Machine Learning Engineer Nanodegree

Definition

Project Overview

Predicting student future earnings based on college.

Can publicly available data help students determine best college selection based on future potential earnings. Choosing a college to attend can be difficult task given vast options available. This choice is ever more important given that the cost of tuition in private colleges is over 50,000 per year. Do private colleges with high tuition cost yield highest returns on investment?
Given student success in high school and their performance on SAT what schools are best suited for their future success?
In this paper explore how various college characteristics influence future student earnings. Do students attending private colleges have an advantage in future earning over public colleges, or is the situation reversed given that students in public colleges have much lower student loans.
This project tries to determine if there is a significant difference in potential earning between public and private colleges. I also try to determine which factors most contribute to future earnings of college graduates based on earnings 10 years after graduations.

Hypotheses

My hypothesis is that we can predict future income 6 years after graduation based on input data namely SAT score of the student, Type of Degree, etc.

Solution

To solve this problem of predicting future student income I am going to use supervised learning regression classifiers. Classifier will be able to predict future earnings given input data. To pick best algorithm I am going to be using Mean Square Error and Accuracy score.

The goal is to analyze college score-card data provided through US Department of Education. Data is provide for period of 1996-2014 :

1. Download and preprocess the DOE Score Card Text data. Import data into SQL Database
2. Using classifier explore data to determine important features
3. Visualize important features to verify correlation
3. Train different classifiers measuring predictive accuracy on Test data
4. Evaluate results and further refine features
5. Pick most suitable classifier for dataset.

The final application is expected to predict future income based on input data.

Metrics
Several metrics are used thought-out the analyses and modeling including:

- Recursive feature elimination (RFE) for feature selection which selects features by recursively considering smaller and smaller sets of features. RFE outputs list of features according to their predictive importance (of future earnings)
- Lasso as a regression method penalizes extra variables producing list of coefficients for each variable. Higher the coefficients, higher is the importance of the variable

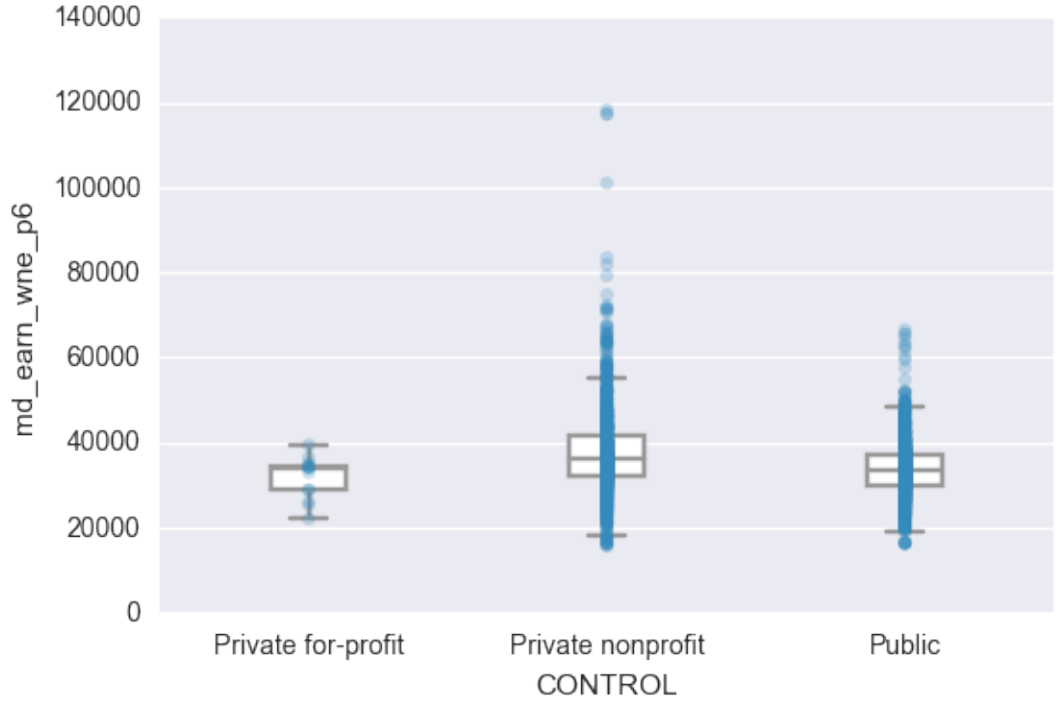
- For predictive leaner models I used Mean squared error, measuring distance or difference of prediction of data the model has not seen (test data) and true test data.

