location, admission rates, SAT test scores, total enrollment, demographics, tuition, family income, loan status, and the target class predicting median income ten years post-graduation (Md10yr). The data was grouped into four classes, splitting on each quartile of Md10yr.

In this project, the performance of machine learners were tested on two different tasks: predicting the university's median income ten years post-graduation with regression models, and predicting the university's Md10yr quartile by classifying into the four classes using classifiers. All estimators were tested using the SciKit Python package.

To perform both tasks, we used several base estimators as well as multiple methods of ensembling these base methods togethers. For the classification task, we tested the accuracy of the following base classifiers: Decision Tree, 1/3/5-Nearest Neighbor, Gaussian Naive Bayes, Adaboost, Random Forest, Recurrent Neural Net, Gradient Boosting, and Logistic Regression. Additionally, we tested an ensemble method, combining all the previous classifiers and assigning class using majority vote. For the regression task, the same base algorithms were used. However, the Naive Bayes classifier was substituted with a linear regression model. Multiple ensembling methods were attempted in the regression analysis, using the stacking process for ensembling. In this process, the results from all of the base regressors were used as inputs to a meta-regressor, which produced the final estimation based on the outputs of each base regressor. Each base regressor was tested as a meta-regressor.

iii. RESULTS

The Gradient Boosting algorithm produced the highest accuracy (0.70) and least error ($8022.33) for both the classification and estimation tasks. Additional analysis shows that none of the attempted ensemble techniques produced significant improvements over any of the base estimators. This suggests

that all of the base estimators are modeling a similar hypothesis function, because there is no difference between any single estimator and a combination of base estimators.

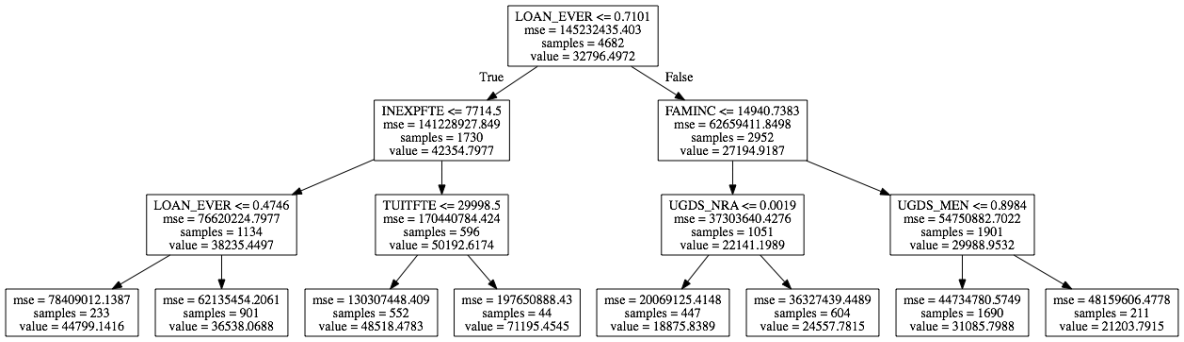
	Classification Accuracy:	Estimation Errors (USD):
ZeroR, Average	0.29	12051.24
Decision Tree	0.61 +/- 0.07	9586.21
1-Nearest Neighbor	0.61 +/- 0.07	10219.39
3-Nearest Neighbor	0.63 +/ 0.10	9180.47
5-Nearest Neighbor	0.65 +/- 0.09	9052.20
Adaboost	0.66 +/- 0.09	9993.38
Random Forest	0.68 +/- 0.14	8411.28
Recurrent Neural Net (MLP)	0.49 +/- 0.10	8346.04
Gradient Boosting	0.70 +/- 0.12	8022.33
Linear Regression	----------	8266.40
Gaussian Naive Bayes	0.58 +/- 0.16	----------
Ensemble: Majority Vote	0.69 +/- 0.11	----------

Table 1. Base Estimator Accuracies and Errors

Meta-Regressor	Estimation Error (USD)
Decision Tree	9629.31
1-Nearest Neighbor	8776.35
3-Nearest Neighbor	8659.53
5-Nearest Neighbor	8650.37
Adaboost	9599.44
Random Forest	9057.69
Recurrent Neural Net (MLP)	9027.13
Gradient Boosting	9039.57

Table 2. Ensemble Regressor Errors

From our decision tree, it can be determined that the most important features are: percentage of students that took out loans, amount of university spending per student, and average student family income.



A decision tree that shows the most important features when predicting Md10yr.

iv. CONCLUSIONS AND FUTURE WORK

Our 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. We were not able to achieve a classification accuracy above 70% however. It would be interesting to incorporate more detailed data such as major/minor choice and academic performance to see if a more accurate income predictor can be made.

v. WORK DISTRIBUTION

We worked on the majority of the project as a group. We spent a lot of time early deciding the different classifiers we wanted to test. In particular, Jared spent some time testing ensembling methods while Isaac and Murphy tested other classifiers and regressors.

