To College or Not to College: Prediction of Post Collegiate Earnings and Debts. An exploration of the College Scorecard Dataset

Rakshitha Bhat and Nithyashri Govindarajan
University of Massachusetts Amherst
Amherst
rbhat@cs.umass.edu, ngovindaraja@cs.umass.edu

1 Introduction

Deciding to pursue a higher education and choosing a college can be a daunting task. There are many factors that influence this process, and the consequences are life changing. As data driven decision making makes great strides in human lives, college decisions can also greatly benefit from analysis of credible data that reflect current trends in education as well as future job prospects. According to the U.S Department of Education, student borrowers owe more than one trillion dollars and about seven million borrowers are in default[1]. With increasing costs of higher education as well as the accompanying rise of student debt, prospective students can greatly benefit from knowing about post collegiate earnings and the correlation of these earnings to the size, quality of education, location of college or any other factors, if any. Additionally, these could be a useful indicator for the recovery of loans. This project aims at using machine learning tasks to do exactly this.

Data driven decision making has made a foray in most fields but surprisingly, there is very little research in this area. Machine Learning has been used to make predictions about earnings in the corporate sector[2]. A collation of census data has been used to make predictions about income of individuals[3]. However, no prior work has explored a machine learning approach to predict debt and earnings post graduation using any college statistics. This is potentially because of lack of authentic data from credible sources. It is important to note that it may be detrimental to higher education institutions if a solely numbers based approach was employed in the college selection process.

The U.S. Department of Education launched College Scorecard in September 2015[4] to gather and credibly process data on institutions of higher education, the demographics of currently enrolled college students, and the details of the graduated or other alumni of these institutions. This was the first step in enabling access to authentic data to students to help aid their college selection process. The data however has raised eyebrows because of the following caveats - it reports findings for only students with federal aid and has omitted about 700 colleges, particularly community colleges. While this may not still make for a completely accurate picture, the general trends are decipherable by applying machine learning methods that can fill in the holes left by omitted data.

The work presented in this report illustrates strategies to impute incomplete data and obtain a general trend for post collegiate debt and earnings over the years for alumni who were recipients of federal financial aid during their enrollment. Our models use currently available information of these institutions regarding enrollment, completion, tuition and acceptance rates to fill in the gaps with the data and unearth interesting factors that actually determine earnings and debt to exact a trend for these hard to predict but very meaningful values. Matching tuition costs and average potential earnings and debt could help determine typical lengths of student loans. In the scope of this work, we limit our findings to trends of debts and earnings over 19 academic years from 1996-1997 to 2014-2015. While our contributions are majorly in completing the College Scorecard Data, we make a valuable finding of rising principal of debt in the face of a stagnant post-collegiate earnings for these years.

2 Related Work

Prediction of post collegiate earnings and debts has been the subject of research in the social sciences since the 1980s and the 1990s. The most relevant work has been in the field of predicting earnings based on college statistics, where the authors examine these statistics and define college "quality" in accordance with the college type to determine earnings of students currently in high school[6]. It uses student demography to make an assertion that the economic returns of attending a private, elite institution were higher and that these returns are increasing across the years.

Another body of work focuses on specific student demography to predict their post college wages as opposed to purely institution statistics [7]. They determined the variables for only male college students, since their research found that the process for female students entailed a different set of variables. There were additional variables introduced to quantify college "experience" and the labor markets. Their conclusions were that private schools in certain geographies correlated to higher wages, but in most cases they were controlled by the college experience variables, and thus asserted that individual college experiences accounted most for post collegiate earnings over particular colleges.

A landmark study in this field incorporated regression to determine factors that predicted earnings[8]. College quality variables, levels of college education, average SAT scores and faculty statistics were determined to be most influential factors for wages after graduation.

Most of the related work has made great progress in aiding the college selection process, and has managed to define quantifiable variables for factors such as college quality and personal college experience. However, most of the work is based off of data of students in high school or of alumni, and these data points are not all encompassing for a particular institution and may not cover all students and scenarios. Further, most of the student demography data came from private and elite institutions and does not cover all strata of schools, particularly of smaller schools. Such schools may actually be beneficiaries of such study. College Scorecard, however, represents anonymized data from institutions that incorporates both student data and college statistics.

Using Machine Learning for predictive tasks for college earnings and debt has not been explored due to the lack of credible encompassing data. The recently launched College Scorecard data has not had much work based off of it. The data set has been widely circulated by Machine Learning enthusiasts on public platforms like Kaggle[5], where a competition has been defined on this data. These exploratory analysis are conducted on the nature of the data, the number of colleges for which data is available across the years, median SAT scores across universities, median earnings etc.

