

#SASGF

The logo for the Virtual SAS Global Forum 2021. The word "VIRTUAL" is written in a large, bold, white, sans-serif font. Each letter of "VIRTUAL" contains a colorful, abstract pattern of diagonal stripes in shades of blue, red, green, and purple. Below "VIRTUAL" is the text "SAS® GLOBAL FORUM 2021" in a smaller, white, sans-serif font. The entire logo is centered on a dark blue background.

**VIRTUAL**  
SAS® GLOBAL FORUM 2021

# Investing In Education Intelligently

---

## Team Path Finders

Oklahoma State University



# Meet the Path Finders



Sam Edison



Akhil Pramod Emani



Nithisha Katta



Rushya Puttam

# Outline

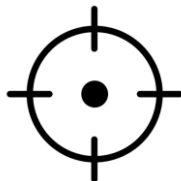
---



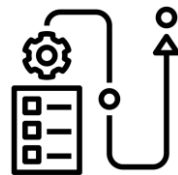
Introduction



Data Description



Scope and Goal



Methodology



Results



Insights

# Introduction

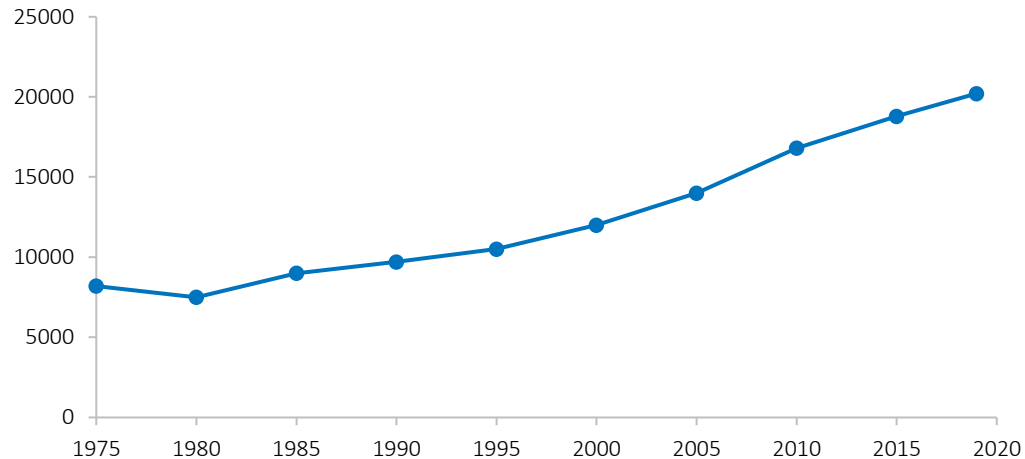
45 Million Borrowers

\$ 1.6 Trillion Debt



Help future students decide on their educational investment

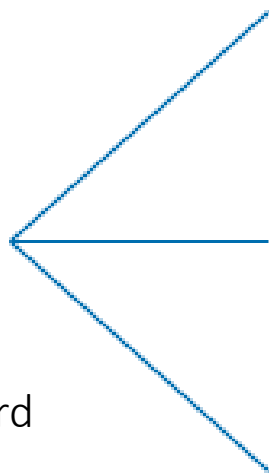
Average Education Costs 1971-2019 in 2018 Dollars



