Predicting Student Default Loans

Nico Van de Bovenkamp
Danny Vilela
Zhenyu Wang

December 10, 2016

Business Understanding

Our data project seeks to determine the key factors of institutions that result in high student default rates on federal student loans. We want to understand what features present among institutions are most likely to result in (high) student default rates on federal student loans in 2012.

Our business problem – often deemed a crisis – is notorious for its impact on the political, social, and economic state of affairs: as of 2016, Americans with federal direct loans owed approximately $911.6 billion across 30.5 million borrowers, a large majority of the $1.26 trillion total U.S. student loan debt[1]. The importance of our business problem cannot be understated—being able to identify and communicate students' risk of defaulting on federal student loans based on their prospective enrollment could save millions of Americans, collectively, billions of dollars.

By mining for data pertaining to both higher-education institutions and the students that self-select into these institutions, we seek to address the business problem by building a model that can estimate the relationship between a potential institution a student would attend and the aggregate default rate for that institution. A scalable model that can help identify features across institutions that contribute to higher federal default rate offers benefits for both students and education legislators. Knowing how their options will likely influence their post-college financial status is a valuable, considerable factor to be included in their choice to pursue higher education. Likewise, legislators will be able to work with schools in their districts to more effectively reduce federal student loan default rate and improve the quality of local educational institutions.

[1] https://studentloanhero.com/student-loan-debt-statistics/

Data Sources

The initial dataset came from the Federal Student Aid Office of the Department of Education[2]. This dataset includes our target variable of per-institution cohort default rates from 2011 to 2013, in addition to a several other features including: city, state, program length, school type (public, private non-profit, proprietary, etc.), ethnic code (if the school identifies as a Historically Black College or University or Native American University, etc.) for approximately 4,760 institutions.

The first dataset contained an OPEID field – a unique identifier from the Office of Postsecondary Education – which we used as the unique link from this first dataset to the 2011 and 2012 datasets from the Integrated Postsecondary Education Data System (IPEDS), which is maintained by the U.S. Department of Education's National Center for Education Statistics. This second dataset included a few more institutional features while also incorporating student-centric features. Most importantly, these datasets contained a UNITID field – a unique identifier for every institution from the Department of Education – which we used to connect to our final dataset from the Delta Cost Project in the Delta Cost Database compiled by the American Institute for Research.

Our final dataset by the Delta Cost Project[3] needed to be merged with the IPEDS data set via their mutual UNITID field. Recall that our first data set is the only one that contains our target variable, hence in order to more fully develop our feature space we join the two via their OPEID.

We note that the data sets from IPEDS and the Delta Cost Project came from mandatory surveys of the over 7,500 institutions that participate in – or are applicants for – participation in any federal student financial aid program (such as Pell grants and federal student loans) authorized by Title IV of the Higher Education Act of 1965, as amended (20 USC 1094, Section 487(a)(17) and 34 CFR 668.14(b)(19))[4].

Dictionaries

We include links to data dictionaries provided by the data sources to clarify the numerous features across each dataset.

- *PEP* by Federal Student Aid Office of the Department of Education: `http://www2.ed.gov/offices/OSFAP/defaultmanagement/instructions.html`

[2] `http://www2.ed.gov/offices/OSFAP/defaultmanagement/cdr.html`

[3] `http://www.deltacostproject.org/delta-cost-project-database`

[4] `http://nces.ed.gov/ipeds/Home/AboutIPEDS`

- *IPEDS* by IES NCES: `http://nces.ed.gov/ipeds/deltacostproject/`
- *Delta* by Delta Cost Project: `http://www.deltacostproject.org/delta-cost-project-database`

Data Understanding

The first dataset comes from the Federal Student Aid Office of the Department of Education. This dataset includes our target variable of cohort default rates, per institution, from 2011 to 2013, in addition to several other features such as the Ethnic Code (if the school identifies as a Historically Black College or University or Native American University, etc.) and the OPEID (a unique identifier from the Office of Postsecondary Education) for 4,760 institutions.

The second dataset is from the Delta Cost Project managed by the American Institutes for research. It contained the majority of the features in our model which are derived from IPEDS (Integrated Postsecondary Education Data System) finance, enrollment, staffing, completions and student aid data for academic years 1986-87 through 2011-12. However, this dataset does not have an OPEID, but a UNITID which shares the same function as the OPEID.

