SAS® GLOBAL FORUM 2021 #SASGF

Paper ###-2021

Investing In Education Intelligently

Sam Edison, Akhil Emani, Nithisha Katta, Rushya Puttam
Oklahoma State University

ABSTRACT

Student loans are a critical resource used by many Americans to help pay for the goal of achieving higher education. As the tuition for post-secondary education continues to rise, the need for understanding student loans and the variables that impact individuals' abilities to pay them is vital. In a recent survey, Country Financial found that only 9% of parents talk to their children about managing student loan debt (PR Newswire, 2019). Many borrowers are teenagers who have little formal financial education and even less experience in such financial matters. The College Scorecard is an online tool assisting students to evaluate higher education options. Created by the United States government, the tool helps address the disconnect experienced by future students and the high cost of advanced education. Using the Statistical Analysis Software (SAS), this study expands on data compiled in the Scorecard to provide further insight into possible student outcomes based on institution-level data.

INTRODUCTION

Many high school students in the United States are excited to continue their education at post-secondary institutions, allowing them more freedom and the option to explore various subjects. This education frequently comes at a premium cost. Students must rely on hard-earned scholarships, grants, parent-funding, student loans, or a combination of these resources. U.S. News reported that 42 million, or one in six American adults, currently carries a federal student loan. Recent calculations for student loan debt in America is $1.6 trillion (U.S. News & World Report, 2019).

Students may not have the financial foresight, education, or awareness of possible career pathways in order to project the outcome of the cost of student loans versus possible earnings after graduation. This study is designed to evaluate post-secondary institution costs and common outcomes to increase the knowledge of current and future students looking to achieve their educational goals.

PROBLEM

The objective of this study is to use visual and predictive analytics to identify likely outcomes for individuals seeking to invest in higher education. In the incurrence of student loan debt, it is important to consider post-graduation income and the ratio of debt-to-income (DTI). These outcomes will be thoroughly explained in the report with examples of how DTI may rise or fall given specific circumstances, such as, degree acquired or U.S. region for each institution.

Analysis within this report uses software, such as, SAS Studio and Enterprise Miner, to build predictive models targeting a debt-to-income ratio.

DATA

The dataset used for this study is compiled annually by the U.S. Department of Education, currently available from academic years 1996-97 to 2018-19 (*College Scorecard*). Each year represents an individual file containing aggregate data for each institution, including information on institutional characteristics, enrollments, student aid, costs, and student outcomes. The analysis provided in this report utilizes the 2014-15 academic year, containing over 7,000 rows (institutions) and nearly 2,000 variables. The dataset compiled for the 2014-15 academic year represented the most completed observations for the target variables. Many data elements within the College Scorecard are only available for federal financial aid recipients. These data are reported at the individual level to the National Student Loan Data System (NSLDS), which is used to distribute federal aid, and published at the aggregate institution level. Any elements containing aggregations with fewer than 30 students in the denominator are represented as "PrivacySuppressed" to ensure data are as representative as possible.

DATA PREPARATION

Data Cleaning, Imputation, and Reduction

Variables containing NULL (empty values) or "PrivacySuppressed" were replaced with blank values to ensure correct interpretation by SAS to define nominal and interval variables. Institutions with greater than or equal to 50% blank values (formerly NULL or "PrivacySuppressed") were dropped from the dataset. Further, institutions which did not belong to one of the 50 U.S. states were removed.

Missing values were handled for both interval and categorical variables by imputing the median for interval variables and the mode for categorical variables. Variables with missing values greater than 50% were removed from the dataset.

Due to the number of variables within the dataset, multicollinearity could be an issue leading to less reliable results in prediction. Multicollinearity is the occurrence of strong relationships between two or more variables in a predictive model (*Hayes*). This will be addressed in the modeling portion of the report.

Data Derivation

Individual state based analysis would result in too many levels to effectively explain results, thus, a region variable was created that buckets states into four regions: Northeast, Midwest, South, and West (Appendix B). The target variable DTI, was also created using SAS code and will be addressed in the analysis section of the report.

Data Transformation

The interval variables in the dataset have values with varying ranges from small to large. The small values range from 0 to 1 as decimals. The large values range from 0 to 100,000. In order to compare all variables in the dataset, they need to be on a similar level. Thus, a series of transformations were applied to the large values to scale them down to be comparable with the small values (Appendix B).

