Capstone Project Rafet Ducic

Machine Learning Engineer Nanodegree

Project Overview

I. Definition

Project Overview

Predicting student future earnings based on college.

Department of Education collects data from universities on students who apply for financial aid, this is publicly available data that can help students determine best college selection. 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.

In this paper explore how various college characteristics influence future student earnings and attempt to answer if 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 create a model capable of predicting future student's earning based on the available data.

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.

Problem Statement

The goal is to analyze college score-card data provided through US

Department of Education. Data is provided for period of 1996-2014. Data is provided in large number of tab delimited files along with column mapping. To enable easier filtering and exploration of data I need to:
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 that there is a relationship with response data
3. Train different classifiers measuring predictive accuracy on Test data
4. Evaluate results and further refine features

Metrics

Due to the large number of features available in the dataset (1731 columns, out of this 397 are numerical) we need a way either programmatically or heuristically picking important variables. Because a programmatic approach is likely going to pick features which are most correlative to the response variable I choose to use the combination of the two approaches. Start by picking variables that related research has shown as important factors to student success, and use programmatic evaluation to confirm importance.

I will use two set of metrics to evaluate models one for feature importance and another for evaluating model performance.

Feature importance:

- 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

Predictive Model Evaluation:

- 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.

II. Analysis

(approx. 2-4 pages)

Data Exploration

Given that the data is very wide (1731 columns) its immediately apparent that in order to have a manageable analysis I need to reduce this dataset. Looking at the previous research on the subject **list research I choose to concentrate on following part of the data set

School related information: 'PREDDEG', 'HIGHDEG', 'CONTROL',

md_earn_wne_p6	Median earnings of students working and not enrolled 6 years after entry
PREDDEG	Predominant undergraduate degree awarded 0 Not classified 1 Predominantly certificate-degree granting 2 Predominantly associate's-degree granting 3 Predominantly bachelor's-degree granting 4 Entirely graduate-degree granting
HIGHDEG	Highest degree awarded 0 Non-degree-granting 1 Certificate degree 2 Associate degree 3 Bachelor's degree 4 Graduate degree
CONTROL	Control of institution (Private NP, Private P, Public)
NUMBRANCH	Number of branch campuses
AVGFACSAL	Average faculty salary
ADM_RATE	Admission rate
SAT_AVG	Average SAT equivalent score of students admitted
TUITFTE	Net tuition revenue per full-time equivalent student
UGDS	Enrollment of undergraduate certificate/degree-seeking students
UGDS_NRA	Total share of enrollment of undergraduate degree-seeking students who are non-resident aliens
PPTUG_EF	Share of undergraduate, degree-/certificate-seeking students who are part-time
UG25abv	Percentage of undergraduates aged 25 and above
PAR_ED_PCT_1STGEN	Percentage first-generation students
DEP_INC_AVG	Average family income of dependent students in real 2015 dollars.
IND_INC_AVG	Average family income of independent students in real 2015 dollars.

COMP_ORIG_YR2_RT	Percent completed within 2 years at original institution
WDRAW_ORIG_YR2_RT	Percent withdrawn from original institution within 2 years
ENRL_ORIG_YR2_RT	Percent still enrolled at original institution within 2 years
COMP_ORIG_YR4_RT	Percent completed within 4 years at original institution
WDRAW_ORIG_YR4_RT	Percent withdrawn from original institution within 4 years
ENRL_ORIG_YR4_RT	Percent still enrolled at original institution within 4 years
OVERALL_YR2_N	Number of students in overall 2-year completion cohort
OVERALL_YR3_N	Number of students in overall 3-year completion cohort
OVERALL_YR4_N	Number of students in overall 4-year completion cohort
OVERALL_YR6_N	Number of students in overall 6-year completion cohort
OVERALL_YR8_N	Number of students in overall 8-year completion cohort
count_nwne_p6	Number of students not working and not enrolled 6 years after entry
DEBT_MDN	Median debt, suppressed for n=30
GRAD_DEBT_MDN	Median debt of completers, suppressed for n=30

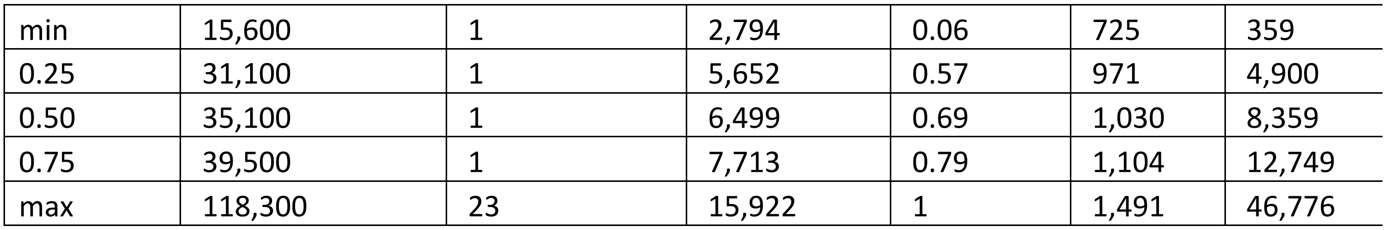
I called this subset of dataset data_reduced having 28281 rows and 31 columns out of which 3 were categorical and 28 numerical.

Number of columns have PrivacySuppressed instead of values, In order for linear models to work correctly all of the data has to be of the same type (numeric) in a single column. I replaced PrivacySuppressed with NaN (Null).