Lastly, our third dataset is from the Integrated Postsecondary Education Data System (IPEDS), which is maintained by the U.S. Department of Education's National Center for Education Statistics. This dataset included a few more institutional features, as well as student-centric features. But most importantly, this dataset contained both the UNITID and the OPEID of each institution, which we used to link the features in the second dataset to the target variable in the first dataset.

Data Preparation

In order to integrate our data sources for data mining, we first joined our first two datasets – PEP and IPEDS – on their 'OPEID' field in order to obtain a more complete dataset describing each institution. We used this preliminary dataset to build our project's proposal model: a decision tree classifier based solely on per-institution features (e.g. institution 'category' and 'size'). At this point, we had reliable data for all reported institutions across years 2011 and 2012.

We opted to reserve our 2012 findings as our final test set due to its applicability: if the year were 2012 and our model is trained on the federal default rate data from 2011, we would use new data from 2012 in order to evaluate our model's generalizability and ability to evaluate new data. Herein, our 'dataset' refers to federal loan default rates and associated institutional features from the reported 2011 fiscal year.

Our target variable 'DRate' is a continuous, numerical representation of students who, one year after graduating, defaulted on their federal student loans. Our target variable was available for the 2011 – 2012 fiscal years, and required little preparation aside from dropping ~2 rows whose target feature was a `NaN` value. The Cohort Default Rates are calculated by the Department of Education at the end of the second fiscal year after a report of students that are in repayment and rely on mandatory reports from each institution via the eCDR process to report loan records[5].

Once we had established our data set and target feature, we underwent some feature engineering in order to properly approach the analysis. Initially, we implemented rudimentary feature selection by going through each feature in the IPEDS and Delta Cost data sets, looking up the corresponding feature definition, and taking note of features that sounded even remotely worthy of being considered in a model. We classified features as 'worthy' based on `NaN` proportion ($\frac{invalid}{total}$, lower is better), features that could imply some causal relationship with students, or any feature that we could interpret with financial or student-level scope.

Next, when preparing to build our baseline model, we realized that many features we identified as potentially promising were filled with `NaN` values. Not all features were equally sparse – some features were missing 11 values, whereas others were missing upwards of 4200 values. Determined to save at least some of our features, we performed a difference of means test on each feature to determine whether we could safely impute the missing feature values with the feature mean. In particular, for each feature with missing values x_m, we:

1. Extracted x_m and our target feature y into a data frame.
2. Split our two-column data frame into two separate data frames: x_v where all values of x_m are valid/non-missing and x_i, where all values of x_m are invalid.
3. We then took the mean of our target feature y for both valid (y_v) and invalid (y_i) data frames. We then performed a difference of mean tests where we determined if the difference of means[1] between the two default rates was statistically significant with a 95% confidence interval.

[5] http://www2.ed.gov/offices/OSFAP/defaultmanagement/ecdr.html

4. If the means were not significantly different, we imputed the missing values with the mean of valid values. Furthermore, we generated a flag feature named $x_v\_missing$ to denote whether that particular row's value had been created via imputation (1) or not (0).
5. If the means were statistically significant, we opted to isolate the features by temporarily dropping them from our feature space. See *Future Work* to see how we would attempt to incorporate these features into a future model via clustering.

Lastly, we noted the abundance of categorical variables in our training set, including census region, ethnic code, program length, school type, and more. All of our categorical features could assume multiple different values, and so in order to improve model relevancy, simplicity, and specificity we dummified all categorical variables via one hot encoding. For a categorical feature x_c with values $v_1, v_2, \ldots, v_i, \ldots, v_n$ we generated a binary categorical variable x_c_i (`ICLEVEL_1`, `ICLEVEL_2`, etc.) denoting whether the categorical took on value i (1) or not (0). This allowed us to more easily interpret and fit our data into both scikit-learn's tree-based (decision tree, random forest) and linear models.