ANALYSIS

TARGET VARIABLE

Median Debt defines the accumulated amount of federal loans received for all student borrowers. The median debt level is the aggregate of students who complete a degree at an institution. Median Earnings is an important consideration for students and an indication of institution value. Although students enroll in diverse programs of study, their earnings reflect the labor market's valuation of the education acquired in school. This measure provides the median earnings 6 to 10 years after enrolling in an institution, while excluding students currently enrolled (*Full Data Documentation*, *36*).

Determining the return on investment of attending a post-secondary institution has certain dependencies that are assumed in this study, such that, the student will graduate and join the work force at least six years after beginning their education at an institution. Furthermore, the interest rate (5%) and term (10 year) used in the student loan payment calculation were generalized to avoid assumptions that rates will be in favor of the future student (studentaid.gov). The target variable, *debt-to-income ratio* (DTI), is calculated by dividing the monthly payment by the monthly income as seen in the formula below:

P = Median Debt
E = Median Earnings
i = interest rate (5%)
t = term (120 months)

Monthly Payment = P/t + (P*i)/12
Monthly Earnings = *E*/12
DTI = Monthly Payment / Monthly Earnings

Table 1. Calculation for DTI

VISUAL ANALYTICS

Target Variable Correlation

This chart displays how individual institutions (bubbles) can be categorized from the target variables: Earnings, Debt, and DTI. The Earnings are shown along the x-axis, while the y-axis displays the Debt. The DTI is represented as both the size of the bubble and a color saturation, where large and red indicates a high DTI, as opposed to, blue and small being low DTI. Furthermore, the chart is broken down into four quadrants (indicated by the black lines). While most of the data appears to cluster near the center of the chart, there are instances of debt rising with earnings, demonstrating a moderate positive relationship (correlation: .3 to .5) between the Earnings and Debt variables.

Quadrant 2 (top left) describes institutions with low earnings but high debt. These institutions students should take more caution and care when considering attending.

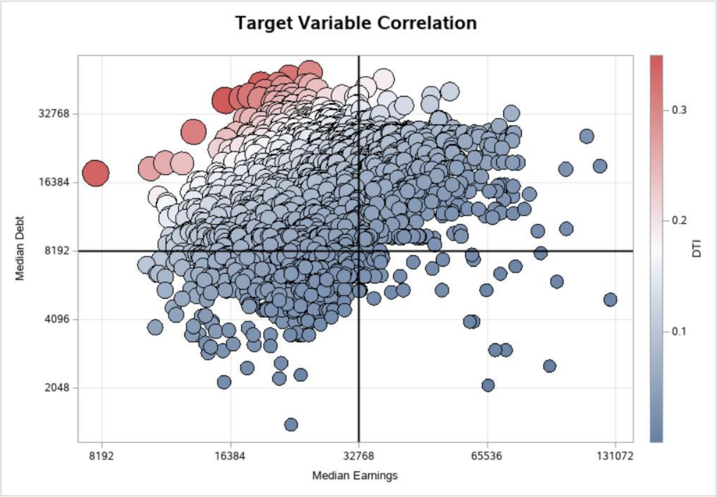


Chart 2. Target Variable Correlation

Quadrant 4 (bottom right) holds the best opportunities from a DTI perspective where each institution has high earnings and low debt. Unfortunately, this quadrant represents the fewest amount of institutions in the dataset.

Median Earnings and Debt by Predominant Degree and Region

The predominant degree (Chart 2) identifies the type of award that an institution primarily allocates. The bars indicate the earnings and the line representing the debt. Ultimately, the desire is for the bar to exceed the line (lower DTI ratio). This occurrence appears in the bar for Master's degrees, meaning that at institutions where Master's degrees are predominantly given, the DTI is lower. Certificates also rank well, as the line and bar are nearly connecting. Meanwhile, Associate's and Bachelor's display a deficit to be considered in weighing options.