A recent analysis on predicting collegiate earnings and debts[13] uses various regression methods on the College Scorecard data. The analysis uses a predetermined set of features that are purely non-categorical which makes assumptions about the relative lesser influence of these features. The prediction task runs two separate weighted linear regressions for debt and earnings on data for the singular year of 2011. We build on this task by not discarding categorical variables in the data and extend the predictions for all 19 years worth of data. This enables the determination of a trend for debt and earnings over two decades. Some categorical variables possess valuable information about the colleges - such as geographical location and degree of urbanization, which at first instance seems a useful indicator for the prediction task.

This data set is raw and poses a lot of data cleaning challenges. Our work in this project mirrors work we present in relevant research. There is a lot of research on handling missing data, and this field is fine grained based on the model of computations preferred.

The overview of most major methods [9] was highly useful for the purpose of handling missing data. As an alternative to the deletion of the data point, the strategies to handle missing data include - mean imputation (setting all missing values to the mean of those values present for a feature), regression imputation (training a regression model on observed values and using this to predict the missing values), multiple imputation (using multiple values generated by models on the the observed values and using multiple iterations of the set of total data) and independent random imputation (assigning random values from a list of valid values for the field). These imputation strategies vary based on the type and the number of missing values for all features.

We examined multiple imputations in the work of Rubin [10] where the setup, environment, and performance of multiple imputations on missing data in a database to capture variability of the data that is lost with single imputation is described. This is a simulation technique that replaces each

missing data point with a set of m > 1 plausible values. We decided that we would not adopt this method in our case for the following reasons - first, the multiple imputation setup is efficient when the ultimate users and the database constructors are different entities; and second, maintaining multiple sets of data and later aggregating models is computationally expensive on the infrastructure we had access to.

Other imputation strategies outlined in research involves application of models that can perform Machine Learning tasks even with missing data. Statistical methods that address missing values other than imputation include likelihood and weighting approaches[12]. Each approach is more complicated when there are many patterns of missing values, or when both categorical and continuous random variables are involved. Incomplete data classification problems are addressed using a logistic regression [11] where a single or multiple imputation for the missing data is avoided by performing analytic integration with an estimated conditional density function (conditioned on the non-missing data). Conditional density functions are estimated using a Gaussian mixture model, with parameter estimation performed using both expectation maximization (EM) and Variational Bayesian EM.

In our case, since the data contains a mixture of categorical and continuous variables across a set of colleges and across a set of years, various imputation strategies were chosen - mean, regression imputation and categorical feature replication. We will address the details of these imputations in the next two sections.

3 Dataset

The dataset used for this project is the College Scorecard Dataset. This dataset is publicly available at https://collegescorecard.ed.gov/data/. College Scorecard consists of 1744 metrics for 10896 degree-granting institutions from 1996 through 2016 [4]. These metrics include 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, percentage of students from each ethnicity, college completion rates and many more. The U.S Department of Education launched this data collection drive through tax submissions of federally aided students as a means to help students make more informed college decisions. The data has raised eyebrows because of the following caveats - it reports findings for only students with federal aid and has omitted about 700 colleges, particularly community colleges. Despite this, the data is useful to provide general albeit approximate trends to aid decision making for all stakeholders.

Our first set of tasks were to select the variables to predict, impute the missing values, transform the data set into pairs of features and prediction variables, and segment the data for evaluation purposes. We chose two values for our prediction variables: median postgraduate debt (principal amount upon entering repayment) and median postgraduate earnings for alumni ten years after graduation (for long term effect analysis). We then went through several steps to reduce the full set of features to an initial feature list. In the first step, we eliminated features which were correlated to the target variables. Next, we removed unrelated features that should not be used to make predictions, such as features that provided number of students in different demographic regions and certain categorical features like zip code and school name. Lastly, we removed the set of features which had NULL values for all colleges across all years. At the end, we were left with 517 features and two prediction variables.

All the selected feature values are numerical. Some features in the data set are categorical fields. We chose to transform these into binary vectors of k dimensions where k is the number of categories of the feature. All values in our data set that were computed using data from fewer than 30 students were listed as "PrivacySuppressed". These values are more common with smaller schools than larger and many privacy suppressed values occurred in potentially useful metrics. We had a lot of "NULL" values present as well. One approach for handling these values was to simply remove all features with any missing entries. However, discarding hundreds of features in this fashion, especially for features with a low percentage of missing values, was undesirable. We mirrored the work prsented in Marlin's paper[9] and performed single mean, median, mode and regression imputation on the missing data.