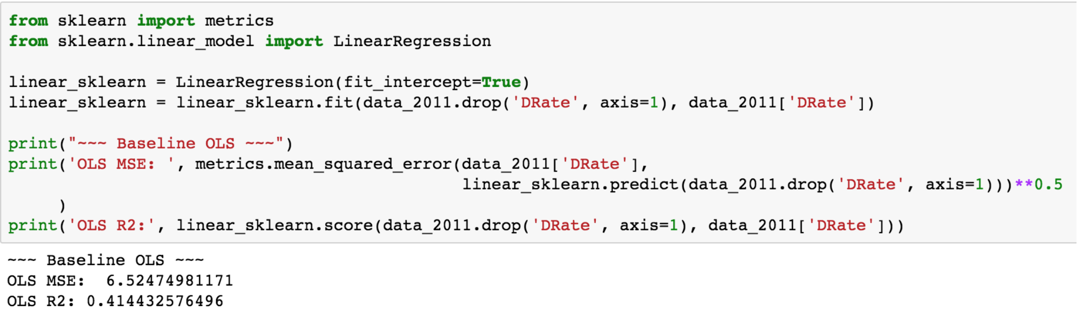
Modeling & Evaluation

When discussing choices for our data mining algorithm, we realized that – given the structure of our dataset and target variable – we had to select a supervised learning algorithm to learn a continuous variable. Thus, we could choose a linear model – like OLS, Ridge, LASSO, or Gradient Descent regressors – or a decision tree or random forest regressor. Taking into account the goal of our project, including the interpretability of the model and its output, we favored a linear model to be used for our final model instead of other, nonparametric regressors.

We acknowledge that random forests and decision trees are significantly less biased than linear models, do not struggle with colinearity of input features, rely much less on assumptions of the distributions of both the error terms and the feature values, and are better structured to model potentially non-linear relationships. That said, we could not sacrifice the interpretability of our model. A non-parametric model would not give as much feature-by-feature insight into the causal relationship of students taking loans and students defaulting. In order to make causal inference, we had to use a linear regression model for the coefficients to 'explain' the relationship between feature and default rate. With that in mind, we set out to build our baseline model.