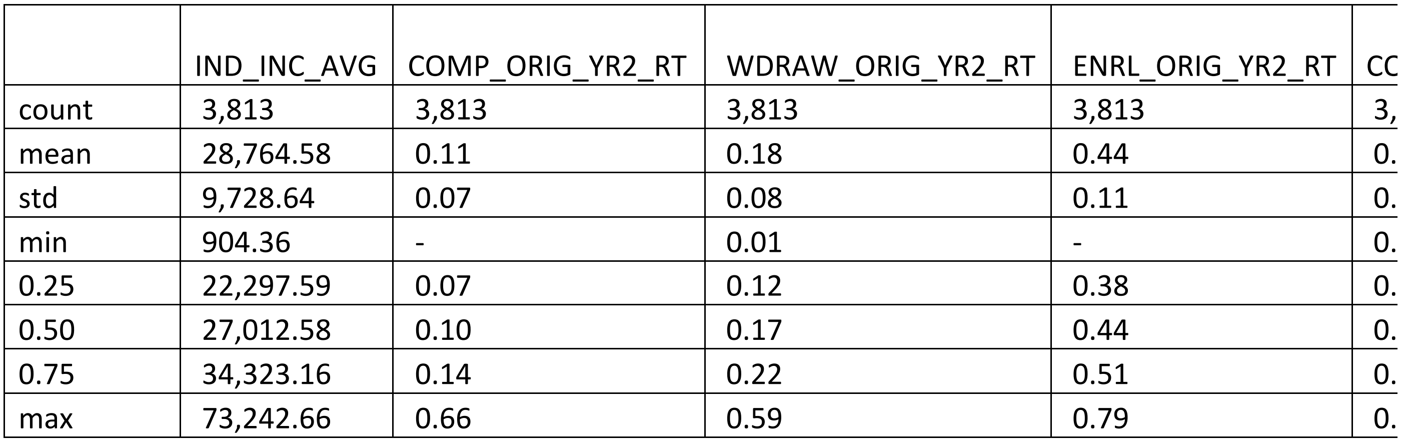
Another requirement of linear models is that it requires non Null data. I originally chose to fill the null values with 0 which had similar results as mean. However as pointed out in refinement section this led to significantly changes in the models to the point where prediction logic produced didn't make sense. Instead I chose to drop all the null values. Doing this further reduces the dataset to 3,813 rows and 31 columns out of which 3 were categorical and 28 numerical.

Data Statistics:

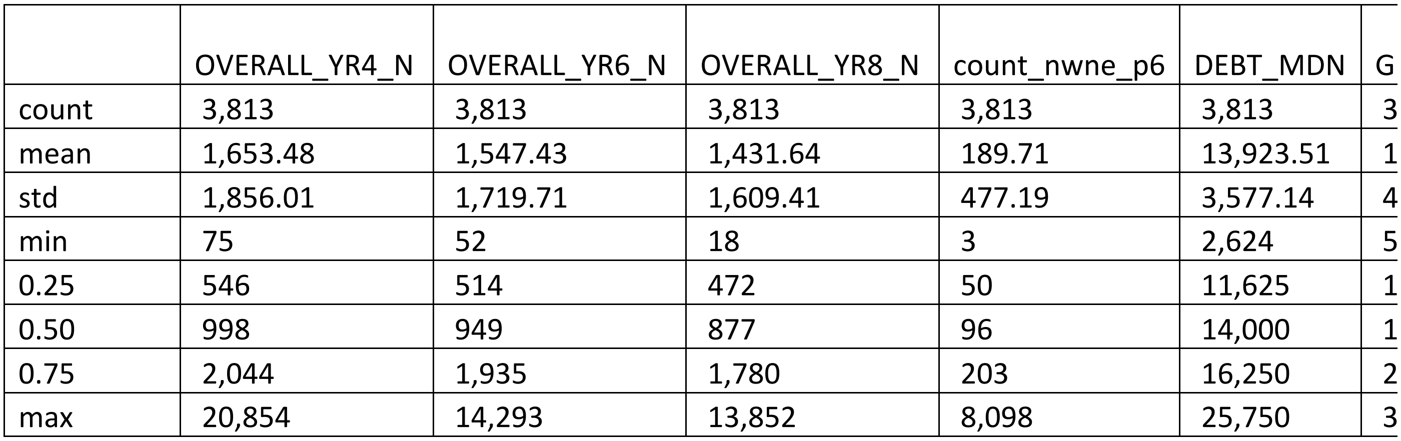
	md_earn_wne_p6	NUMBRANCH	AVGFACSAL	ADM_RATE	SAT_AVG	TUITFTE
count	3,813	3,813	3,813	3,813	3,813	3,813
mean	35,929.16	1.36	6,809.88	0.67	1,043.97	9,465.49
std	8,127.25	1.95	1,684.15	0.18	114.15	5,682.89



min	15,600	1	2,794	0.06	725	359
0.25	31,100	1	5,652	0.57	971	4,900
0.50	35,100	1	6,499	0.69	1,030	8,359
0.75	39,500	1	7,713	0.79	1,104	12,749
max	118,300	23	15,922	1	1,491	46,776



	IND_INC_AVG	COMP_ORIG_YR2_RT	WDRAW_ORIG_YR2_RT	ENRL_ORIG_YR2_RT	CC
count	3,813	3,813	3,813	3,813	3,
mean	28,764.58	0.11	0.18	0.44	0.
std	9,728.64	0.07	0.08	0.11	0.
min	904.36	-	0.01	-	0.
0.25	22,297.59	0.07	0.12	0.38	0.
0.50	27,012.58	0.10	0.17	0.44	0.
0.75	34,323.16	0.14	0.22	0.51	0.
max	73,242.66	0.66	0.59	0.79	0.