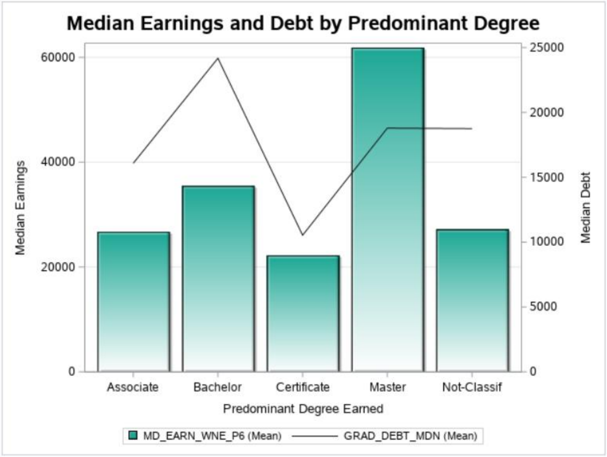


Chart 2. Predominant Degree

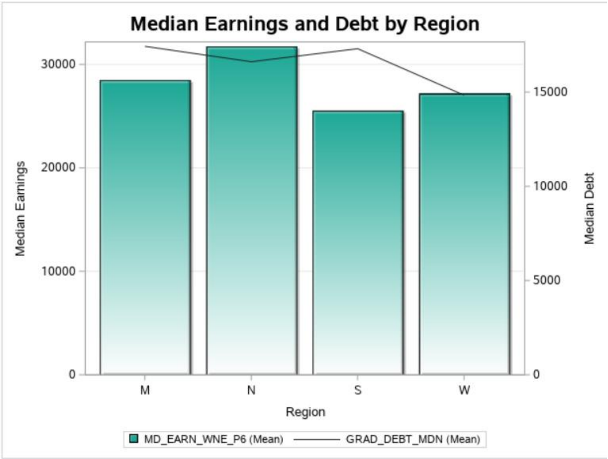


Chart 3. Target Variable Correlation

Various Regions (Chart 3) were created from the State variable within the dataset (see appendix for detail). Using the same logic as the previous chart, the Northeast (N) and West (W) bars represent the best DTI across all regions. The Northeast shows the highest amount of debt but alternatively, the highest earnings. Meanwhile, the Midwest (M) and South (S) maintain higher debt with lower earnings.

MODELING

The variable selection node was used to identify the variation explained by each predictor variable in explaining the debt-to-income target. The variable selection yielded 30 variables remaining for modeling. To understand the effect of variable selection in the rejection model, two baseline regression models are run with and without variable selection:

Sl. No	Model	Mean Square Error
1	Regression without variable Selection	0.0149
2	Regression with variable selection	0.037

Table 2. Model Evaluation

The desire is to have a lower mean square of error (MSE). The MSE measures the average of the squared distance between the predicted values and the actual values. The lower MSE is found in the model run after variable selection. Thus, this criterion is considered to build a series of models and to select the champion model.

On the reduced dataset (30 variables), a series of models such as regression (with forward, backward and stepwise), random forest, neural network with default setting, neural network with 2 hidden layers, neural network with five neurons and high-performance tree were built. All these models are run and compared using Model compare node with the criteria as Average Squared Error. Random Forest turned out to be the best model with an average squared error of 0.00021. To interpret the model results, the most important variables given by this model are passed into the regression model to understand the variables effect on the debt-to-income ratio.

The top three important variables resulted from random forest model are *Percent who transferred to a 4-year institution and withdrew within 4 years*, *Three-year repayment rate for first-generation students* and *Percent of not-first-generation students who completed within 2 years at original institution* (see Appendix C for the full list of important variables).

The parameter estimates of these important variables resulted from the regression shows which factors influence the debt-to-income ratio positively and negatively. Here, the variables which have the negative estimates coefficients are favorable in decreasing the debt-to-income ratio, the desired outcome. The variables that decrease the debt-to-income ratio are *Graduate Degree*, *Certificate Degree*, *Percentage of degrees awarded in Personal and Culinary Services*, and more (Appendix C). Furthermore, the factors that increase the debt-to-income ratio are *Instate tuition fees*, and *Bachelor's degree in Communications Technologies/Technicians And Support Services*.

SUGGESTIONS FOR FUTURE STUDIES

In this study, the focus was solely on post-graduation outcomes for debt and earnings for graduating students. Further assessment of an institution's acceptance and completion rates used in conjunction with the student's test scores and high-school GPA could suggest a future student's likelihood to succeed, regardless of cost. Furthermore, the College Scorecard has become a widely used tool which is still being expanded. A secondary dataset, *field of study*, began forming in 2019 to include disaggregated data elements describing post-graduation earnings and cumulative loan debt of graduates by field of study and degree earned. This disaggregated data would provide more granular insights into the student outcomes for each individual field of study and degree types at a specific university, as opposed to, the aggregated version of the Scorecard used in this study. For example, a student could compare an Engineering major to a Computer Science major at a given institution. Lastly, the Department of Education offers an API to directly access the Scorecard in a SAS compatible interface, such as SAS Viya. An API could benefit a company targeting students with loans by designing a process which models the Scorecard data and markets to students that may need help with their debt (i.e. loan consolidation).