Naïve Ordinary Least Squares

Following our decision to utilize a linear model, we approached our baseline model with relative ease: attempting to fit an ordinary least squares (OLS) linear model onto every feature in our training set. This gave us an easily interpretable understanding of our dataset while retaining model clarity (not a black box) and simplicity. Fitting our linear model onto every feature, we arrived at the following results:



```
from sklearn import metrics
from sklearn.linear_model import LinearRegression

linear_sklearn = LinearRegression(fit_intercept=True)
linear_sklearn = linear_sklearn.fit(data_2011.drop('DRate', axis=1), data_2011['DRate'])

print("~~~ Baseline OLS ~~~")
print('OLS MSE: ', metrics.mean_squared_error(data_2011['DRate'],
                                              linear_sklearn.predict(data_2011.drop('DRate', axis=1)))**0.5
     )
print('OLS R2:', linear_sklearn.score(data_2011.drop('DRate', axis=1), data_2011['DRate']))
```

```
~~~ Baseline OLS ~~~
OLS MSE:  6.52474981171
OLS R2: 0.414432576496
```

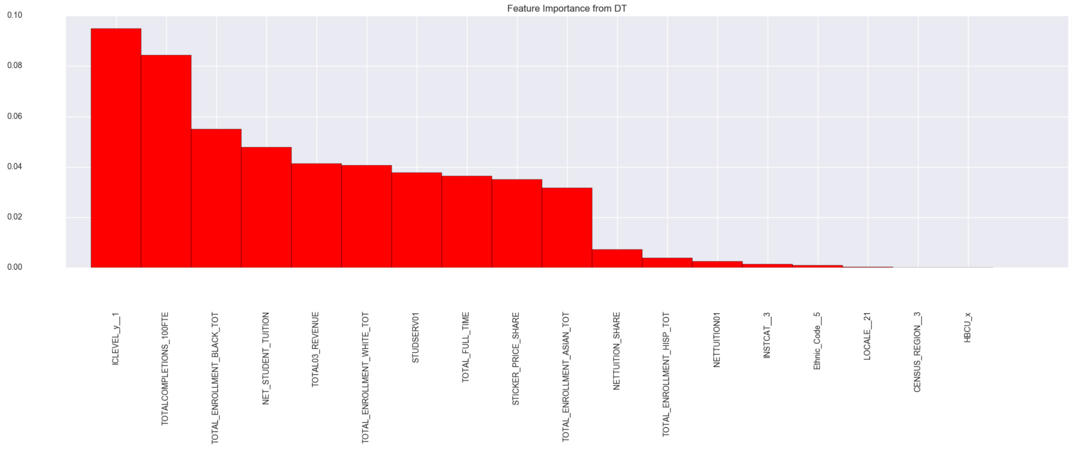
Our initial results were promising: a naïve linear model was able to explain ~41% of the variance of the underlying data, and was able to approximate a particular institution's federal loan default rate among new graduates to within ~6.5%.

A natural improvement upon our baseline model was feature selection: determining which features would be most informative to our understanding of federal default rate would allow for an even simpler model while retaining the clarity behind a simple linear model. In order to improve on the baseline, we sought to implement robust feature selection via the random forest regressor, and then re-evaluate our OLS model on the lower-dimensional space.

Feature Selection with Random Forest Regressors

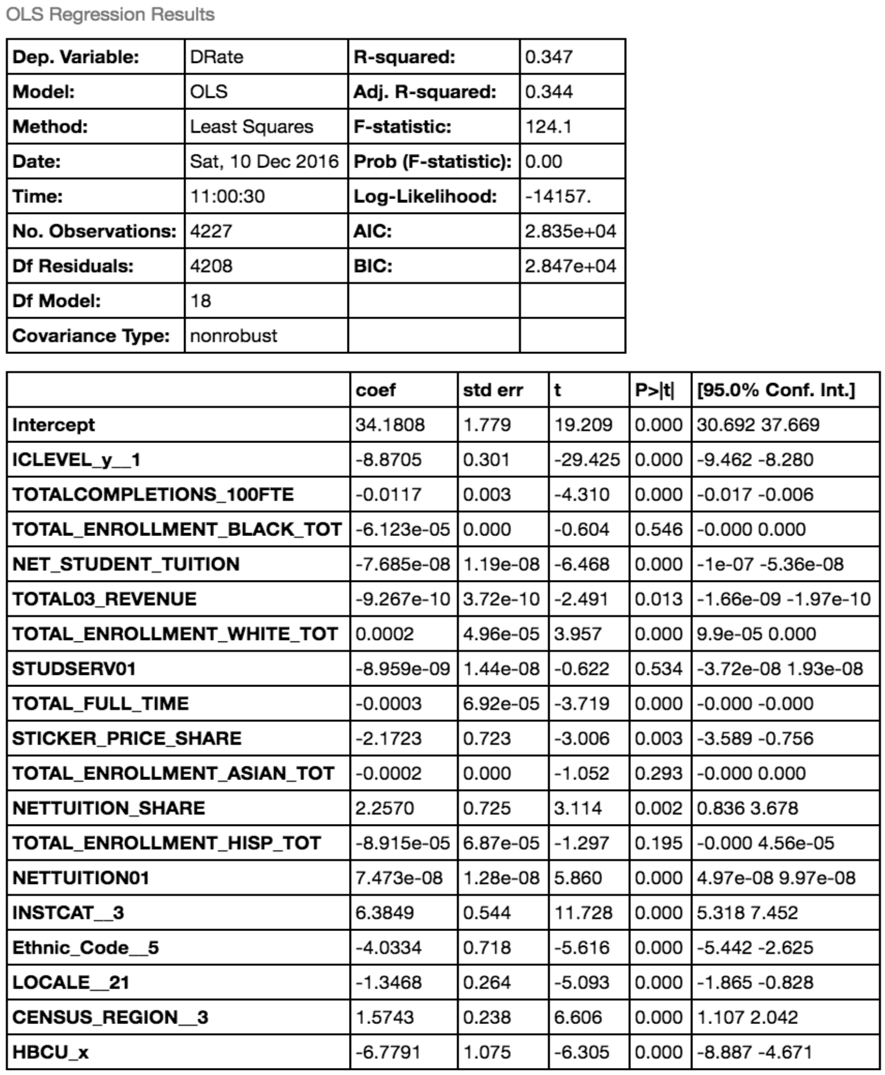
To understand our model's features, we elected to use an ensemble approach – specifically a random forest regressor [2] – in order to extract the most informative features as determined by a large number of decision trees. We limited the random forest's black-box interpretability by splitting our training dataset into further 80/20 train/test splits and fitting a random forest regressor model onto the training set. This way, we could introduce a little bias into our model, but reduce variance greatly by constructing many trees based on random samples of features to yield the most informative features.