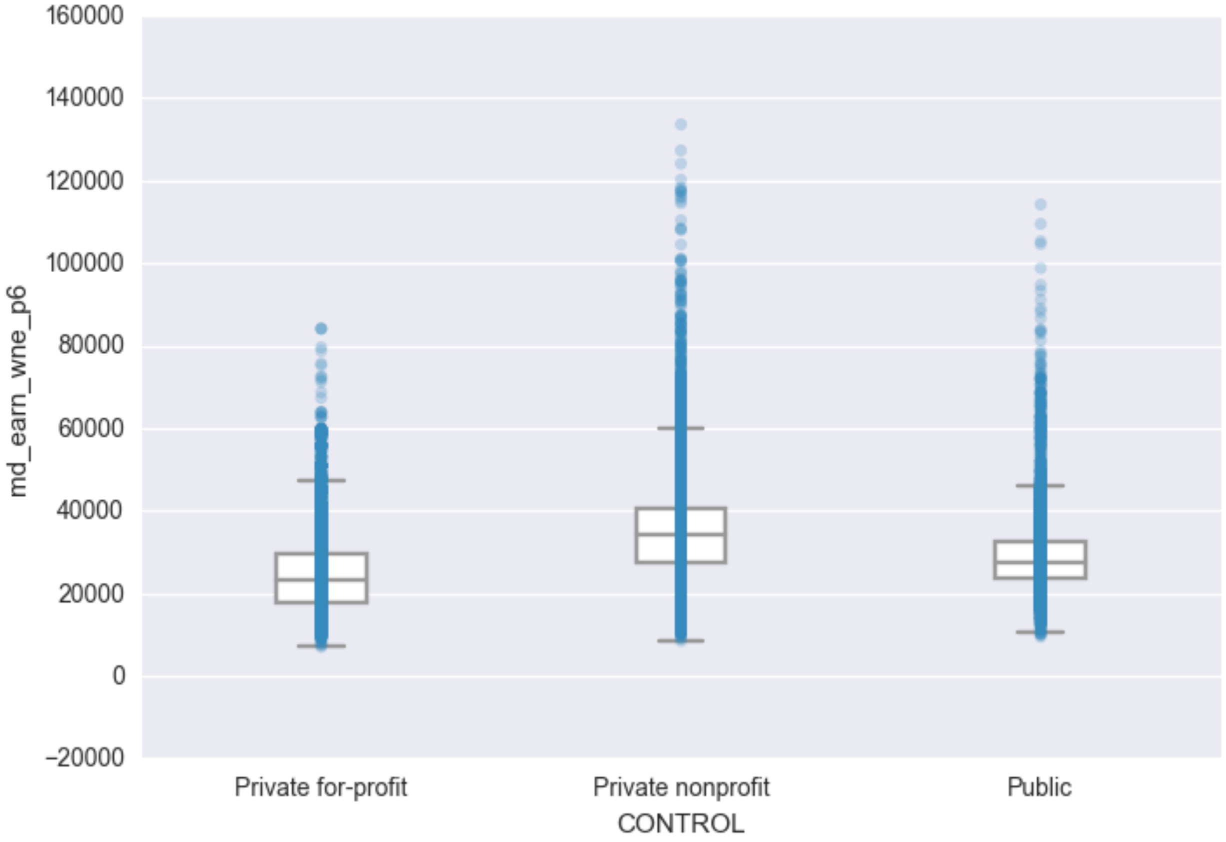


	OVERALL_YR4_N	OVERALL_YR6_N	OVERALL_YR8_N	count_nwne_p6	DEBT_MDN	G
count	3,813	3,813	3,813	3,813	3,813	3
mean	1,653.48	1,547.43	1,431.64	189.71	13,923.51	1
std	1,856.01	1,719.71	1,609.41	477.19	3,577.14	4
min	75	52	18	3	2,624	5
0.25	546	514	472	50	11,625	1
0.50	998	949	877	96	14,000	1
0.75	2,044	1,935	1,780	203	16,250	2
max	20,854	14,293	13,852	8,098	25,750	3

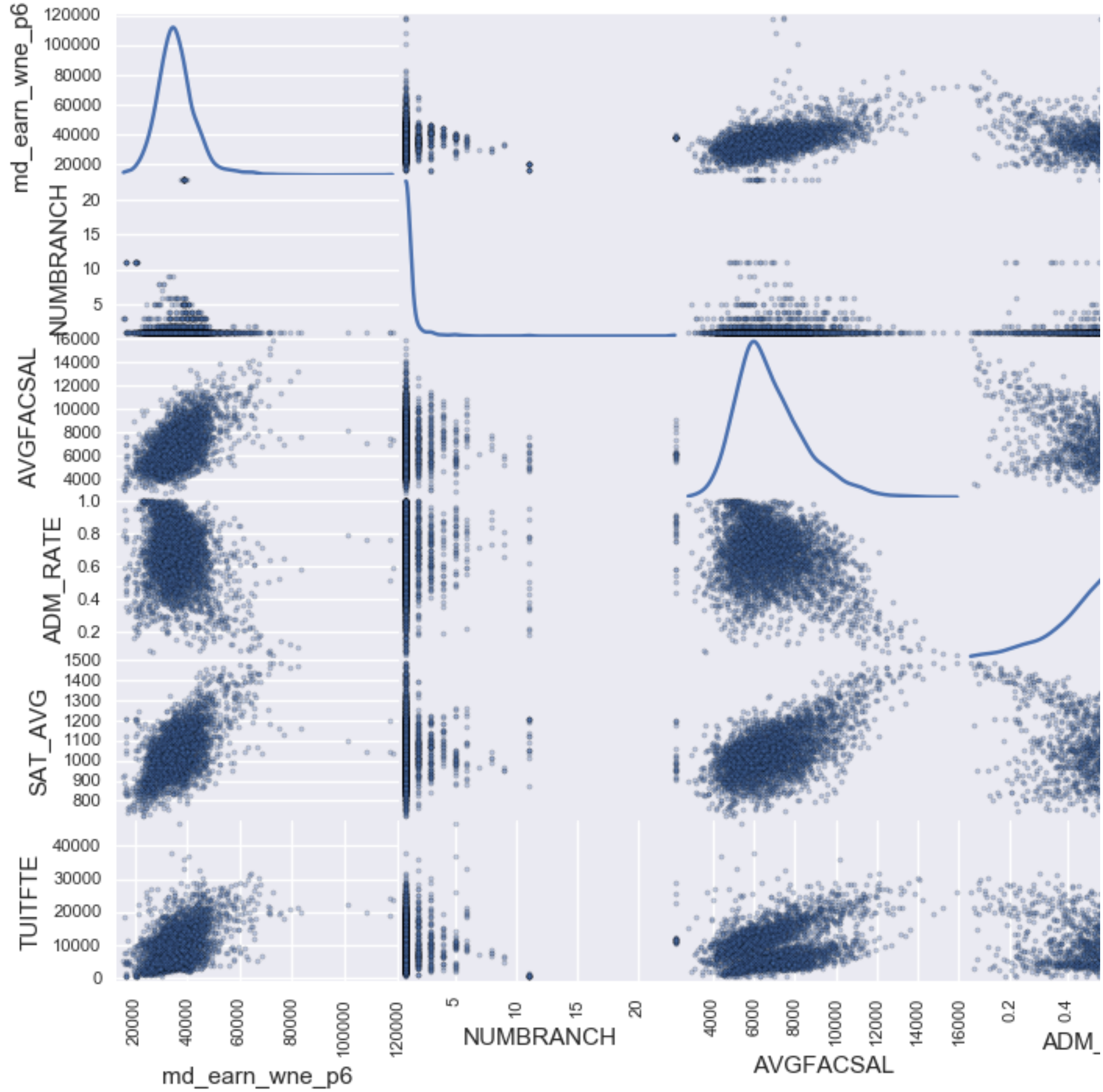
- *If a dataset is present for this problem, have you thoroughly discussed certain features about the dataset? Has a data sample been provided to the reader?*
- *If a dataset is present for this problem, are statistics about the dataset calculated and reported? Have any relevant results from this calculation been discussed?*
- *If a dataset is **not** present for this problem, has discussion been made about the input space or input data for your problem?*
- *Are there any abnormalities or characteristics about the input space or dataset that need to be addressed? (categorical variables, missing values, outliers, etc.)*