# Data Sources



College Scorecard

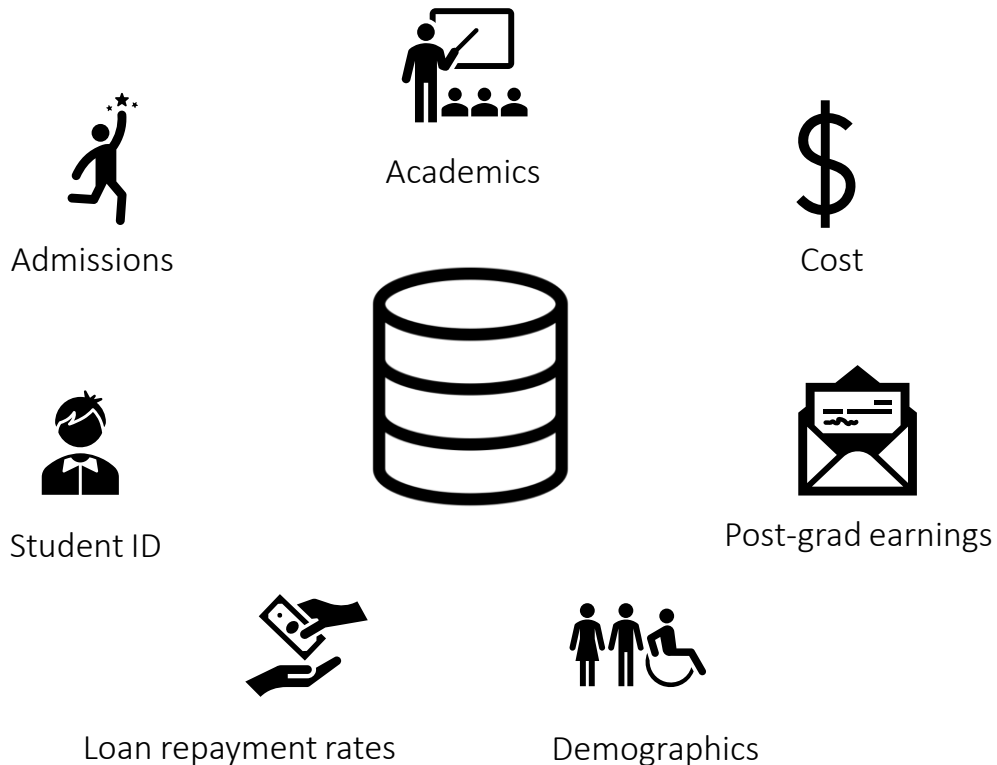


U.S Department of Education

Annual reports on Graduate, Undergraduate,  
Community colleges from 1996-2018

1800+ variables and 7000 observations per year