4 Methodology

In this section, we describe our complete pipeline. Much of our work is in handling missing and privacy suppressed values to ensure sufficient complete data can be fed to the regression model. We

Table 1: Dataset Description

Total Number of Features	Categorical	Continuous
517	20	497

enumerate the various methods we use in handling missing information for each type of data, and describe techniques to reduce the dimensionality of the problem. We finally describe our regression pipeline for the prediction task.

4.1 Data Pre-processing

The data set consists of a large number of missing values taking the form of either null rows or privacy suppressed rows. Discarding the data points containing missing values severely reduces the training set and important information may be lost. For this reason we perform imputations.

4.1.1 College-Wise Imputation

The data set provides a list of CSV files containing yearly information of the features across colleges. In order to obtain college based trends across the years, we grouped the features based on colleges instead of years. This provides a more accurate representation of the trend in the features. For example if we were to fill in the value of percentage enrollment for a particular college, imputing the value by observing the trend in enrollment across years for that college is more accurate than performing the imputation across colleges for that year. Hence we chose to perform college-wise imputation on all the features.
The data set contains two types of features: Categorical and Continuous valued features. Most of the categorical features like Locale, Region do not change over time for a college. Hence we decided to perform mode imputation for categorical values. Initial analysis performed on a few colleges gave an accuracy of 98% for mode imputations which suggested that our choice of method was appropriate. For continuous valued features, we chose to fit a regression line across the values of a feature over the years. Our assumption was that this method provides a good representation of the trend in features. In order to verify our assumption, we performed a train test split and measured the Root Mean Square Error (RMSE) scores for missing features for a college. Since the RMSE scores obtained were low, we decided to go ahead with our assumption. The RMSE scores are mentioned in the Results section.

For features which had only one available value across years, the above method cannot be used because a minimum of two points are required to plot a regression line. Hence for such features we use an average trend line of that feature for all colleges across all years. We plot the feature values available for all years and all colleges against the years and fit this to a regression. We take the one value we know for a particular college and shift the line vertically to fit through the data point. This gives us a shifted line to make predictions for other years for this feature. The assumption here is that if the college displayed a positive or negative trend above or below the average curve for one year, the same applies for the other years. The coefficients of the regression line remain the same whereas the intercepts change (increase or decrease by the amount of shift in that dimension) in this shifting.

4.1.2 Year-Wise Imputation

Some of the features for certain colleges had null values across all years. Hence these values could not be imputed using college wise trends due to the absence of data points. For such features, we performed a year wise imputation. The set of non-null data points for a feature were grouped together. These data points were then split into train and test sets and mean, median and mode imputation were performed.The RMSE scores were recorded. The imputation strategy which gave the least RMSE score was chosen and imputation for all the data points in that year (including the ones with missing values) was performed. Features which had missing values for all data points in a year were discarded since there was no method of estimation to move forward. Features which had existing values for all data points were left untouched.

Table 2: Categorical variable encoding for a 3-category variable

Category	Original Value	Encoded Value
Public	1	[1, 0, 0]
Private non-Profit	2	[0, 1, 0]
Private for-profit	3	[0, 0, 1]

4.2 Data Transformation

The completely filled data, segmented year wise, contains filled categorical features that cannot be directly used for regression. The categorical variables are coded into a k-dimensional binary vector. This maps a column of category indices to a column of binary vectors, with exactly a single one-value per row that indicates the input category index. We chose to perform coding retaining all categories as opposed to dropping any (last) category for a k-1 binary vector. This ensures that there is no reference category that would have been represented under the drop-one system of k-1 bit binary vector as all zero values. Though certain implementations drop one category to avoid complete collinearity, we choose to retain all k dimensions to reduce any omitted variable bias. We perform this at the worst case for twenty categorical features with ranges between 2 categories and 20 categories. With 517 features in total, this meant a worst case increase to 597 features. At the end of this encoding step, we have data segmented year wise with all features suitable for regression.

Table 2 illustrates the encoding process for one of the categorical variables "CONTROL" that describes the ownership of each institution.

4.3 Dimensionality Reduction

Prior to performing the regression for the actual prediction task, we choose Principal Component Analysis (PCA) to perform dimensionality reduction for our features. The goal of PCA is to identify the directions of maximum variance contained in the data. It converts a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components in a multivariate setting.

