Stat 159 - Project 3

Kartik Gupta
Lauren Hanlon
Morgan Smart
Dakota Lim

November 20, 2016

Abstract

In this paper we will consider the data provided in collegescorecard.ed.gov to consult on the following issue:

"The CEO of a biotech startup is looking for candidates. She is interested in diversifying the workforce in regards to women in STEM (Science, Technology, Engineering, and Math). Where should the startup focus their recruitment and outreach for maximum impact?"

To ensure we use the newest data on recent college graduates we will only consider the most recent datafiles, ignoring older data. Next we will create a scoring system to rank colleges based on where we believe our client will have the most success in recruiting top tier female STEM majors.

Intrduction

We begin by cleaning the data. Since our raw data file contains hundreds of features, we create a new data matrix with only the desired predictors (see the Data section for further details). We also create several new features by combining columns, which imroves the regression model's interpretability by creating more consise predictors. All this is done through the `code/scripts/data-processing-script.R` script.
Next we consider a cross-validated linear regression to predict the mean earnings of female students working and not enrolled 6 years after entry

(MN_EARN_WNE_MALE0_P6) using the new data matrix, which we will use to create a scoring system for all schools. We use the mean earnings as a metric to represent the overall quality of a college's graduates, the intuition being that smarter more qualified candidates will be paid more. After preforming the regression, we check that all features are significant and use the weights of all significant terms.
We then use the regression equation, with only the significant terms included, to create a score for each school. After giving each collge a score, we will rank them by said score and present the top institutions to our client.

Data

"HIGHDEG4"
For our consultation we consider the following predictors from the scaled data set produced from the 'data-processing-script.R':

1. $MN\_EARN\_WNE\_MALE0\_P6$ = Mean earnings of female students working and not enrolled 6 years after entry
2. $SATMTMID$ = Median SAT math score
3. ADM_RATE = Admission rate
4. $STEM\_DEG\_WOMEN$ = Approximate number of female STEM majors
5. $WOMENOONLY$ = Indicator if schools that are only femal
6. $HIGHDEG4$ = Indicator if the school offers Bachelors degrees
7. $COUNT\_WNE\_MALE0\_P6$ = Number of female students working and not enrolled 6 years after entry

These predictors will be used to create a regression model predicting and, ultimately, create a score for each school.

Methods

In order to properly consult our client, we build score function and apply it to each school, recommending the top institutions to our client. To create the score function we use ordinary least squares regression (OLS) and cross validation.

OLS is an approach to predicting a quantitative response Y based on a multiple predictor variables X_1 through X_p, where Y and X_1 through X_p are vectors and each value in each X_{ij} (x_{ij}) has a corresponding value in Y (y_i). The model assumes that the relationship between every X_i and Y is linear and that each X_j isn't correlated with any other X_j. OLS can be written as $Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon$, where β_0 is the intercept and β_1 through β_p are the slopes of their corresponding predictor variable X_j. The beta values are all constants, unknowns, and together are the model coefficients. The interpretation of β_0 is the expected mean value of Y without a predictor variable and the interpretation of β_1 through β_p is the change in Y for a unit increase in the beta's corresponding X_j. Although the betas are unknown, we can estimate them using the OLS model: solving for the intercept and slopes that produce the plane closest to each point (x_{ij},y_i) in each X_j,Y, which is minimizing the residual sum of squares (RSS). Once we have an estimate for the betas, we can compute using OLS to determine the strength of the relationship between each X_j and Y, if the relationship is statistically significant, and asses how accurately the model predicts the relationship.

Cross-validation is a model validation technique used to assess how the results of an analysis will generalize to an independent data set. To cross-validate a model, we fit its required estimated parameters using k parts of a training dataset and then test the k created models on an unseen test dataset. We then chose the model that minimizes the predictive error. If the model performs well using the test dataset (meaning it has a low predictive error), we are more confident that this model is accurate (since it produces good prediction of data not used to fit the model). For our purposes, we will cross validate the resulting model from our OLS regression and choose the β vector that has the minimal cross-validated mean squared error.

Analysis

After running our regression script we observed the following values for our $\hat{\beta}_{OLS}$ estimator:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.00	0.02	-0.00	1.00
ADM_RATE	-0.07	0.03	-2.77	0.01
SATMTMID	0.61	0.03	21.75	0.00
STEM_DEG_WOMEN	-0.15	0.03	-5.05	0.00
COUNT_WNE_MALE0_P6	0.15	0.03	5.44	0.00
WOMENONLY	0.06	0.02	2.49	0.01
HIGHDEG4	0.21	0.02	8.55	0.00

Table 1: Predictor Significance

It's clear from the final column of the tabel that all the regressors are significant (with the exception of the intercept). Thus, we will utilize all values *except* the intercept in our score function.

Results

From our 10-fold cross validation we observe a mean squared error of: 0.6064

Next we build the score function, utilizing the above $\hat{\beta}_{OLS}$ values as the weights for our function. This yields the following expression:

$f(X_{.,i}) \coloneqq$ (-0.0715) $X_{1,i}$ + 0.6094 $X_{2,i}$ + (-0.1537) $X_{3,i}$ + 0.1469 $X_{4,i}$ + 0.0582 $X_{5,i}$ + 0.2076 $X_{6,i}$

Where $X_{j,i}$ represents the j-th regressor value from our OLS regression of the i-th observation in the data matrix. Using this equation, we apply it to all schools in our data matrix and sort by their relative score to yield the top 20 colleges we recomend the client recruits from:

SCORE	INSTNM	CITY
2.14	Massachusetts Institute of Technology	Cambridge
2.06	University of Chicago	Chicago
2.04	Washington University in St Louis	Saint Louis
2.04	Yale University	New Haven
2.03	Vanderbilt University	Nashville
2.02	Harvard University	Cambridge
2.01	Princeton University	Princeton
1.99	Rice University	Houston
1.98	Columbia University in the City of New York	New York
1.96	Northwestern University	Evanston
1.94	Carnegie Mellon University	Pittsburgh
1.93	Stanford University	Stanford
1.93	Duke University	Durham
1.92	Claremont McKenna College	Claremont
1.91	University of Pennsylvania	Philadelphia
1.85	Johns Hopkins University	Baltimore
1.80	Dartmouth College	Hanover
1.79	Williams College	Williamstown
1.78	Brown University	Providence
1.75	University of Notre Dame	Notre Dame

Table 2: Top 20 Colleges to Recruit From

Conclusions

After considering the request of our client, we filtered the data, built an OLS regression model, considered the significance of the coefficients, and built a comprehensive scoring function for any given college. Using this we were abel to recommend the following 20 universities to recruit from:
It should be noted that the structure of this paper allows for it to be reproduced with each new data upload on collegescorecard.ed.gov