Exploratory Visualization

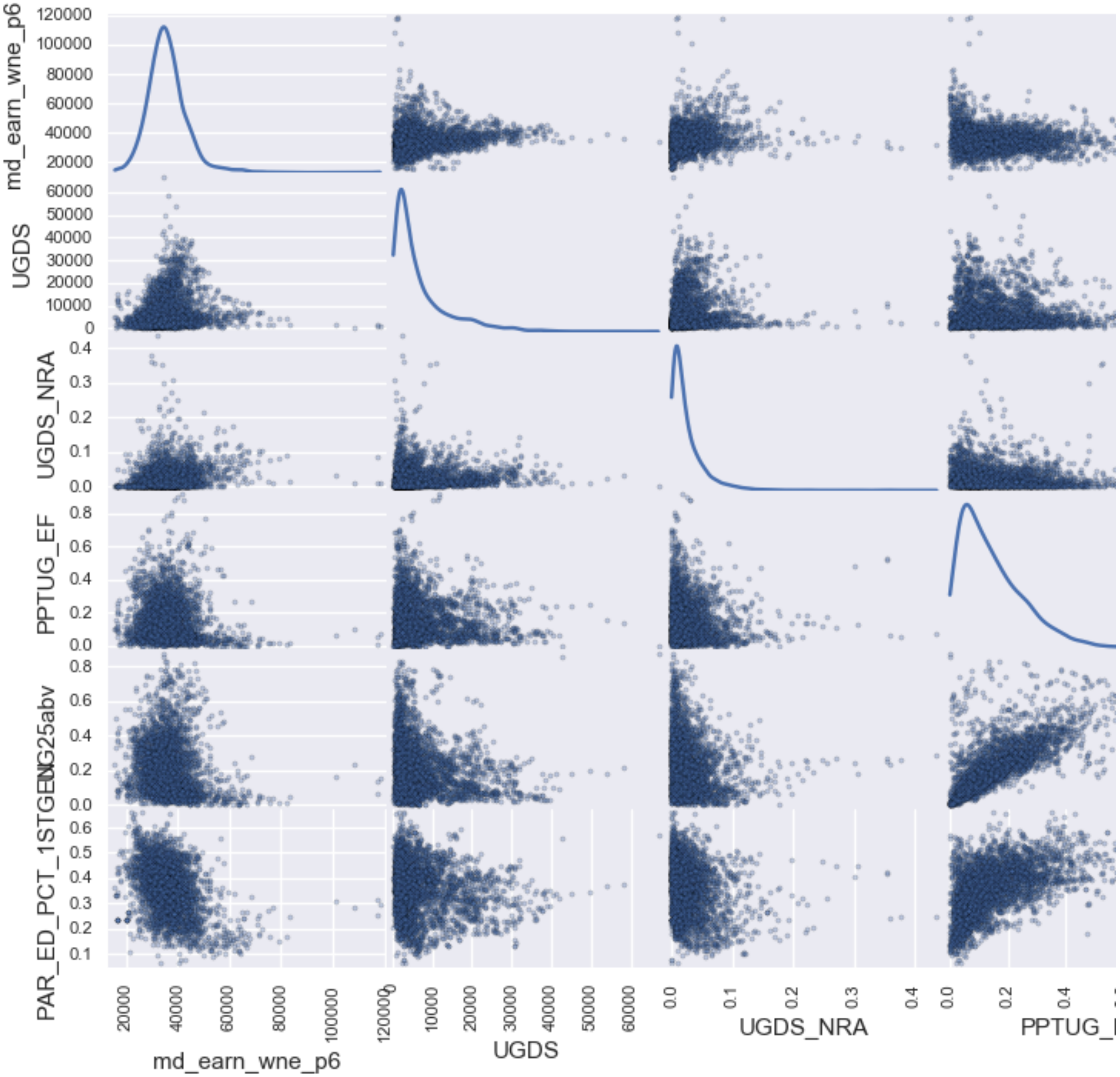
Response Variable



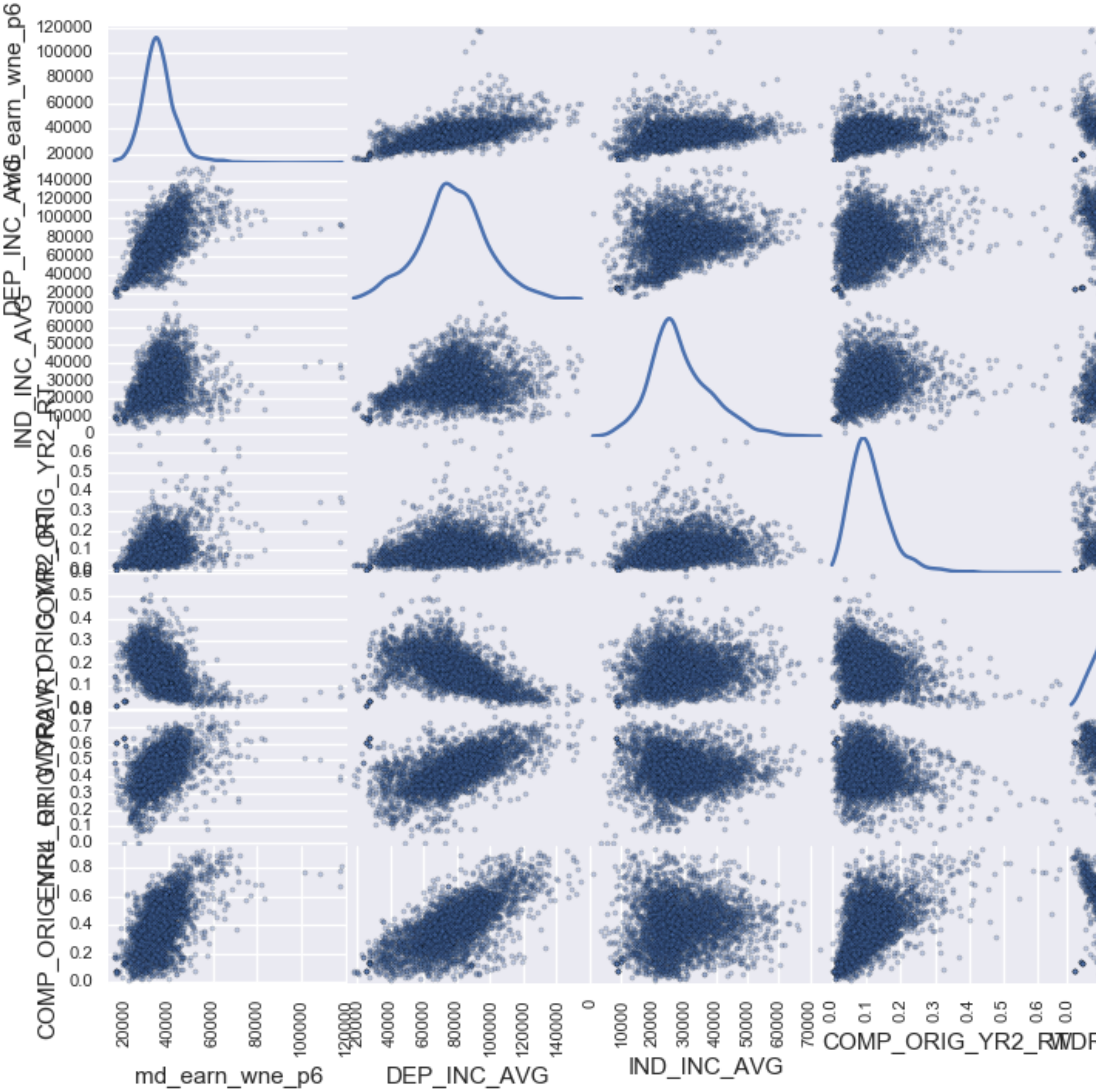
- 'md_earn_wne_p6', 'NUMBRANCH', 'AVGFACSAL', 'ADM_RATE', 'SAT_AVG', 'TUITFTE