Dimensionality reduction with PCA:

- Given the centered data matrix $X \in \mathbb{R}^{N \times D}$, with column-wise zero empirical mean, we compute the unscaled sample covariance matrix $\Sigma = X^T X$.
- We compute the K leading eigenvectors $w_1, ..., w_K$ of Σ where $w_k \in \mathbb{R}^D$.
- We stack the eigenvectors together into a $D \times K$ matrix W where each column k of W corresponds to w_k.
- We project the matrix X into the rank-K sub-space of maximum variance by computing the matrix product $Z = XW$.
- To reconstruct X given Z and W, we use $\widehat{X} = ZW^T$.

Thus PCA produces principal components that are orthogonal and have variances in decreasing order. For this dataset, when choosing the number of principal components (k), we choose k to be the smallest value so that 99% of variance is retained.

4.4 Regression Pipeline

The data run through PCA is passed to the regression. We use a baseline regressor and a predicting regressor to compare the values produced.

4.4.1 Model Selection

Ridge Regressor
We use Ridge regressor as the baseline regressor. Ridge regression is the name given to regularized least squares when the weights are penalized using the square of the $l2$ norm, which is $||w||_2^2 = w^T w = \Sigma_{d=1}^{D} w_d^2$.

The regression function is defined as: $f_{ridge}(x) = (\Sigma_{d=1}^{D} w_d.x_d) + b = \mathbf{x}.\mathbf{w} + b$ where $\mathbf{x}$ is a linear function with parameters $w = [w_1, ..., w_D]^T$ and the optimized regularized weights are obtained using the equation: $w^* = (X^T X + \alpha I)^{-1} X^T Y$. This shrinks coefficients towards zero.

The advantages of Ridge is that it solves the problem of needing at least D cases to learn a model with a D dimensional feature vector. It also solves the problem of co-linear features and the regularization can reduce the possibility of very large weights overfitting to outliers. It particularly does well when there is a subset of true coefficients that are small. $\alpha \geq 0$ is the tuning hyperparameter, which controls the strength of the penalty term. It controls the amount of shrinkage - the regularization strength. Larger α values specify stronger regularization.

Random Forest Regressor

We use a Random Forest regressor for the prediction task. Random Forest is a random decision forests based ensemble regressor that constructs multiple regression trees and outputs the class that is the mean of the all classes predicted by each estimator.

The base estimator for a Random Forest regressor is a Decision Tree Regressor. The regression tree is a tree based classifier that uses a combination of rules on attributes of data to make predictions. Every internal node comprises rules that compare a single data dimension against a threshold and pass down the data point to either the left or right subtree according to the data value. Trees use recursive binary splitting to make predictions – where this process of comparing a data value (or a dimension) against a threshold is done recursively from the root of the tree to the leaf node of the tree – where each leaf node is assigned a class label and all the data points that flow through to that leaf node is assigned the class of the leaf node. A tree based regressor thus uses a top-down approach beginning from the root to the leaf, and makes greedy choices - a choice for splitting at any step is made based on the best split at that step and not affected by the future steps. Decision Trees have linear boundaries – they divide a prediction space into on-overlapping regions, where all values in a region have the same output label.

Random Forests help avoid overfitting (as opposed to using just a decision tree regression), essentially being a bagging technique.

They requires very little data pre-processing or normalization, make fast predictions and if some data is missing, they still make a prediction by averaging all the leaves. They work well with cases where the true regression surface is not smooth, which we expect for our data set, hence inspiring this choice of regressor. The key hyperparameters are the maximum depth of tree – that represents the maximum height any tree can have before being cutoff and labels attached to each leaf, minimum number of samples at split - that decides the minimum number of samples needed to split at an internal node, and minimum number of samples at a leaf node - if a node has enough samples flowing through to it to be tagged a leaf node. We only optimize for maximum depth and number of base estimators.

We do not perform feature selection because we have pre-processed the data using PCA to reduce the dimensions, and the number of PCA components are computationally tractable to this problem. Further, PCA captures variability and these components are better equipped to handle the trends in data.