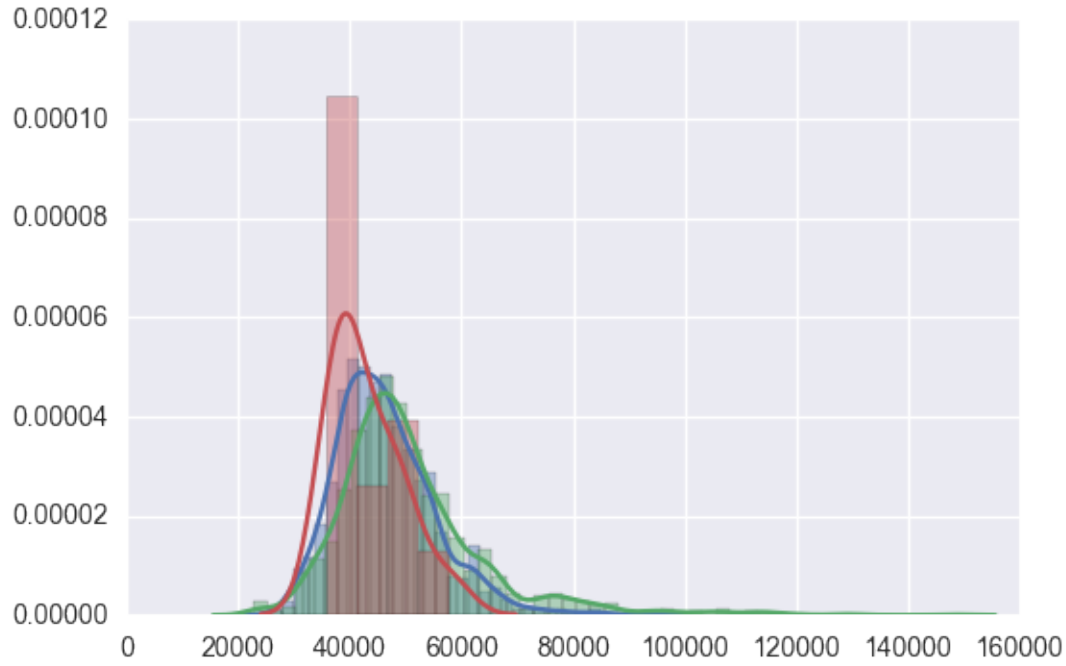
- Accuracy score measuring percentage of correct predictions.

Exploratory Data Analyses and Visualization

Starting with median income 6 years after graduation I broke it down in tow graphs. First r showing distribution of incomes across different school types.
Second graph shows density graph of salaries, we can see that then salaries are normally distributed with long tail on the right. Most notable is distribution of private for profit collages which report high number of students with the same salary values. We can see this from the boxplot as well that there is not a lot of variation in data reported for private for-profit colleges.

Dataset consists 28281 of rows across 1731 columns, out of this 397 are numerical. To work with regression, I have filtered data set to 3813 non null values and 31 columns.





Algorithms and Techniques

Data Preprocessing
Because linear models don't perform well with missing data or data on different scales pre-processing is necessary.
The preprocessing done in the "Feature Selection" notebook consists of the following steps:

1. Filter out all the “Privacy Suppressed” and null values
2. Scale the data around mean and calculate distance from mean

Future Selection

To Evaluate useful variables, I created a matrix of regressors and measured feature importance.