- 'md_earn_wne_p6', 'UGDS','UGDS_NRA', 'PPTUG_EF', 'UG25abv', 'PAR_ED_PCT_1STGEN'

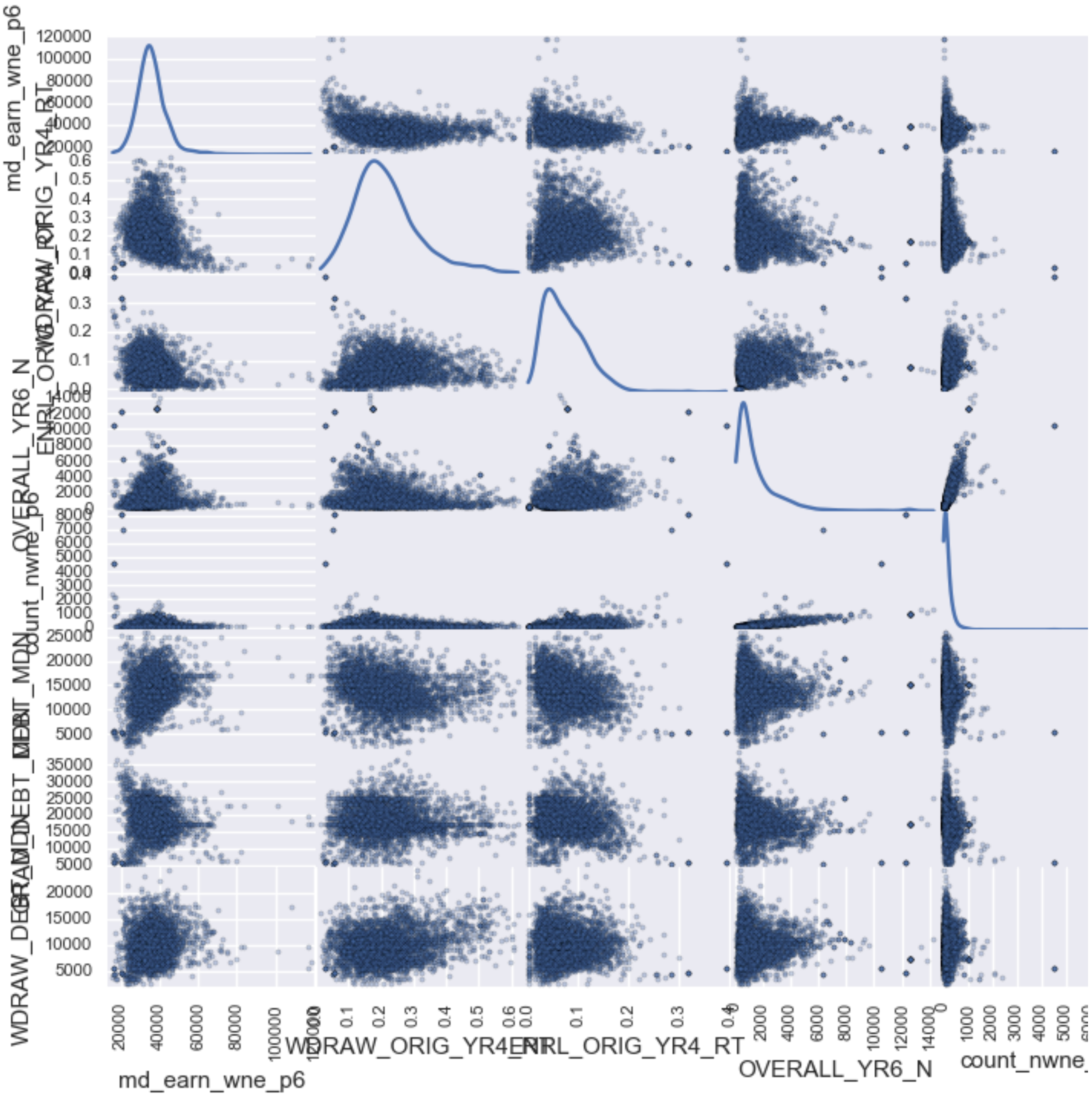


- 'md_earn_wne_p6', 'DEP_INC_AVG','IND_INC_AVG', 'COMP_ORIG_YR2_RT', 'WDRAW_ORIG_YR2_RT', 'ENRL_ORIG_YR2_RT', 'COMP_ORIG_YR4_RT'