# Data Overview



# Scope & Goal

---



## Scope

2014-2015  
school year



## Feature Engineering

Debt-to-Income (DTI)  
Ratio

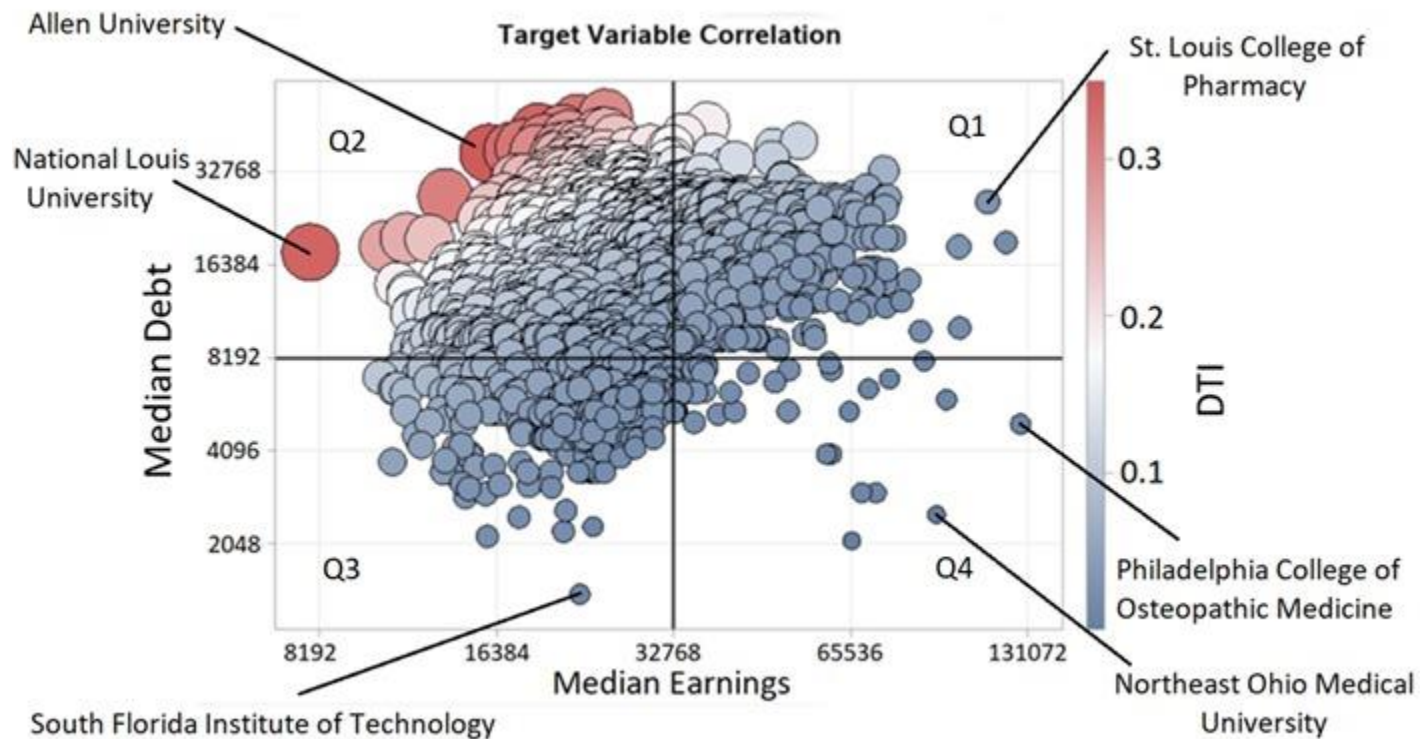


## Goal

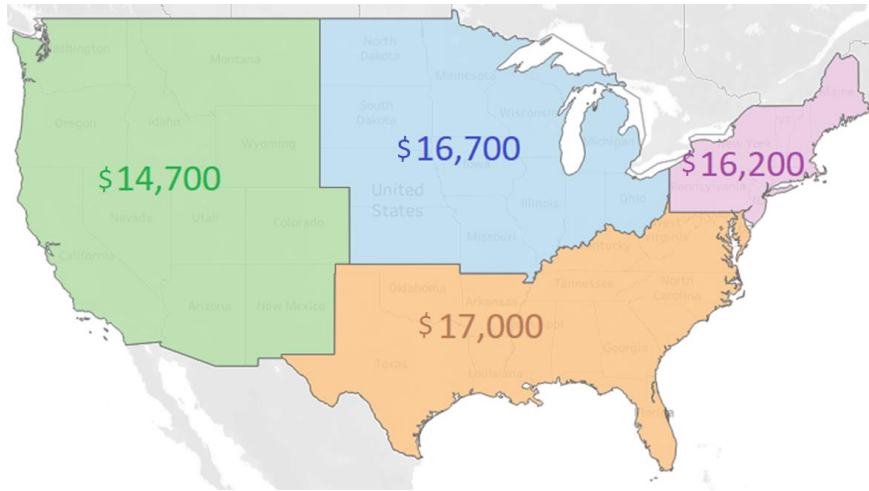
Factors impacting debt-  
to-income ratio