After fitting our random forest regressor, we were able to extract and determine their relative feature importances, like so:



OLS Redux

Armed with our reduced feature list, we elect to revisit our baseline model to perform an apples-to-apples comparison of an ordinary least squares linear regression model with all features versus a reduced feature shortlist. We arrived at the following results:



OLS Regression Results

Dep. Variable:	DRate	**R-squared:**	0.347
Model:	OLS	**Adj. R-squared:**	0.344
Method:	Least Squares	**F-statistic:**	124.1
Date:	Sat, 10 Dec 2016	**Prob (F-statistic):**	0.00
Time:	11:00:30	**Log-Likelihood:**	-14157.
No. Observations:	4227	**AIC:**	2.835e+04
Df Residuals:	4208	**BIC:**	2.847e+04
Df Model:	18		
Covariance Type:	nonrobust		

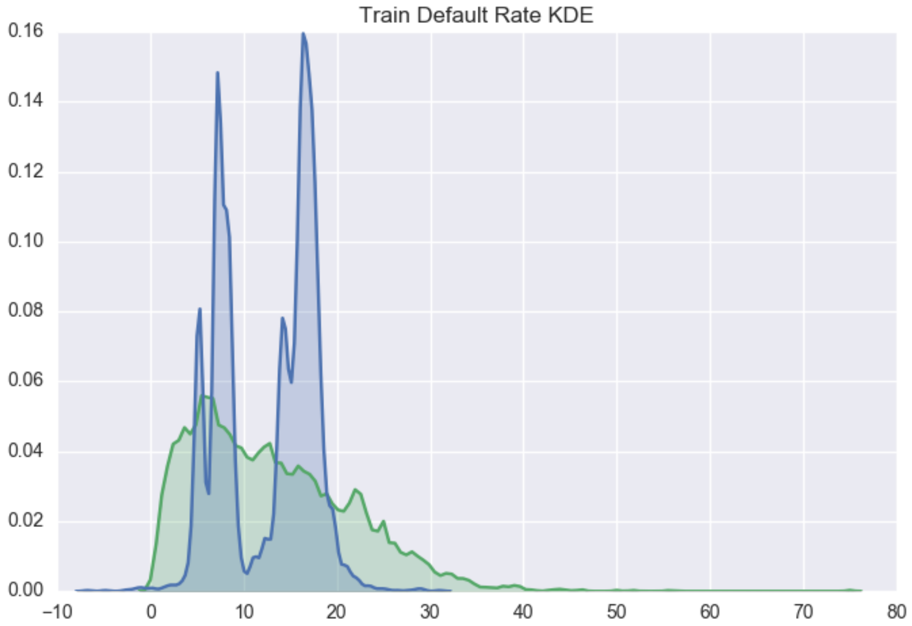
	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Intercept	34.1808	1.779	19.209	0.000	30.692 37.669
ICLEVEL_y__1	-8.8705	0.301	-29.425	0.000	-9.462 -8.280
TOTALCOMPLETIONS_100FTE	-0.0117	0.003	-4.310	0.000	-0.017 -0.006
TOTAL_ENROLLMENT_BLACK_TOT	-6.123e-05	0.000	-0.604	0.546	-0.000 0.000
NET_STUDENT_TUITION	-7.685e-08	1.19e-08	-6.468	0.000	-1e-07 -5.36e-08
TOTAL03_REVENUE	-9.267e-10	3.72e-10	-2.491	0.013	-1.66e-09 -1.97e-10
TOTAL_ENROLLMENT_WHITE_TOT	0.0002	4.96e-05	3.957	0.000	9.9e-05 0.000
STUDSERV01	-8.959e-09	1.44e-08	-0.622	0.534	-3.72e-08 1.93e-08
TOTAL_FULL_TIME	-0.0003	6.92e-05	-3.719	0.000	-0.000 -0.000
STICKER_PRICE_SHARE	-2.1723	0.723	-3.006	0.003	-3.589 -0.756
TOTAL_ENROLLMENT_ASIAN_TOT	-0.0002	0.000	-1.052	0.293	-0.000 0.000
NETTUITION_SHARE	2.2570	0.725	3.114	0.002	0.836 3.678
TOTAL_ENROLLMENT_HISP_TOT	-8.915e-05	6.87e-05	-1.297	0.195	-0.000 4.56e-05
NETTUITION01	7.473e-08	1.28e-08	5.860	0.000	4.97e-08 9.97e-08
INSTCAT__3	6.3849	0.544	11.728	0.000	5.318 7.452
Ethnic_Code__5	-4.0334	0.718	-5.616	0.000	-5.442 -2.625
LOCALE__21	-1.3468	0.264	-5.093	0.000	-1.865 -0.828
CENSUS_REGION__3	1.5743	0.238	6.606	0.000	1.107 2.042
HBCU_x	-6.7791	1.075	-6.305	0.000	-8.887 -4.671