Appendix A

The following are the features the College Scorecard dataset included for over 7,000 different institutions.

Feature Name	Feature Definition
LATITUDE	Geographical latitude
LONGITUDE	Geographical longitude
ADM_RATE	Admission rate
ADM_RATE_ALL	Admission rate for all campuses
SAT_AVG	Average SAT Score
UGDS	Undergraduate Enrollment
UGDS_WHITE	Share of undergraduate degree-seeking students who are white
UGDS_BLACK	Share of undergraduate degree-seeking students who are black

UGDS_HISP	Share of undergraduate degree-seeking students who are Hispanic
UGDS_ASIAN	Share of undergraduate degree-seeking students who are Asian
UGDS_AIAN	Share of undergraduate degree-seeking students who are American Indian/Alaska Native
UGDS_NHPI	Share of undergraduate degree-seeking students who are Native Hawaiian/Pacific Islander
UGDS_2MOR	Share of undergraduate degree-seeking students who are two or more races
UGDS_NRA	Share of undergraduate degree-seeking students who are non-resident aliens
UGDS_UNKN	Share of undergraduate degree-seeking students whose race is unknown
UGDS_WHITENH	Share of undergraduate degree-seeking students who are white non-Hispanic
UGDS_BLACKNH	Share of undergraduate degree-seeking students who are black non-Hispanic
UGDS_API	Share of undergraduate degree-seeking students who are Asian/Pacific Islander
UGDS_AIANOLD	Share of undergraduate degree-seeking students who are American Indian/Alaska Native
UGDS_HISPOLD	Share of undergraduate degree-seeking students who are Hispanic
UG_NRA	Share of undergraduate students who are non-resident aliens
UG_UNKN	Share of undergraduate students whose race is unknown
UG_WHITENH	Share of undergraduate students who are white non-Hispanic
UG_BLACKNH	Share of undergraduate students who are black non-Hispanic
UG_API	Share of undergraduate students who are Asian/Pacific Islander
UG_AIANOLD	Share of undergraduate students who are American Indian/Alaska Native
UG_HISPOLD	Share of undergraduate students who are Hispanic
TUITIONFEE_IN	Tuition for in-state full-time students
TUITIONFEE_OUT	Tuition for out-of-state full-time students

TUITFTE	Net revenue per full-time equivalent student
INEXPFTE	Instructional expenditures per full-time equivalent student
AVGFACSAL	Average faculty salary
PFTFAC	Proportion of faculty that is full-time
PCTPELL,PCTFLOAN	Percent of students who received Pell Grants in a given year
COUNT_ED	Count of students in the earnings cohort
LOAN_EVER	Share of students who received a federal loan while in school
PELL_EVER	Share of students who received a Pell Grant while in school
AGE_ENTRY	Average age of entry, via SSA data
AGEGE24	Percent of students over 23 at entry
FEMALE	Share of female students via SSA data
MARRIED	Share of married students
DEPENDENT	Share of dependent students
VETERAN	Share of veteran students
FIRST_GEN	Share of first-generation students
FAMINC	Average family income in real 2015 dollars
MD_FAMINC	Median family income in real 2015 dollars
FAMINC_IND	Average family income for independent students in real 2015 dollars
PCT_WHITE	Percent of the population from students' zip codes that is White
PCT_BLACK	Percent of the population from students' zip codes that is Black
PCT_ASIAN	Percent of the population from students' zip codes that is Asian
PCT_HISPANIC	Percent of the population from students' zip codes that is Hispanic
PCT_BA	Percent of the population from students' zip codes with a bachelor's degree over the age 25
PCT_GRAD_PROF	Percent of the population from students' zip codes over 25 with a professional degree

PCT_BORN_US	Percent of the population from students' zip codes that was born in the US
MEDIAN_HH_INC	Median household income
POVERTY_RATE	Poverty rate, via Census data
UNEMP_RATE	Unemployment rate, via Census data
LN_MEDIAN_HH_INC	Log of the median household income
DEBT_MDN_SUPP	Median debt, suppressed for n=30
GRAD_DEBT_MDN_SUPP	Median debt of completers, suppressed for n=30
GRAD_DEBT_MDN10YR_SUPP	Median debt of completers expressed in 10-year monthly payments
UGDS_MEN	Share of undergraduate degree-seeking students who are men
UGDS_WOMEN	Share of undergraduate degree-seeking students who are women
MEDIAN_HH_INC2	Median household income