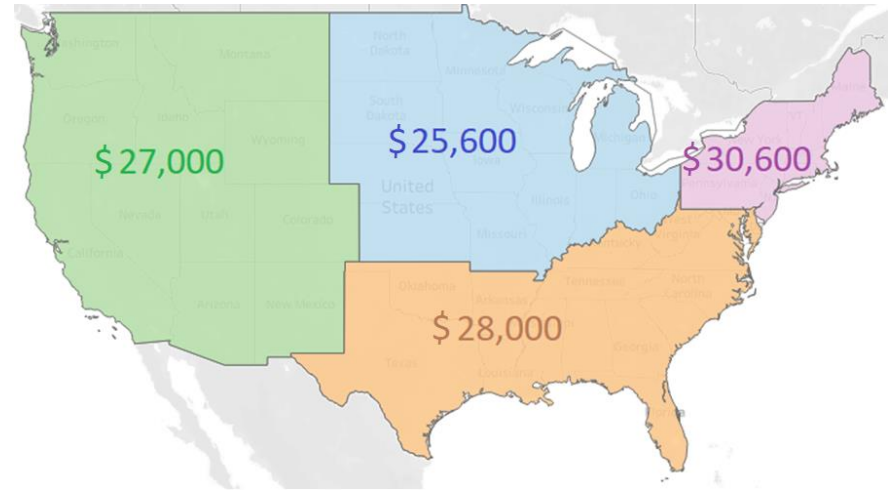
# Exploratory Data Analysis



# Exploratory Data Analysis

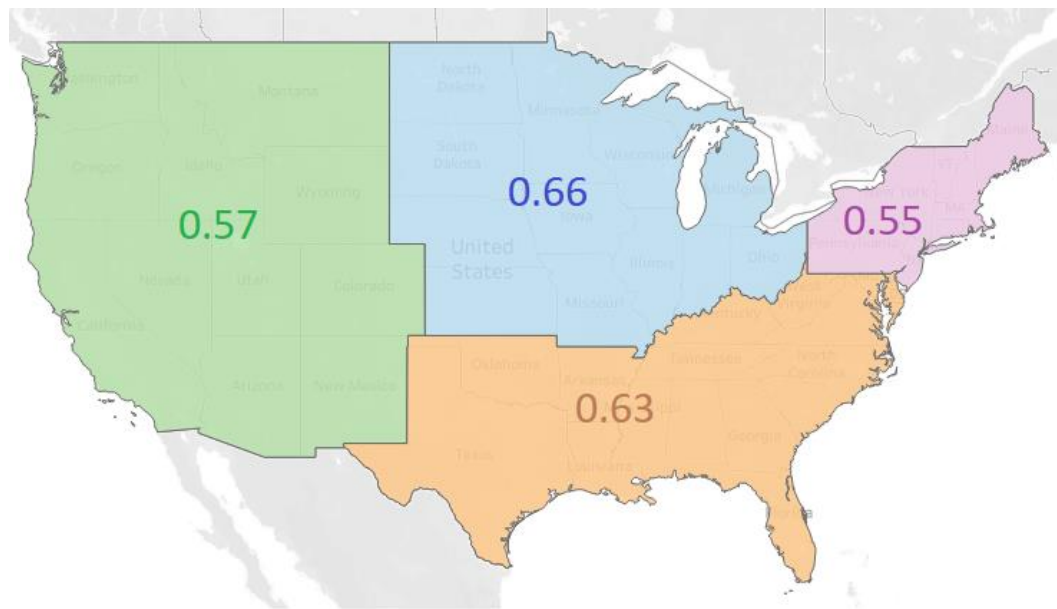


Median Debt by Region



Median Earnings by Region

# Exploratory Data Analysis



Average Debt to Income Ratio by Region

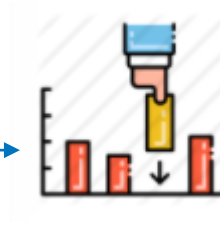
# Approach



Variable  
Creation



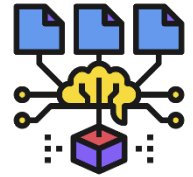
Data  
Reduction



Data  
Imputation &  
Transformation



Variable  
Selection



Model  
Building

# Data Preprocessing

Derived variables

DTI

Region

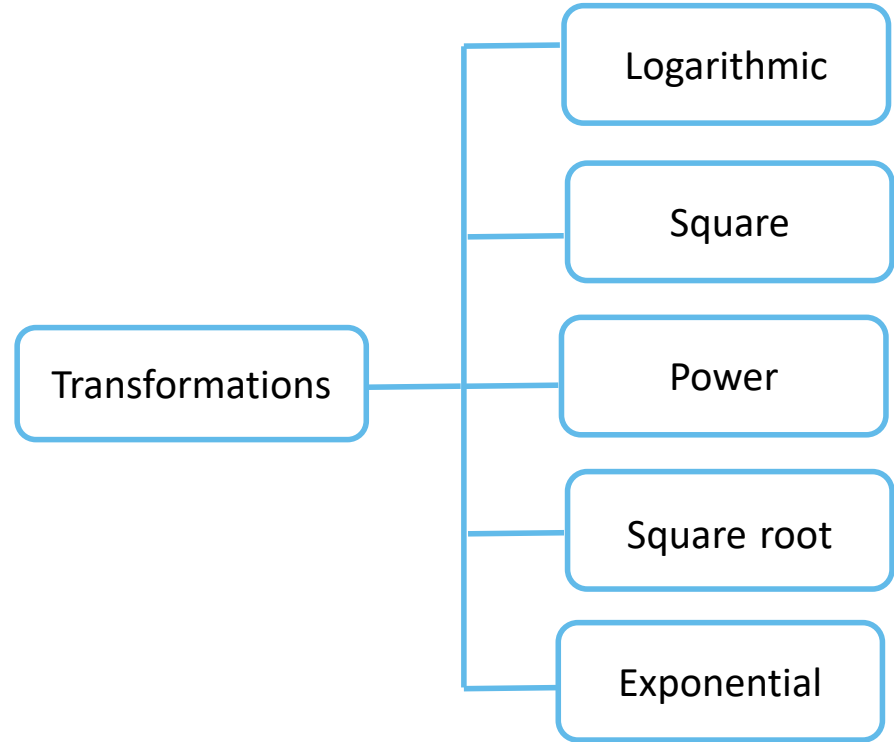
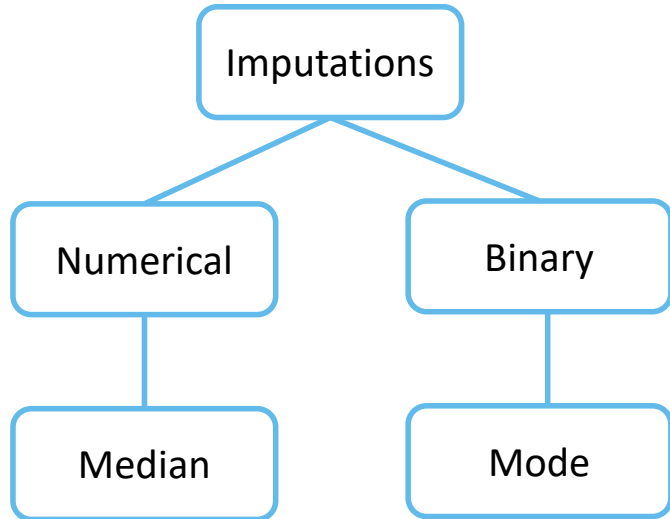
Data Reduction