- Corr F-regression just computes the F statistic and pick the best features. F-regression does the following:
  Start with a constant model, M0
  Try all models M1 consisting of just one feature and pick the best according to the F statistic
  Try all models M2 consisting of M1 plus one other feature and pick the best ...
- Recursive Feature Elimination (RFE) Recursive feature elimination is based on the idea to repeatedly construct a model (for example an SVM or a regression model) and choose either the best or worst performing feature (for example based on coefficients), setting the feature aside and then repeating the process with the rest of the features. This process is applied until all features in the dataset are exhausted. Features are then ranked according to when they were eliminated. As such, it is a greedy optimization for finding the best performing subset of features. I used following models with RFE
  - Linear Regression
  - Ridge Regression
  - Lasso
  - Randomized Lasso

- Random Forest feature importance

	Corr.	Lasso	Linear reg	RF	RFE	Ridge	Stability
ADM_RATE	0.09	0.15	0.75	0.03	0.3	0.75	1
SAT_AVG	0.87	0.51	0.04	0.17	1	0.04	1
TUITFTE	0.59	0.58	0.13	0.07	1	0.13	1
UGDS	0.03	0.13	0.88	0.1	0.15	0.88	1
UGDS_NRA	0.07	0.06	0.96	0.04	0.05	0.96	1
PPTUG_EF	0.02	0.13	0.79	0.04	0.25	0.79	1
UG25abv	0.04	0	1	0.06	0	1	0.6
PAR_ED_PCT_1STGEN	0.36	0.38	0.08	0.08	1	0.08	1
DEP_INC_AVG	1	0.59	0	1	1	0	1
IND_INC_AVG	0.13	0.2	0.54	0.2	0.55	0.54	1
COMP_ORIG_YR2_RT	0.28	0.23	0.67	0.09	0.4	0.67	1
WDRAW_ORIG_YR2_RT	0.41	0.28	0.46	0.1	0.65	0.46	1
ENRL_ORIG_YR2_RT	0.42	0.18	0.71	0.14	0.35	0.71	1
COMP_ORIG_YR4_RT	0.79	0.63	0.17	0.55	1	0.17	1
WDRAW_ORIG_YR4_RT	0.11	0.34	0.42	0.04	0.7	0.42	1
ENRL_ORIG_YR4_RT	0.12	0.54	0.21	0.03	0.95	0.21	1
OVERALL_YR2_N	0	0.16	0.58	0.02	0.5	0.58	0
OVERALL_YR3_N	0	0.27	0.63	0	0.45	0.63	0.68

OVERALL_YR4_N	0	0.34	0.5	0	0.6	0.5	1
OVERALL_YR6_N	0	0.81	0.29	0.01	0.85	0.29	1
OVERALL_YR8_N	0	1	0.25	0.02	0.9	0.25	1
count_nwne_p6	0.04	0.61	0.33	0.05	0.8	0.33	1
DEBT_MDN	0.12	0.1	0.92	0.02	0.1	0.92	0.94
GRAD_DEBT_MDN	0.02	0.46	0.37	0.1	0.75	0.37	1
WDRAW_DEBT_MDN	0.05	0.17	0.83	0.01	0.2	0.83	1

From the list of columns and its scores I see some interesting observations:

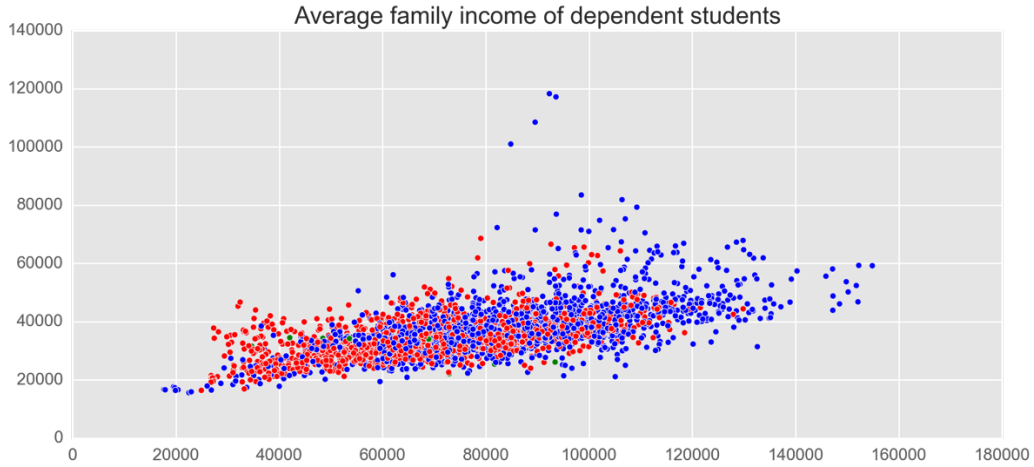
- Stability. Using Ridge regression for data interpretation due to its stability and the fact that useful features tend to have non-zero coefficients.