4.4.2 Hyperparameter Optimization

Trees have a tendency to overfit, but the aim is to produce models that generalize well without overfitting. Since decision trees are rule based in essence, the overfitting can happen because of either noisy data or lack of sufficient representative samples. There is a danger of making an incorrect prediction due to noise, or simply not enough representative samples at the leaf node. We pick K-Fold Cross Validation to tune hyperparameters. K-Fold Cross Validation helps by having the training set split into K folds, where each fold is trained on K-1 times and tested on only once, and the optimal tree which minimizes the error is chosen. This way, trees are exposed to a varying combination of training data and tend to not overfit the entire training set – they are expected to generalize over the 1 fold held out for testing in each iteration. The test set is non-overlapping across all the iterations. A good value of K will control high bias (low value of K) and high variation (high value of K), and thus find a balance in bias versus variance of the tree. Thus K-Fold is a more tractable solution. GridSearchCV is used in the implementation and it does the cross validation while fitting over a range of hyperparameter values.

We set up three regression tasks with the PCA reduced data - prediction of debt alone, prediction of earnings alone and prediction of both together. All three are processed use the same pipeline. We use Root Mean Square Error (RMSE) as a measure of accuracy of the regression task. It represents the square root of the average of the square of all of the error. This choice was picked because RMSE amplifies and severely punishes large errors.

5 Experiments and Results

We perform our experiments in Python 2 with scikit-learn package of 0.18.

5.1 Data Pre-Processing

During the imputation process, we perform linear regression college wise. For this process, we use a train-test split of 70%-30% for all observable values. As an example, the RMSE scores reported for 5 features for a college ID 100654 were [0.001157, 0.0002432, 0.012684, 0.012000, 0.0001829]. These scores are considerably low and hence we decided that this method gives a valid approximation of the missing features.

For year-wise imputation, all three strategies were explored (based on mode, median and mean) and the one with the least RMSE value is picked and run on that feature.

5.2 Dimensionality Reduction

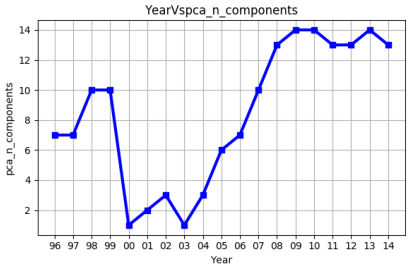


Figure 1: Number of Components with greater than 99% variance in each case

We perform PCA with the number of components set to 0.99, which is indicative of picking the smallest value of number of components where 99% of variance, is retained. The number of components with this cutoff for each year's data is illustrated in Figure 1.

5.3 Regression

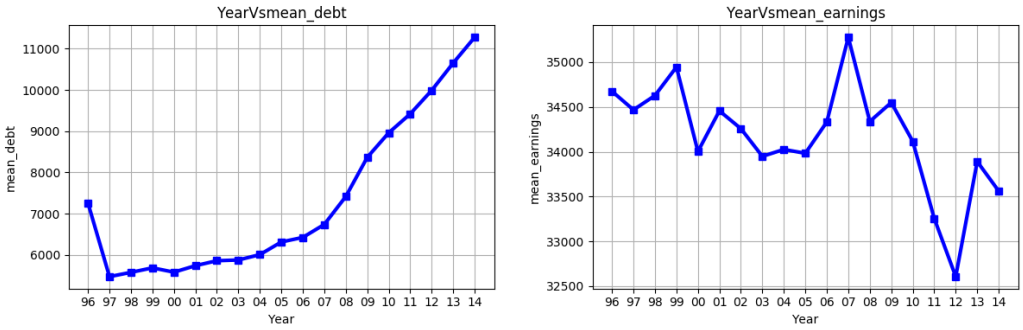
The setup for regression uses a 5 fold cross validation for hyperparameter values. This was performed on a train set of 70% of the total data, with 30% heldout as a test set. The RMSE scores are reported for this heldout test set. RMSE scores for debt are lower than for earnings, but since the dollar amounts for debt are typically lower than those of earnings, there will be a higher percentage error for debt prediction. Hence a decrease in RMSE value for debt does not necessarily mean that the regressor performs better on debt predictions. Figure 2 shows that Random Forest outperforms the baseline regressor across the years for all prediction variables. We also compute the mean values of the predictions - the median debt (principal amount upon entering repayment) and the median post ten year earnings.

6 Discussion and Conclusions

Figure 3a depicts an increase in the mean median debt values (principal amount) over the years. A possible explanation for this is the increase in the cost of education over the years. This would have led to an increase in the loan principal thereby increasing the debt.