Row Reduction

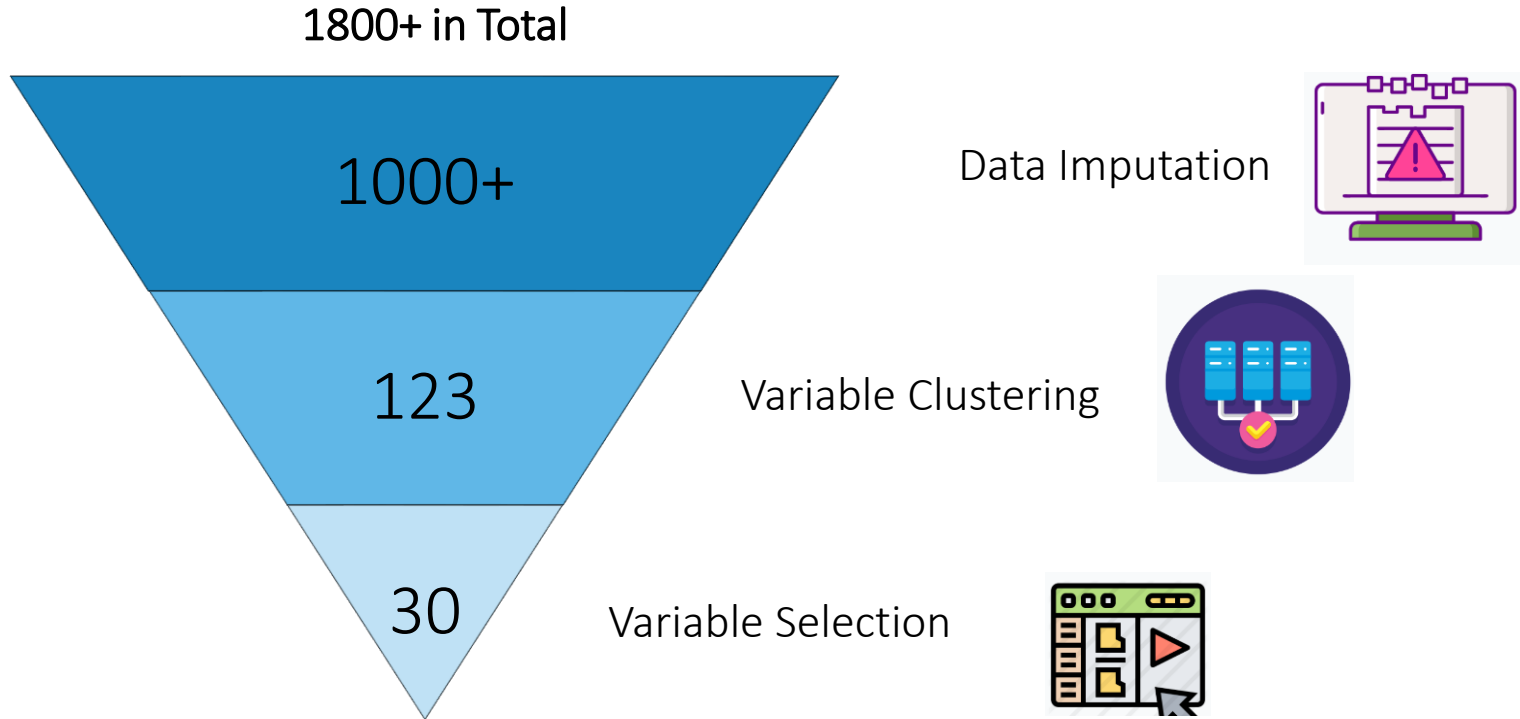
Variable  
Reduction

Missing values  
(>50%)