-

- 'md_earn_wne_p6', 'WDRAW_ORIG_YR4_RT', 'ENRL_ORIG_YR4_RT', 'OVERALL_YR6_N', 'count_nwne_p6', 'DEBT_MDN', 'GRAD_DEBT_MDN', 'WDRAW_DEBT_MDN'



•

Graphs above show comparison of each variable to our response variable. Kernel Distribution is shown Diagonally.

I chose to compare the response variable to each feature which shows collinearity between them Strongest relationship can be observed between future income and

- PAR_ED_PCT_1STGEN Percentage first-generation students
- AVGFACSAL Average faculty salary

- SAT_AVG Average SAT equivalent score of students admitted
- DEP_INC_AVG Average family income of dependent students in real 2015 dollars.
- IND_INC_AVG Average family income of independent students in real 2015 dollars.

Algorithms and Techniques

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.
I will use following regression algorithms

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

Because Ridge Regression and Lasso require passing of **alpha parameter** I will use Cross-Validation to find a best parameter.

In this section, you will need to discuss the algorithms and techniques you intend to use for solving the problem. You should justify the use of each one based on the characteristics of the problem and the problem domain. Questions to ask yourself when writing this section:

- *Are the algorithms you will use, including any default variables/parameters in the project clearly defined?*
- *Are the techniques to be used thoroughly discussed and justified?*
- *Is it made clear how the input data or datasets will be handled by the algorithms and techniques chosen?*

Benchmark

Outside benchmark is not available for this problem so I will use liner regression trained on Average SAT score and has very high MSE score and 0% accuracy. I am using SAT score because of its strong correlation to response variable and because of assumption that higher SAT score will lead to acceptance into highly ranked school, which according to traditionally held belief will result in high earnings after graduation.

After running Liner Regression on available data following benchmark values are obtained:
Mean squared error: 45721737.61
Variance score: 0.25

III. Methodology

(approx. 3-5 pages)

Data Preprocessing

Due to different scales the data is on preprocessing is required for most regression algorithms. I chose to use centering around the mean as method for data scaling. *Based on the* ***Data Exploration*** I detected some outlier is tuition fee. I will be removing tuition fees higher than 1Million.

Further pre-processing was performed on PREDDEG, HIGHDEG and CONTROL features by Encode labels with value between 0 and n_classes-1. Original Features were dropped after visualization was performed.

Once dataset only contained numerical values, each feature was filtered for PrivacySuppressed values and where found PrivacySuppressed was replaced with null.

One of the requirement of regression models is that data can't contain null values I have dropped nulls.

Implementation

To predict future earnings I chose to use following classifiers

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

Linear Regression is useful predictor, but it can suffers from collinearity. With each model that exposes Coefficient estimates I will report on them. Ordinary Least Squares rely on the independence of the model terms. When terms are correlated and the columns of the design matrix X have an approximate linear dependence, the design matrix becomes close to singular and as a result, the least-squares estimate becomes highly sensitive to random errors in the observed response, producing a large variance. This situation of multicollinearity can arise, for example, when data are collected without an experimental design.

Ridge regression addresses some of the problems of Ordinary Least Squares by imposing a penalty on the size of coefficients. The ridge coefficients minimize a penalized residual sum of squares.

The **Lasso** is a linear model that estimates sparse coefficients. It is useful in some contexts due to its tendency to prefer solutions with fewer parameter values, effectively reducing the number of variables upon which the given solution is dependent.

A **random forest** is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and use averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is always the same as the original input sample size but the samples are drawn with replacement if bootstrap=True

To prevent overfitting and to test accuracy of the models I have separated data into training and test datasets, retaining 40% of data for test. `16286 data points across columns were available for training and 10858 data points for test.`

After running each of the models above following are results of the tests:

Model	Mean squared error	Accuracy
Benchmark	45,721,737	0.25
Linear Regression	25,256,789	0.60
Ridge Regression	25,337,600	0.60
Lasso	25,258,307	0.60
RandomForestRegressor	18,003,431	0.72

As we can see from the results above, each of the algorithms over performed the benchmark model, which in truth was not difficult to do since accuracy of the benchmark model was only 25%.

In this section, the process for which metrics, algorithms, and techniques that you implemented for the given data will need to be clearly documented. It should be abundantly clear how the implementation was carried out, and discussion should be made regarding any complications that occurred during this process. Questions to ask yourself when writing this section:

- *Is it made clear how the algorithms and techniques were implemented with the given datasets or input data?*
- *Were there any complications with the original metrics or techniques that required changing prior to acquiring a solution?*
- *Was there any part of the coding process (e.g., writing complicated functions) that should be documented?*

Refinement

In hopes of improve accuracy of the models and since number of models used require parameter to be passed specified I will use Cross-Validation with Ridge Regression and Lasso Lars. For Random forest I will use bootstrapping.

Model	Mean squared error	Accuracy
Benchmark	45,721,737	0.25
Linear Regression	25,256,789	0.60
Ridge Regression	25,337,600	0.60
Ridge Regression Cross-Validation a=0.1	25,264,868	**0.60**
Lasso	25,258,307	0.60
LassoLarsCV a=0.2	**25,191,547**	**0.60**
RandomForestRegressor	18,003,431	0.72
RandomForestRegressor with bootstrap	**17,655,114**	**0.72**

After comparing the results of final model and examining the feature importance it became clear that filling missing values with 0 was significantly affecting the results.