CONCLUSION

While the aggregated dataset has its limitations not allowing granular detail for modeling, many results were identified that can further educate individuals on the subject of student loans. It was seen from the visuals that institutions predominantly offering Master's and Certificate's degrees represented a lower DTI than institutions predominantly offering Associate's and Bachelor's degrees. Regions, such as the Midwest and South demonstrate higher DTI's due lower earnings than the Northeast and higher debt than the West.

The Random Forest Model reiterated that future students should seek Master's and Certificate degrees to lower DTI. Additionally, pursuing a degree in Personal or Culinary services offers a lower expected DTI. The factors increasing DTI were partly due to institutions with high *Instate tuition and fees*, as well as, institutions predominantly offering Bachelor's degrees.

REFERENCES

Financial, COUNTRY. “Parents Just Don't Understand (Finances): 3 in 5 Americans Rely on Uncertain Parents for Financial Guidance.” *PR Newswire: News Distribution, Targeting and Monitoring*, 18 Sept. 2019, www.prnewswire.com/news-releases/parents-just-dont-understand-finances-3-in-5-americans-rely-on-uncertain-parents-for-financial-guidance-300920509.html.

Hayes, Adam. “Multicollinearity.” *Investopedia*, Investopedia, 22 Oct. 2020, www.investopedia.com/terms/m/multicollinearity.asp.

Federal Student Aid, studentaid.gov/understand-aid/types/loans/interest-rates.

Federal Student Aid, studentaid.gov/manage-loans/repayment/plans.

“Student Debt Explained: Breaking Down the $1.6T in Loans.” *U.S. News & World Report*, www.usnews.com/news/elections/articles/2019-11-01/student-debt-explained-breaking-down-the-16t-in-loans.

College Scorecard, Office of Planning, Evaluation and Policy Development (OPEPD), collegescorecard.ed.gov/.

Full Documentation, Office of Planning, Evaluation and Policy Development (OPEPD), https://collegescorecard.ed.gov/assets/FieldOfStudyDataDocumentation.pdf.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

