Predicting Median Post Graduate Earnings

Arun Babu (110338501) Shreya bhatia (110619150)
Vijay Teja (110622613)

September 3, 2016

1 Motivation

Our goal is to build a model that would accurately predict the earnings of a college's graduates after ten years based on the specific features of the college like average cost of attendance, acceptance rate, average test scores etc. Because of the increasing cost of higher education and the accompanying rise of student debt, it is important to understand the factors that contribute to the post graduation earnings and ability of the students to repay loans. This model will increase the transparency and will help students to weigh the trade-offs of different colleges and make more informed decisions.

2 Background

We are using the College Scorecard Data [1] curated by US Department of Education. This dataset which spans nearly 20 years, contains approximately 2000 features for 7805 degree-granting institutions. These features include different information like demographic data, test scores, family income data, data about the percentages of students in each major, financial aid information, debt and debt repayment values, earnings of alumni several years after graduation.

3 Data preprocessing

The dataset contains information of about 8000 colleges, and for each college there are around 2000 features. But the data is also sparse, containing a large number of missing and privacy suppressed values. So we filtered out the columns whose fraction of bad (missing or privacy suppressed) values is greater than a certain threshold. Even though we have data from 1996 to 2013, we are using data from year 2011 as it is of the largest size.

4 Discovery

4.1 Discovering significant features

Since we have a huge number of features, it is unlikely that all of them would affect the post-graduation earnings. So to see which are the most significant features for predicting the post-graduation earnings, we do feature selection. We use forward stepwise selection to select top 100 features and then run a least squares ordinary linear regression on the post-graduate earnings. After getting the beta values from the regression model, we test the significance of each individual predictor.

Since we are performing multiple hypothesis tests, we use Benjamini-Hochberg correction to correct for the false discovery rates. We use alpha=0.05 for the significance test. Below is the table for the top 5 and bottom 5 significant features by p-values

Top significant 5 features

Feature	Coefficient	Description
PCIP12	-7038.52	Percentage of degrees awarded in Personal And Culinary Services.
UGDS_ASIAN	24312.37	Total share of enrollment of undergraduate degree-seeking students who are Asian.
IND_INC_PCT	-14787.35	Percentage of students who are financially dependent and have family incomes between $0-30,000.
HCM_2	13898.49	Schools that are on Heightened Cash Monitoring 2 by the Department of Education.
PCIP50	-7851.42	Percentage of degrees awarded in Visual And Performing Arts.

Least significant 5 features

Feature	Coefficient	Description
CIP25BACHL	1642.96	Bachelor's degree in Library Science.
CIP51ASSOC	-489.81	Associate degree in Health Professions And Related Programs.
COMP_ORIG_YR4_RT	-506.79	Percent completed within 4 years at original institution
STABBR_IN	-972.43	State postcode Indiana
STABBR_DE	-2829.25	State postcode Delaware

5 Prediction Models

We use four different machine learning models in our analysis for predicting the target variable(the median income 10 years after joining the institution).
We divide the entire dataset into training and test set in the ratio of 7:3. We train the model on training data and evaluate the model on the test data. We use RMSE as the metric for evaluating the performance of the model.

5.1 Baseline model

Our baseline model is the MLE estimate for the earnings, which is nothing but the mean.

5.2 Linear Regression

In simple linear regression we model the relationship between a single independent variable and a target variable. In this analysis, since we have multiple candidate features affecting the target variable, we use a multivariate linear regression model.

5.3 Regularization

One of the major drawbacks in using a linear regression model for high dimensional data, is the complexity of the resulting model: due to both large coefficients, and non-zero coefficients for many irrelevant features. Due to this high complexity, the model is more prone to over-fitting and poor generalization.

5.3.1 Lasso Regression

Lasso (Least Absolute Shrinkage and Selection Operator) overcomes the above problems by adding a L1 norm penalty term to the objective function, and thus forcing the coefficient of irrelevant features to go to zero, and also regularizing the high coefficient values. This way Lasso helps in both feature selection and regularization.

5.3.2 Ridge Regression

Ridge regression is another way of regularization done by adding a L2 norm penalty term to the objective function. It helps in avoiding large coefficient values for features and thus helps in building simpler models.

5.4 Principal Component Analysis

Principal Component Analysis is a method of dimensionality reduction where data is transformed into a new k dimensional subspace where the basis vectors for the subspace are the top k eigenvectors of the original data. Due to the high dimensional nature of our data, PCA might be a good pre-processing step to handle the prediction task.

6 Utilization of course topics

6.1 Probability

The derivation for linear regression is based on MLE estimation i.e. maximizing probability of data, given parameters.

6.2 Discovery

We use forward stepwise algorithm to select top k features and then run linear regression to get beta values for each of the individual predictor. We perform multiple test correction on each of the predictor using Benjamini–Hochberg correction.

6.3 Prediction

We deploy several regression models, including regularized regression models, for predicting the earnings. We use test-train split for training and testing, and perform cross validation for searching of optimal hyper parameters like alpha value, in regularized regression.

7 Evaluation and Results

Below are the results for the prediction models we built. Our models performed significantly better than the baseline model.

Results with bad items threshold as 0.25:

Model	Hyper parameter	RMSE
Baseline	NA	9082.38
Linear Regression	NA	4178.33
Ridge Regression	256	4203.7
Lasso	64	4279.01
PCA with LR	256	4404.58
PCA with ridge	256	4267.38

Results with bad items threshold as 0.125:

Model	Hyper parameter	RMSE
Baseline	NA	11806.49
Linear Regression	NA	5451.63
Ridge Regression	256	5408.6
Lasso	64	5388.03
PCA with LR	256	5520.6
PCA with ridge	256	5477.62

8 Conclusion

We used College Scorecard dataset to predict the median earnings 10 years after graduation. The dataset had large number of missing values so we did preprocessing and reduced the number of features and colleges. We found the most significant features for predicting median earnings using multiple hypothesis testing and benjamini hochberg correction. For prediction, we used linear regression and penalized linear regression (Lasso and Ridge) methods.

References

[1] https://collegescorecard.ed.gov/data/