For example examining feature importance, which I cover later in the section it would appear that the most important variable for predicting future salaries is “Number of students in overall 2-year completion cohort” When I examined the graph it showed that 0 was the highest predictor, this of course doesn’t make sense.   At this point I have gone back and modified pre-processing step to **Drop rows with null values. This was of course** a major setback as I had to re-evaluate all of the points I have made.

In this section, you will need to discuss the process of improvement you made upon the algorithms and techniques you used in your implementation. For example, adjusting parameters for certain models to acquire improved solutions would fall under the refinement category. Your initial and final solutions should be reported, as well as any significant intermediate results as necessary. Questions to ask yourself when writing this section:

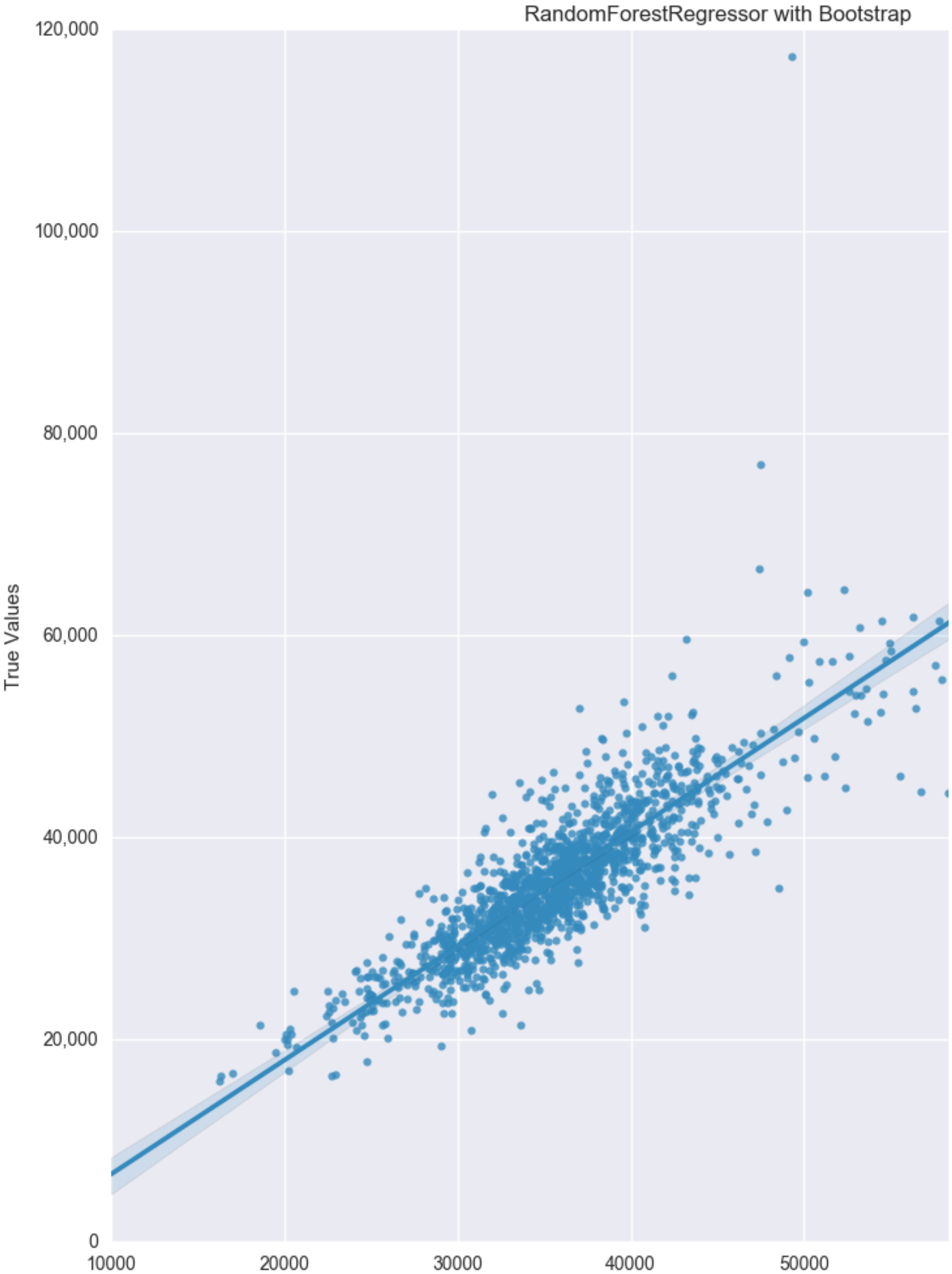
- *Has an initial solution been found and clearly reported?*
- *Is the process of improvement clearly documented, such as what techniques were used?*
- *Are intermediate and final solutions clearly reported as the process is improved?*

IV. Results

(approx. 2-3 pages)

Model Evaluation, Validation and Justification

As can be observed from results obtained in model training and refinement **RandomForestRegressor with bootstrap** performs the best in term of accuracy. In visualizing the predicted vs true values we can see that model predicts well for salaries between 25,000 and $50,000 after which data becomes sparse which results in more inaccurate prediction. The line and the shading represents margin of error of predictions.


RandomForestRegressor with Bootstrap
True Values
120,000
100,000
80,000
60,000
40,000
20,000
0
10000
20000
30000
40000
50000

Questions:

- *Is the final model reasonable and aligning with solution expectations? Are the final parameters of the model appropriate?* Model is well aligned with true data points as can be see from the graph above.
- *Has the final model been tested with various inputs to evaluate whether the model generalizes well to unseen data? Model is trained on the 60% of the data and tested on 40%*
- *Is the model robust enough for the problem? Do small perturbations (changes) in training data or the input space greatly affect the results? This has not been observed.*
- *Can results found from the model be trusted? Model can be trusted to predict salaries with accuracy and pf 72 percent of accuracy*
- *Are the final results found stronger than the benchmark result reported earlier? Yes, results are significantly Stogner than benchmark model*
- *Have you thoroughly analyzed and discussed the final solution? Significant detail is provided.*
- *Is the final solution significant enough to have solved the problem? Predictability of 72% is good but not excellent as it leaves 28% chance of incorrectly predicting final salary.*
-

V. Conclusion

(approx. 1-2 pages)