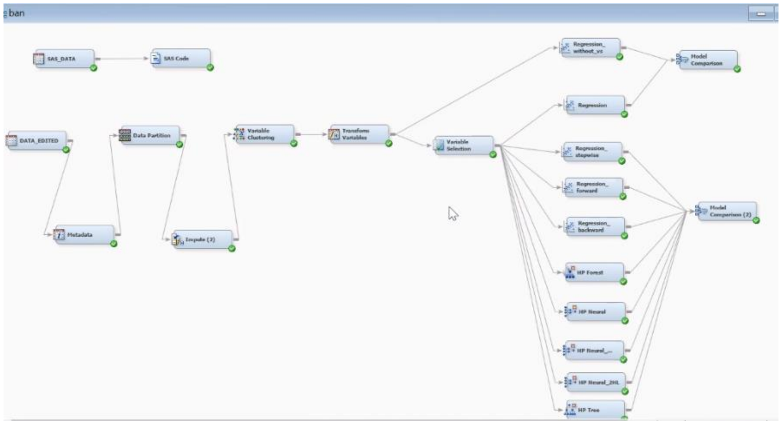
Samuel Edison
Samuel.edison@okstate.edu

Akhil Emani
aemani@okstate.edu

Nithisha Katta
nkatta@okstate.edu

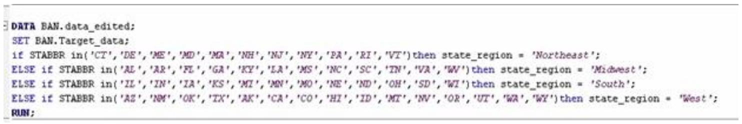
Rushya Puttam
rputtam@okstate.edu

APPENDIX A: WORKFLOW

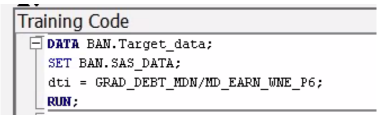


APPENDIX B: DATA PREPARATION

DERIVED VARIABLES

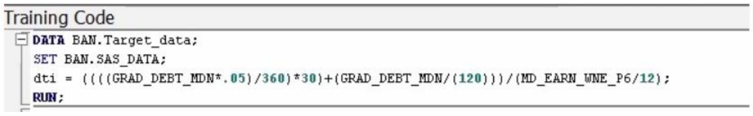


```
DATA BAN.data_edited;
SET BAN.Target_data;
if STABBR in('CT','DE','ME','MD','MA','NH','NJ','NY','PA','RI','VT')then state_region = 'Northeast';
ELSE if STABBR in('AL','AR','FL','GA','KY','LA','MS','NC','SC','TN','VA','WV')then state_region = 'Midwest';
ELSE if STABBR in('IL','IN','IA','KS','MI','MN','MO','NE','ND','OH','SD','WI')then state_region = 'South';
ELSE if STABBR in('AZ','NM','OK','TX','AK','CA','CO','HI','ID','MT','NV','OR','UT','WA','WY')then state_region = 'West';
RUN;
```



Training Code

```
DATA BAN.Target_data;
SET BAN.SAS_DATA;
dti = GRAD_DEBT_MDN/MD_EARN_WNE_P6;
RUN;
```



Training Code

```
DATA BAN.Target_data;
SET BAN.SAS_DATA;
dti = ((((GRAD_DEBT_MDN*.05)/360)*30)+(GRAD_DEBT_MDN/(120)))/(MD_EARN_WNE_P6/12);
RUN;
```

DATA REDUCTION

Variable Clustering is used to address the multicollinearity assumption, whereby predictor variables should not be collinear with each other. This node helps to reduce the data as well as to remove the collinear variables. The settings used here are determining clusters by hierarchical clustering method, not confining any value to maximum clusters. Two stage clustering is set to "yes" as the dataset is large, with more than 200 variables. The Cluster split criterion is set to Second Max Eigen Value > 1. Typically, this node is for numeric variables, however, class variables can be used after changing to dummy variables. Variable clustering is performed to identify the variables which fall into similar clusters. Global clusters are formed with the following formula:

$$\text{Number of clusters} = \text{INT}\ ((\text{number of variables} / 100) + 2)$$

After the data imputation node, there are 714 variables, and substituting this in the above formula gives the number of clusters as 9.

$$\text{Number of Clusters} = \text{INT}\ ((714/100) +2) = \text{INT}\ (7.14+2) = 9$$

Variable clustering is then performed on the global clusters. The 9 global clusters comprise of 43 clusters. Variables from each cluster are selected using 1- R2 ratio. The lower the ratio, the higher the chance of selection. The below plot shows the selection of variables within each global cluster with 1- R2 criterion.