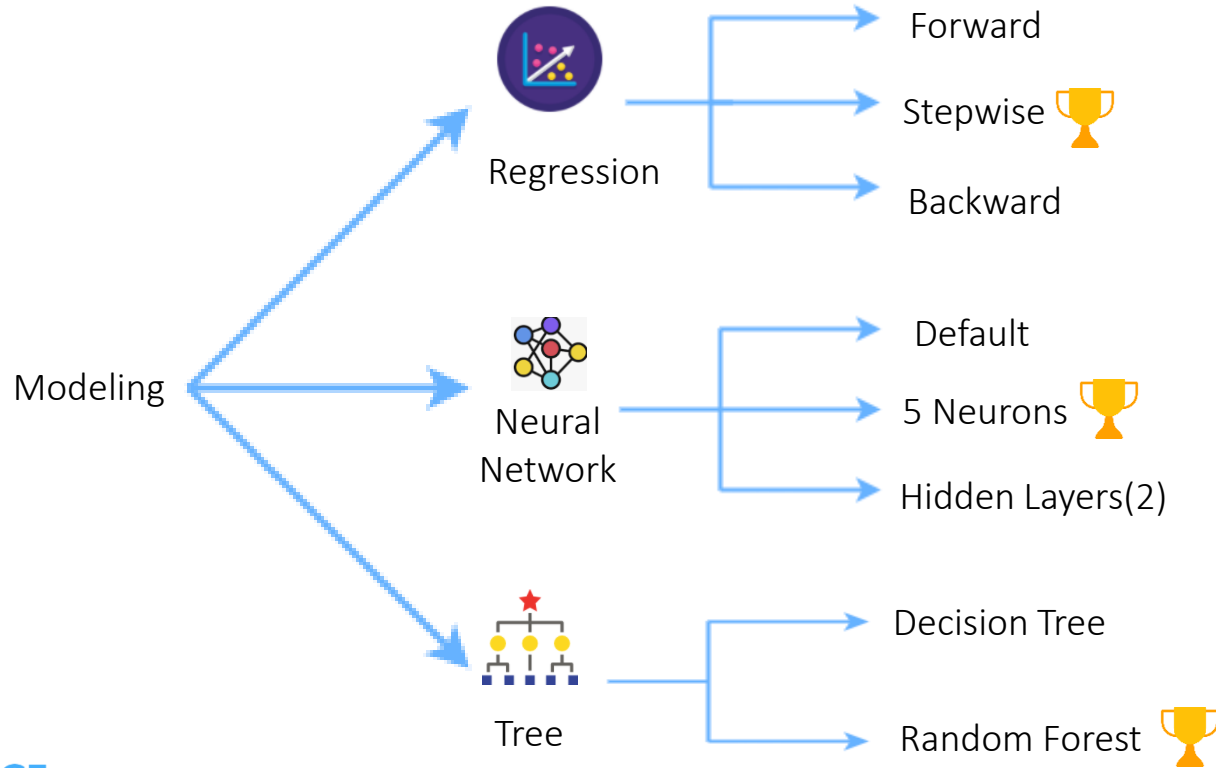
# Data Preprocessing



# Variable Selection

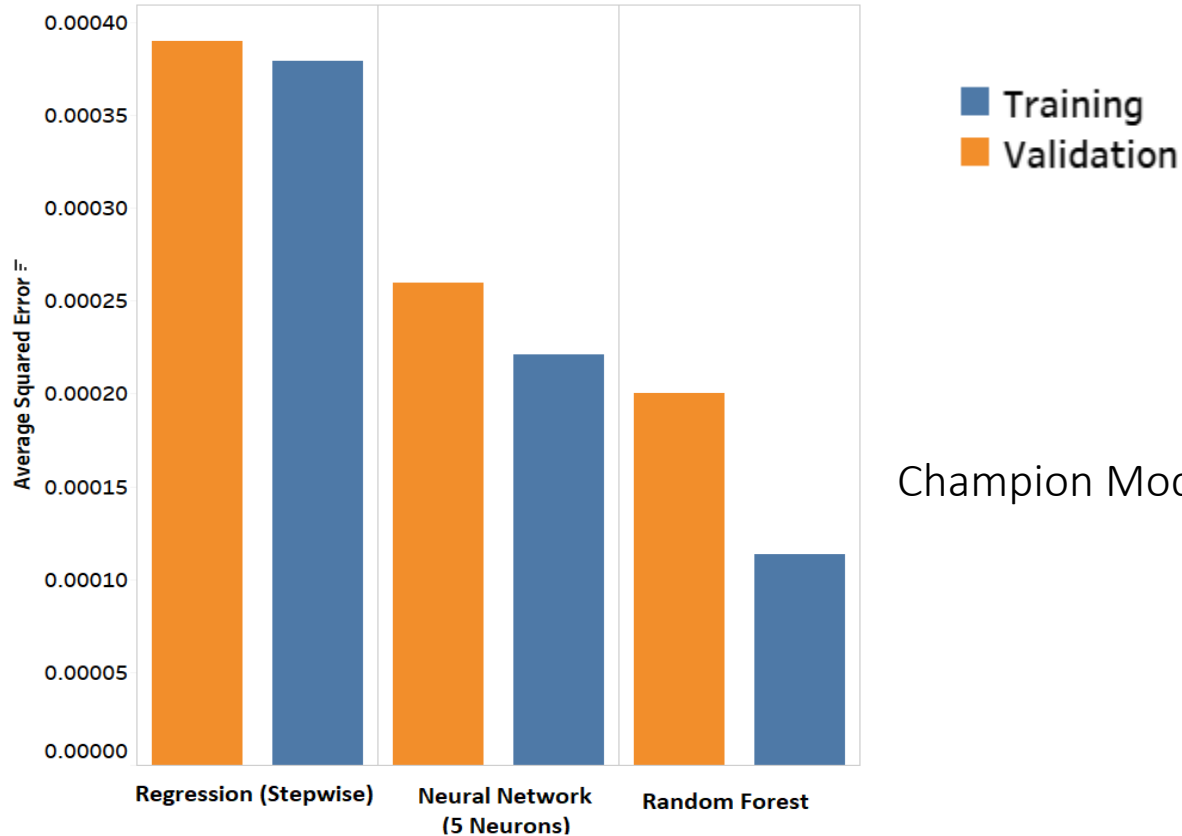


# Modeling





# Performance Metrics

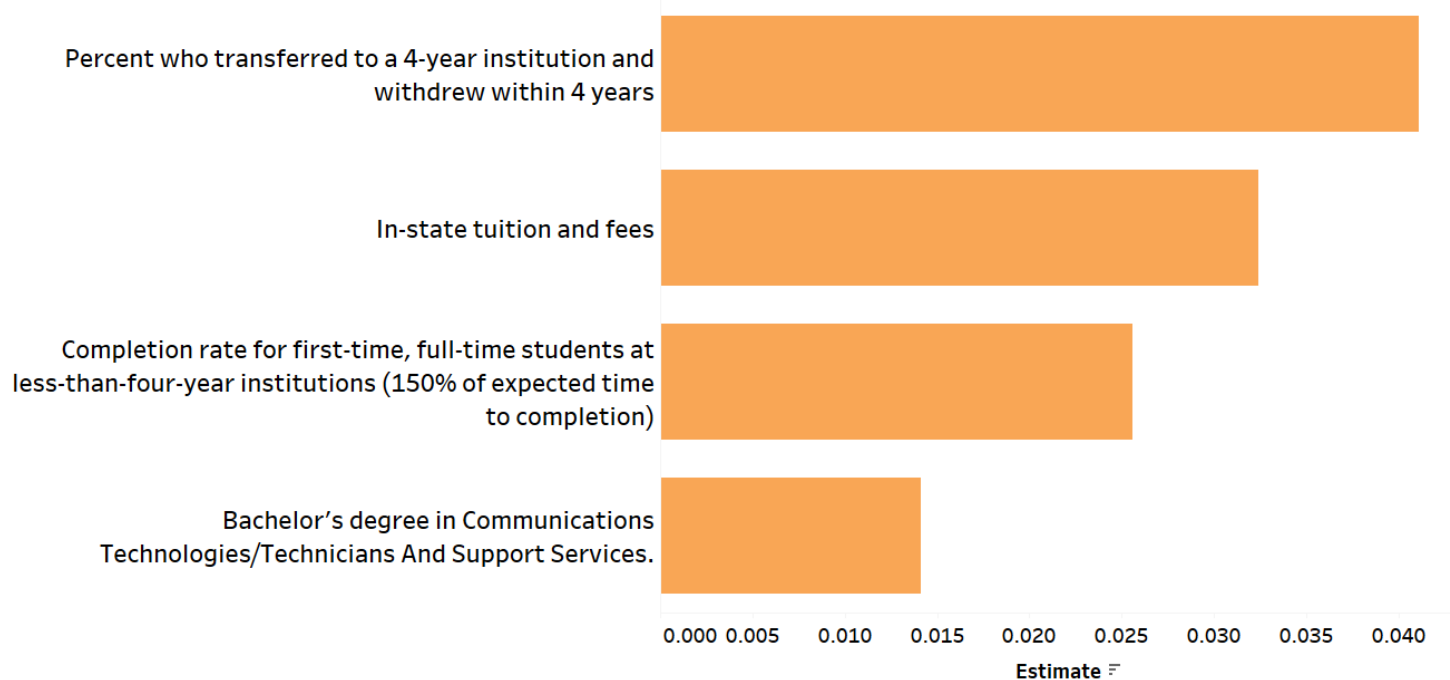


Champion Model : Random Forest



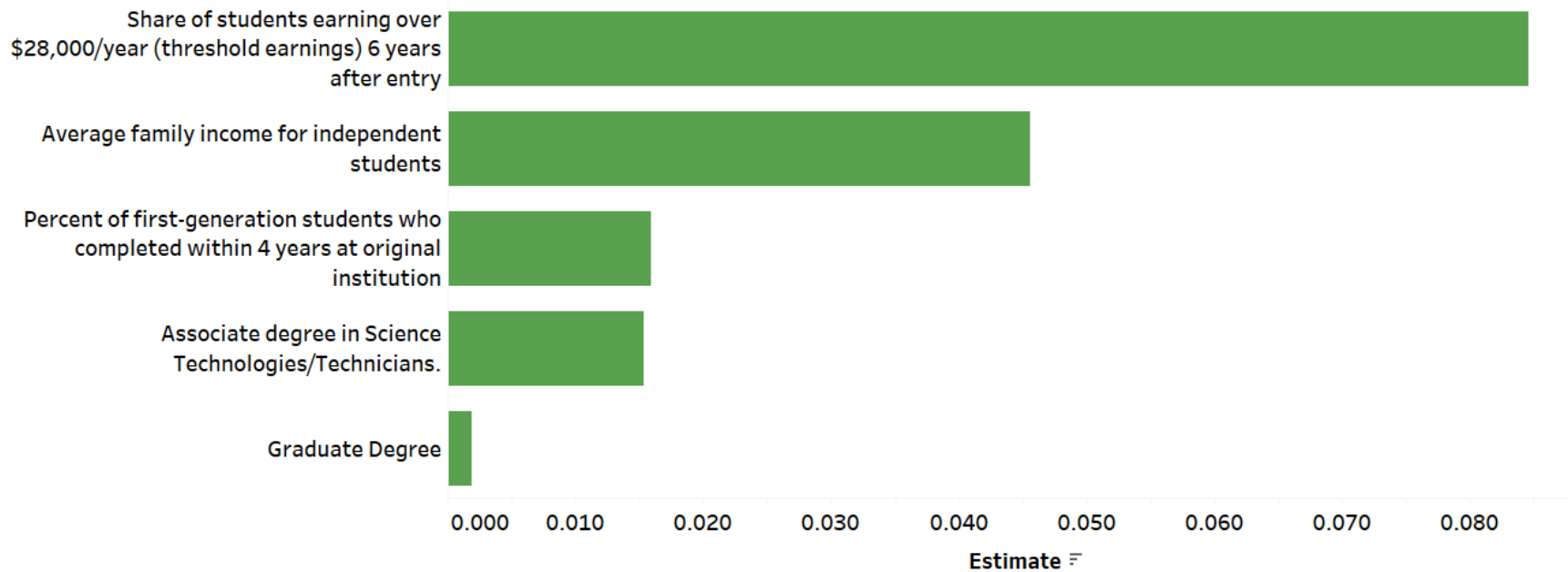
# Results

## Variables increasing Debt-To-Income ratio