```
In [10]: # Obtain RMSE of our OLS on training data
         mean_squared_error(data_2011['DRate'], lm.fittedvalues) ** 0.5

Out[10]: 6.9187815737197012
```

It seemed as though our model's overall Adjusted R^2 decreased, however, we also note a decrease in our root mean squared error. Interpreted, this shows that our OLS linear model is able to use our limited features to produce a default rate for each institution that is within ~7% of the true value.

Taking a look at the distribution of the training data and our model's fitted values, we see our model's performance is not ideal, but performant nonetheless:

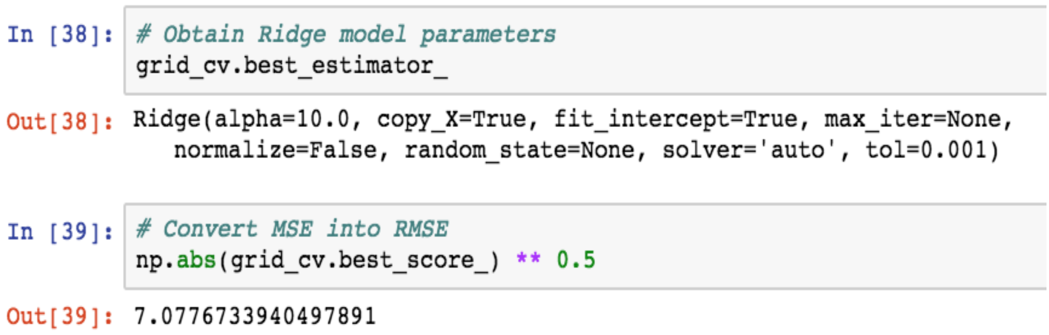


Note: although quite evident, our model's fitted values are shown in blue whereas the true distribution is in green.

Ridge Regression

RidgeCV

RidgeCV implements ridge regression with built-in cross-validation of the alpha (penalty) parameter. The object works in the same way as scikit-learn's GridSearchCV, and returns the ideal configuration of our Ridge model fit onto our training data. After running RidgeCV with the intent of hyperparameter optimization, we arrive at a model that, surprisingly, underperforms relative to our baseline:



```
In [38]: # Obtain Ridge model parameters
         grid_cv.best_estimator_

Out[38]: Ridge(alpha=10.0, copy_X=True, fit_intercept=True, max_iter=None,
            normalize=False, random_state=None, solver='auto', tol=0.001)