Regressor	Debt RMSE Scores (Ridge)	Debt RMSE Scores (RF)	Earnings RMSE Scores (Ridge)	Earnings RMSE Scores (RF)	Multivariate RMSE Scores(Ridge)	Multivariate RMSE Scores(RF)
Year 2010	3179.67	2710.26	11363.03	10081.66	8343.23	7306.46
Year 2011	3361.00	2867.83	10447.58	8734.85	7761.83	6409.07
Year 2012	3341.17	2915.80	11162.99	9567.25	8248.14	7064.82
Year 2013	3675.35	3150.62	9392.83	8307.33	7132.09	6303.75
Year 2014	3886.16	3275.38	10257.05	8595.76	7767.93	6367.91

Figure 2: RMSE Values For All Regressors



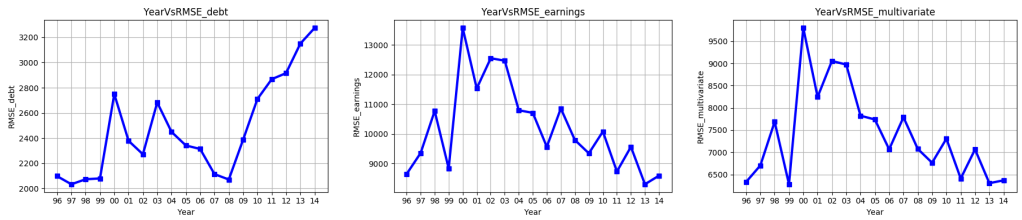
(a) Mean Predicted Debt Values (b) Mean Predicted Earnings Values

Figure 3: Mean Prediction Values Versus Years

Figure 3b depicts a fluctuation in the median earnings of alumni ten years after graduation. We could assume that this variability is dependent on the economy. The peaks coincide with the years of known good economy performance and the dips coincide with the years of known economic struggles and recession. As an example, the median earnings steadily decreased following the recession of 2008.

Figure 4 shows the RMSE scores on the test set for each year of computation. The RMSE scores for debt are lower than for earnings, and those of the multivariate regression are in between. Notice that 4c mirrors the trend in 4b, indicating a greater influence of earnings. This is because of the greater dollar amount of earnings compared to debt. There is lesser missing data for the later years, and this is reflected in the lower RMSE scores since there is a lesser chance of imputation errors creeping in.

Our models successfully impute data to over 25,000 missing values per year and produce explainable, expected trends in results for debt and earnings with a computationally tractable pipeline. Thus, given a proper disclaimer, they can be used to fill in gaps in the current College Scorecard data set for a deeper analysis and an exhaustive exploration.



(a) RMSE for Debt Prediction (b) RMSE for Earnings Prediction (c) RMSE for Both Prediction

Figure 4: RMSE Versus Years

References

[1] https://www.wsws.org/en/articles/2015/08/26/debt-a26.html, August 2015.

[2] Kamp, Michael, Boley, Mario & Gartner, Thomas. (2013) Beating Human Analysts in Nowcasting Corporate Earnings by using Publicly Available Stock Price and Correlation Features.

[3] Center for Machine Learning and Intelligent Systems. Census Income Data Set, 1996.

[4] U.S. Department of Education. College Scorecard, 2015.

[5] https://www.kaggle.com/benhamner/d/kaggle/college-scorecard/exploring-the-us-college-scorecard-data . https://www.kaggle.com/kaggle/college-scorecard.

[6] Brewer, Dominic J., Eide, Eric R.& Ehrenberg, Ronald G (1999) Does It Pay to Attend an Elite Private College? Cross-Cohort Evidence on the Effects of College Type on Earnings. *The Journal of Human Resources*.

[7] James, Estelle, Alsalam, Nabeel, Conaty,Joseph C,& To Duc-Le (1989) College Quality and Future Earnings: Where Should You Send Your Child to College?: *The American Economic Review*.

[8] Solmon, Lewis C (1975) The Definition of College Quality and Its Impact on Earnings. *The National Bureau of Economic Research*.

[9] Marlin, Benjamin M (2008) Missing Data Problems in Machine Learning. University of Toronto.

[10] Rubin, Donald B. (1996) Multiple Imputation After 18+ Years. *Journal of the American Statistical Association*.

[11] Williams, Daniel, Liao, Xuejun, Xue, Ya, & Carin, Lawrence (2005) Incomplete-data classification using logistic regression.*International Conference on Machine Learning* 972-979.

[12] Horton, Nicholas J. & Kleinman, Ken P. (2007) Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician* **61**(1):79-90.

[13] http://cs229.stanford.edu/proj2015/212report.pdf