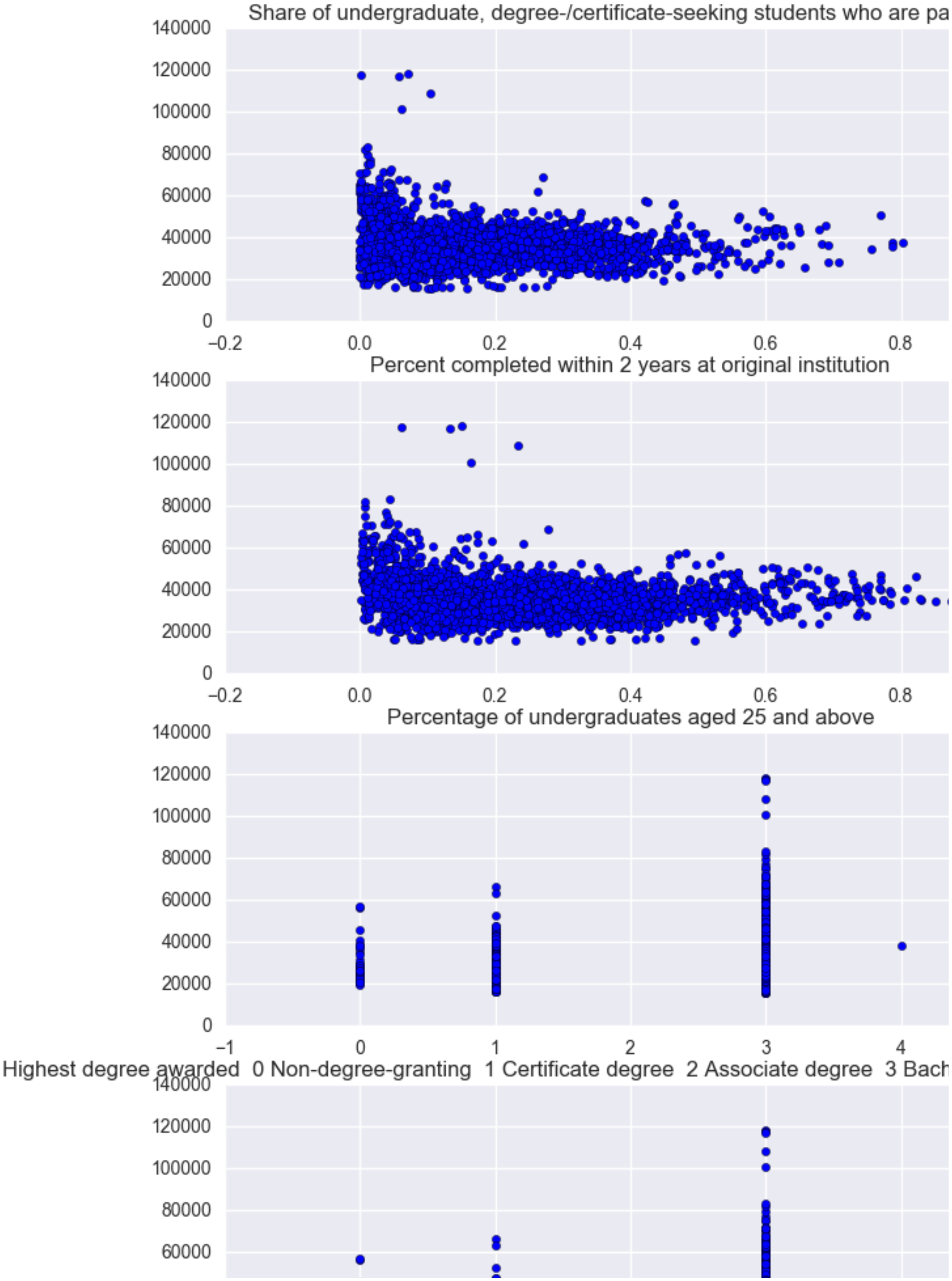
Free-Form Visualization

Feature Importance

Random Forest Regressors with boot strapping was shown to have highest accuracy, and while most linear models expose coefficients, Random Forest doesn’t but instead it exposes Feature importance. I will examine top 10 features according to their importance.

Model	RFR
PPTUG_EF	1
COMP_ORIG_YR2_RT	0.59
UG25abv	0.25
HIGHDEG_N	0.25

In next series for visualizations I will compare response variable to each of the important features:



Share of undergraduate, degree-/certificate-seeking students who are pa

Percent completed within 2 years at original institution

Percentage of undergraduates aged 25 and above

Highest degree awarded 0 Non-degree-granting 1 Certificate degree 2 Associate degree 3 Bach

What graphs and the important features suggest is that for high post-graduation salary most important features (from the ones examined here are :

- Highest salaries are for
- Share of undergraduate, degree-/certificate-seeking students who are part-time **that are close to 0**. Meaning Students that were not part time were more likely to have higher salaries
- Percent completed within 2 years at original institution **that are close to 0.** Meaning Students who continued studying after 2 were more likely to have higher salaries
- Percent of undergraduates aged 25 and above that at 3 percent or above. Meaning Students who were 25 or older were more likely to have higher salaries
- Highest degree awarded. With Bachelor degree having highest earnings

Reflection

This was definitely a very challenging project, in many aspects. From data import with large number of files to process, to very large feature space over 1700 columns.

For all of the reasons listed above and in general predicting student income after graduation is a very difficult task, I am glad that I was able to come close to predicting it with 72% accuracy. Looking at the predictions and factors that influence high income after graduation and tying it back to conventional wisdom it makes sense.

What data appears to suggest is concentrate on studying (don’t do it part time), Stay in school for more than 2 years and graduate with Bachler degree.

Improvement

There are number of areas of improvements in this project in following areas:

- Feature Selection. In this project I have analyzed fraction of all the available features. One idea is to programmatically select features to use by starting with an empty model and them looping through list of features adding one by one and based on gain in predictive performance keeping the feature.

- Feature engineering and categorization. Majority of features are non-numerical (397 columns, out of 1731). I would like to transform all the columns to numerical values.
- Missing data. This dataset is very sparse with only 13% of the dataset having all the values for columns I selected. I would like to examine each column and choose appropriate fill value (mean, min, max, 0, etc.)
- Models. There are number of other regression models that I would like to try out. This is in addition to using other models such as Using Neural Networks with Regression that I would like to try out.