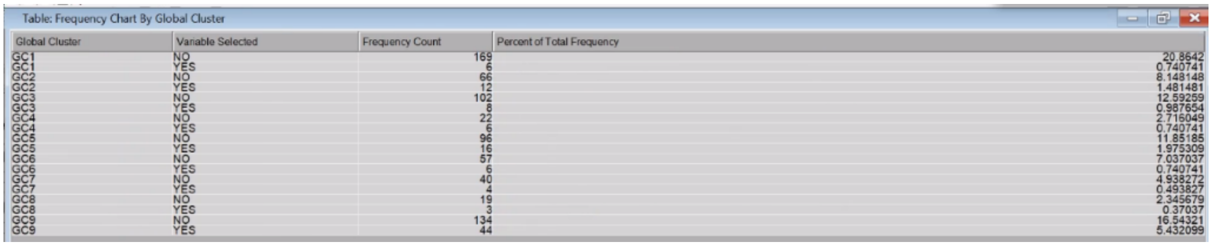


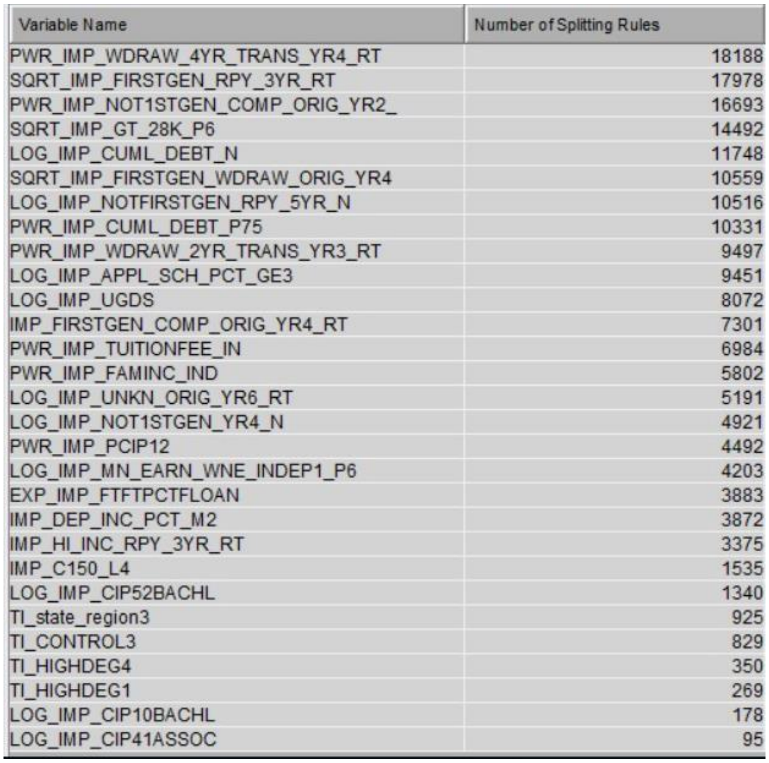
Table: Frequency Chart By Global Cluster

Global Cluster	Variable Selected	Frequency Count	Percent of Total Frequency
GC1	NO	169	20.8642
GC1	YES	6	0.740741
GC2	NO	66	8.148148
GC2	YES	12	1.481481
GC3	NO	102	12.59259
GC3	YES	8	0.987654
GC4	NO	22	2.716049
GC4	YES	6	0.740741
GC5	NO	96	11.85185
GC5	YES	16	1.975309
GC6	NO	57	7.037037
GC6	YES	6	0.740741
GC7	NO	40	4.938272
GC7	YES	4	0.493827
GC8	NO	19	2.345679
GC8	YES	3	0.37037
GC9	NO	134	16.54321
GC9	YES	44	5.432099

DATA TRANSFORMATION

A transformation is done by feeding all remaining variables into the transformation node in Miner. The transformation is set to Max Normal to get the best possible transformation. Various transformations, such as, exponential, log, power, sqrt, and square were applied to many variables. Furthermore, dummy variables are created for the binary and categorical variables such as region, main campus flag etc. to use in the modeling techniques.

APPENDIX C: MODELING



Variable Name	Number of Splitting Rules
PWR_IMP_WDRAW_4YR_TRANS_YR4_RT	18188
SQRT_IMP_FIRSTGEN_RPY_3YR_RT	17978
PWR_IMP_NOT1STGEN_COMP_ORIG_YR2_	16693
SQRT_IMP_GT_28K_P6	14492
LOG_IMP_CUML_DEBT_N	11748
SQRT_IMP_FIRSTGEN_WDRAW_ORIG_YR4	10559
LOG_IMP_NOTFIRSTGEN_RPY_5YR_N	10516
PWR_IMP_CUML_DEBT_P75	10331
PWR_IMP_WDRAW_2YR_TRANS_YR3_RT	9497
LOG_IMP_APPL_SCH_PCT_GE3	9451
LOG_IMP_UGDS	8072
IMP_FIRSTGEN_COMP_ORIG_YR4_RT	7301
PWR_IMP_TUITIONFEE_IN	6984
PWR_IMP_FAMINC_IND	5802
LOG_IMP_UNKN_ORIG_YR6_RT	5191
LOG_IMP_NOT1STGEN_YR4_N	4921
PWR_IMP_PCIP12	4492
LOG_IMP_MN_EARN_WNE_INDEP1_P6	4203
EXP_IMP_FTFTPCTFLOAN	3883
IMP_DEP_INC_PCT_M2	3872
IMP_HI_INC_RPY_3YR_RT	3375
IMP_C150_L4	1535
LOG_IMP_CIP52BACHL	1340
TI_state_region3	925
TI_CONTROL3	829
TI_HIGHDEG4	350
TI_HIGHDEG1	269
LOG_IMP_CIP10BACHL	178
LOG_IMP_CIP41ASSOC	95

SL. No	Model Name	Average Squared Error
1	Regression Stepwise	0.017
2	Forward Regression	0.017
3	Stepwise regression	0.017
4	HP Tree	0.013
5	HP Random Forest	0.0085
6	HP Neural network 2hidden layers	0.013
7	HP Neural network 5 neurons	0.0119
8	HP Neural	0.0133