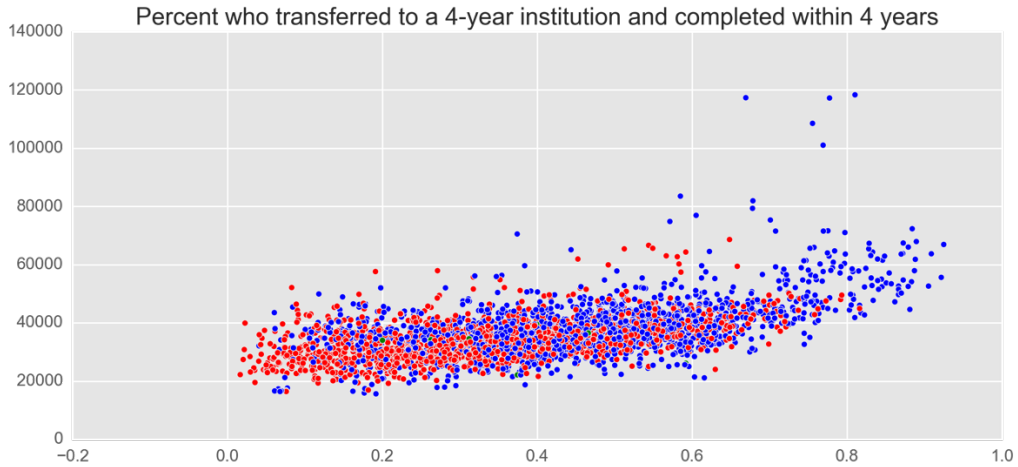
Following features have low stability (coefficients close to 0) and can be removed:

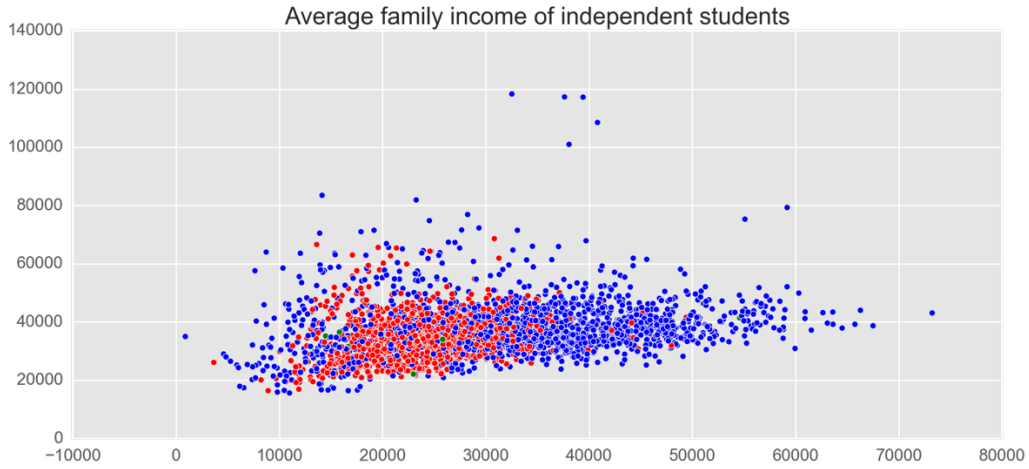
- Percentage of undergraduates aged 25 and above (UG25abv)
- Number of students in overall 3-year completion cohort(OVERALL_YR3_N)