In [39]: # Convert MSE into RMSE
         np.abs(grid_cv.best_score_) ** 0.5

Out[39]: 7.0776733940497891
```

In an attempt to improve our model, we wanted to see if we could decrease the variance in our OLS by introducing some bias via regularization. However, upon using a grid-search with 5-Fold cross-validation, we found that the introduction of regularization only increased the bias in our model and, thus, increased error instead of reducing variance.

Evaluation on test data

Given our feature-limited OLS linear model performed better than our RidgeCV, we set out to evaluate its performance on our test data: all federal loan default rates across new graduates in 2012. Evaluating, we see the following:

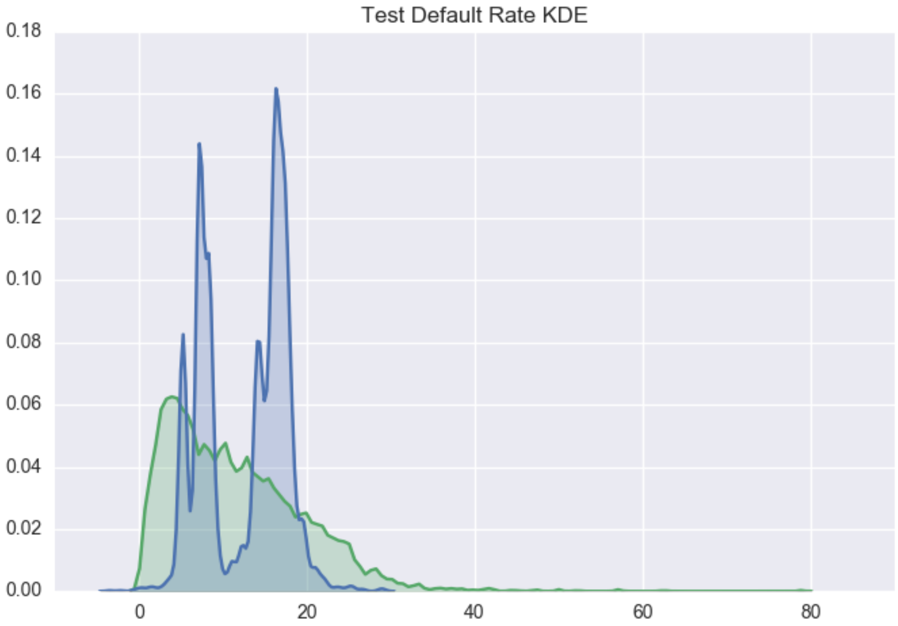
OLS Regression Results

Dep. Variable:	DRate	**R-squared:**	0.367
Model:	OLS	**Adj. R-squared:**	0.365
Method:	Least Squares	**F-statistic:**	177.6
Date:	Fri, 09 Dec 2016	**Prob (F-statistic):**	0.00
Time:	16:26:07	**Log-Likelihood:**	-14079.
No. Observations:	4309	**AIC:**	2.819e+04
Df Residuals:	4294	**BIC:**	2.828e+04
Df Model:	14		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[95.0% Conf. Int.]**
Intercept	30.6890	1.623	18.909	0.000	27.507 33.871
ICLEVEL_y__1	-9.1155	0.234	-38.983	0.000	-9.574 -8.657
TOTALCOMPLETIONS_100FTE	-0.0173	0.002	-7.182	0.000	-0.022 -0.013
TOTAL_ENROLLMENT_BLACK_TOT	0.0003	8.4e-05	3.624	0.000	0.000 0.000
TOTAL03_REVENUE	-1.064e-09	4.07e-10	-2.618	0.009	-1.86e-09 -2.67e-10
TOTAL_ENROLLMENT_WHITE_TOT	0.0002	4.41e-05	4.498	0.000	0.000 0.000
STUDSERV01	-8.936e-09	1.17e-08	-0.762	0.446	-3.19e-08 1.41e-08
TOTAL_FULL_TIME	-0.0001	5.05e-05	-2.653	0.008	-0.000 -3.5e-05
TOTAL_ENROLLMENT_ASIAN_TOT	-0.0002	0.000	-1.161	0.246	-0.000 0.000
TOTAL_ENROLLMENT_HISP_TOT	6.905e-05	5.57e-05	1.239	0.215	-4.02e-05 0.000
INSTCAT__3	6.1310	0.464	13.199	0.000	5.220 7.042
HBCU_y	-5.6916	0.982	-5.798	0.000	-7.616 -3.767
Ethnic_Code__5	-3.6906	0.660	-5.594	0.000	-4.984 -2.397
CENSUS_REGION__3	1.7024	0.232	7.346	0.000	1.248 2.157
CENSUS_REGION__1	-1.9960	0.254	-7.849	0.000	-2.495 -1.497

```
In [43]: # Obtain RMSE of our OLS on testing data
         mean_squared_error(data_2012['DRate'], lm_predictions) ** 0.5

Out[43]: 6.5203009031565111
```



Note: although quite evident, our model's fitted values are shown in blue whereas the true distribution is in green.

Back to the business problem

By looking at the coefficients of each features in the regression model, we can find out which features 'caused more' of the default. Thus, students with loans can look at the historical data about these specific features and decide whether or not they should consider a particular class of schools.

By looking at the coefficients of each feature in the regression model, we can find out which features 'caused more' students to default. Thus, student with loans can take a look at the historical data about these specific features and decide whether or not they should accept the offer. Our results show that the categories describing a school's program lengths, the category of degrees the school grants, net-share of tuition that students pay, and if the school is a historically black college or university are most informative. Essentially, our overall model tells us that schools granting associate's degrees, with shorter program lengths, and that have tuition fees that share the majority of the total cost tend to have higher average default rates. In particular, schools in the southern census region appear to cluster at higher default rates, which indicates there are some demographic, cultural, or economic factors at hand that tend to lead to higher default rates. Our final result found that HBCUs tend to have lower default rates, which could indicate that students attending HBCUs tend to have better financial and academic support.

To choose our best algorithm, we chose to use the root mean squared error (RMSE) as an indicator of algorithm optimality. Notably, the RMSE also gives our model an easy, natural interpretation: the aggregate features associated with this institution are more likely to result in any given student defaulting on their federal student loans to within ~7%.

Deployment

The main goal of our model is to advise students when selecting institutions for secondary education and to better inform legislators about the current conditions of student default rates. Deployment, in the sense of information, would be in the form of reporting this information and allowing end users to access predictions of default rates a year out. In particular, if an institution matches the criteria of factors that lead to higher default rates, then a warning could be brought up and we could direct people to some literature on taking loans, scholarships, or alternative lines of credit!

While monitoring the model, those responsible for informing students (if not the students themselves) would have to be aware of concept drift. We expect deployment to be iterative: given shifting political, economic, and educational landscapes we expect student default rates to be influenced by numerous factors outside of their direct control. Or, for instance, if our new president decides to reduce the amount of money they spend on education this will lead to less money for student loans. Thus, it is important to keep retraining the model at different points in time to account for such variation.

With respect to ethical concerns: the largest drawback in making recommendations or causal inference is the lack of student level information. The claims we can make only dive as deep as the aggregate institutional characteristics that lead to default rates. With this said, we may be unethically, and unfairly, telling students to be aware of default rates based on their interest in an institution alone. Additionally, we have little insight as to why an individual may default. According to our model, if we were to take it quite literally, we may tell a poor student that wants to go to a short-term vocational school to avoid taking a federal loan, when in fact we don't actually know if that individual has any reason to default other than that large portions of students that attend these schools have higher default rates.

The main approaches to mitigating these risks would include transparency with students on the strict definitions of our findings (the lack of student level data) and to search for student level data. When we 'deploy' the model and inform students of potential results, the lack of causal inference at the student level must be clear. And, as we discuss in *Future Work* section, we would search for student-level data to integrate into our model for further testing.

Future Work

In our model, we elected to impute missing values based on a difference of means test for each feature. Another potential path would have been to craft a more engineered set of features by clustering missing values. The clustering could have been accomplished using either k-means or hierarchical clustering. However, we chose not to go this route primarily due to time constraints and in order to avoid an overly engineered solution. In the end, also would have been introducing much more bias into our model by training a clustering algorithm on such a small subset of the institutions that had labels.

Likewise, one of the biggest takeaways from our data mining was the need for additional data – a classic finding. In the end, we have found some causal relationships for what institutional factors potentially attract or potentially influence students at the aggregate level to default. There is an essential break between the causality of a student taking a loan, then attending an institution, and then defaulting on their loans due to the lack of features that directly describe students interaction with taking on loans. Furthermore, given most of the student demographic features were either limited by missing values or proved insignificant in our model, our models are lacking in the student-level features we even obtained. There is a clear need for deeper, student-centric level data. In particular, data on individual students at these universities and their respective financial, demographic, and other qualities such as: other student loans taken, other lines of credit, credit history, family credit history, family income, personally paying for college or not, scholarships, employment history, parental occupation, and more. [4]

References

1. Skipper Seabold and Josef Perktold. Statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.