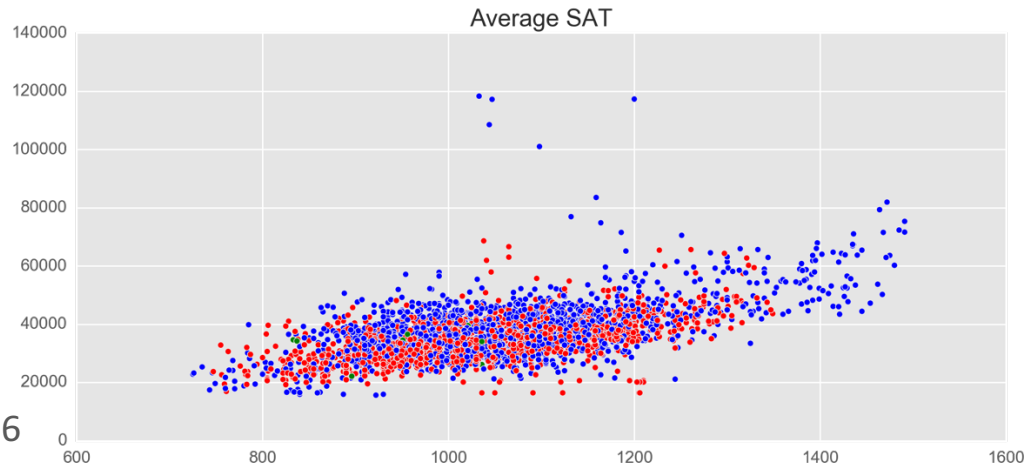
- In previous tests I observed that Random Forests have the best predictive power measured by Median SQ Error MSE. Using Random Forest Feature Importance I can see that following Features are most important: Average family income of dependent students in real 2014 dollars. DEP_INC_AVG
- Percent who transferred to a 4-year institution and completed within 4 years COMP_ORIG_YR4_RT
- Average family income of independent students in real 2015 dollars. IND_INC_AVG
- Average SAT equivalent score of students admitted SAT_AVG
- Percent still enrolled at original institution within 2 years ENRL_ORIG_YR2_RT
- Enrollment of undergraduate certificate/degree-seeking students UGDS
- Percent withdrawn from original institution within 2 years WDRAW_ORIG_YR2_RT
- The median debt for students who have completed GRAD_DEBT_MDN
- Percent completed within 2 years at original institution COMP_ORIG_YR2_RT
- Percentage first-generation students PAR_ED_PCT_1STGEN
- Tuition revenue per full-time equivalent student TUITFTE

I wanted to Visualize important variables against income broken down by school type concentrating on Public vs private non profit schools.
Public schools are shown in red and Private nonprofit blue


Average family income of dependent students
140000
120000
100000
80000
60000
40000
20000
0
0
20000
40000
60000
80000
100000
120000
140000
160000
180000


Percent who transferred to a 4-year institution and completed within 4 years
140000
120000
100000
80000
60000
40000
20000
0
−0.2
0.0
0.2
0.4
0.6
0.8
1.0


Average family income of independent students
140000
120000
100000
80000
60000
40000
20000
0
−10000
0
10000
20000
30000
40000
50000
60000
70000
80000


Average SAT
140000
120000
100000
80000
60000
40000
20000
0
600
800
1000
1200
1400
1600

Model Evaluation

Benchmark model

- To evaluate other models, I have selected LinearRegression. Benchmark model was trained on 2 columns and has very high MSE score and 0% accuracy.

```
Mean squared error: 58064335.19
Accuracy score: -0.04
```

Having selected the variables to train the models on and to test the accuracy of the model I reserved 40% of data for test scenarios. Meaning that I will train the models on 60% of the data and test on the rest.

I trained a model on following classifiers:

- LinearRegression
- Ridge Regression including Cross-Validation
- Lasso including Lasso Lars Cross-Validation
- RandomForestRegressor

All 3 leaner repressors had similar accuracy result of 60%. Random Forest Repressor had best performance of 71% accuracy.

	Linear reg	RandomForestClassifier	Ridge
Mean squared error	25,255,203.66	18,263,105.34	25,266,928.04
Accurancy	0.60	0.71	0.60

While 71% accuracy is promising, it leaves room for improvement. Also Mean Squared error appears to be fairly high.
Top 5 Important features ranked by model that had highest accuracy are:

DEP_INC_AVG	1	Average family income of dependent students in real 2015 dollars.
COMP_ORIG_YR4_RT	0.59	Percent completed within 4 years at original institution
AVGFACSAL	0.26	Average faculty salary
IND_INC_AVG	0.26	Average family income of independent students in real 2015 dollars.
SAT_AVG	0.22	Average SAT equivalent score of students admitted
ENRL_ORIG_YR2_RT	0.13	Percent still enrolled at original institution within 2 years