# Results

## Variables decreasing Debt-To-Income ratio



# Insights



Good idea to join Main campus



Low Debt-to-Income ratio in Northeast



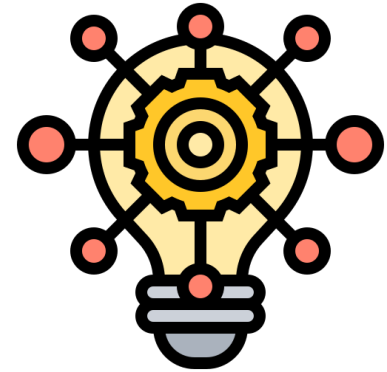
High Debt-to-Income universities in Midwest



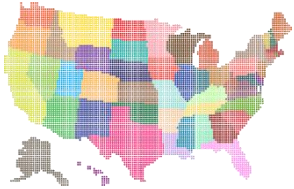
Prefer universities which have the low dropout rate



Master's and PhD degrees: low DTI



# Future Scope



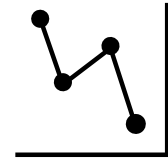
Inclusion of States



Expand to further years



Low Debt College  
Analysis



Model Biases



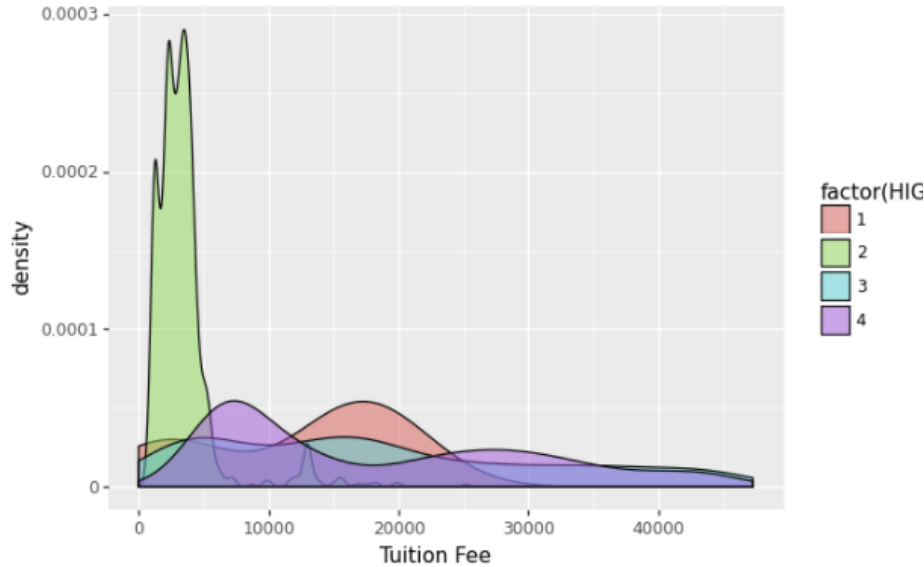
# Thank you!

Team Path Finders

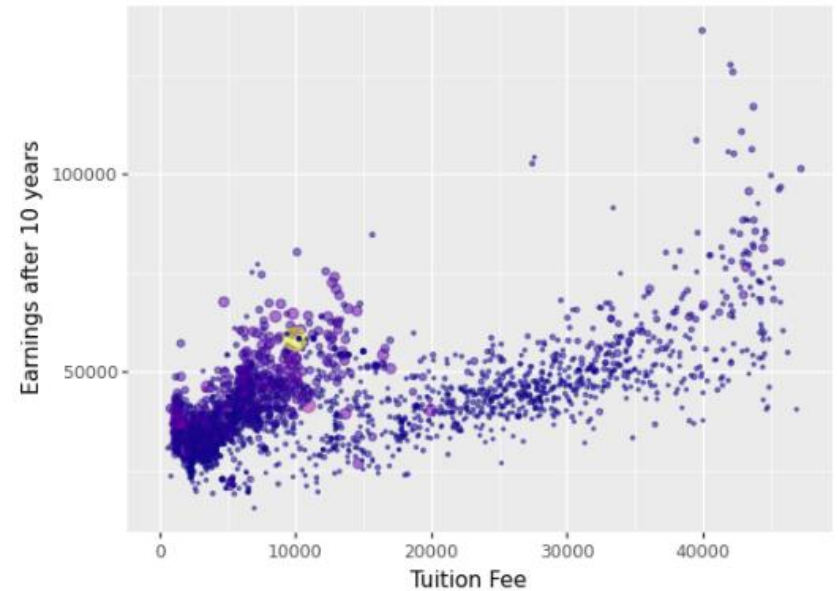
# Appendix

# EDA

Density of in-state tuition by highest degree

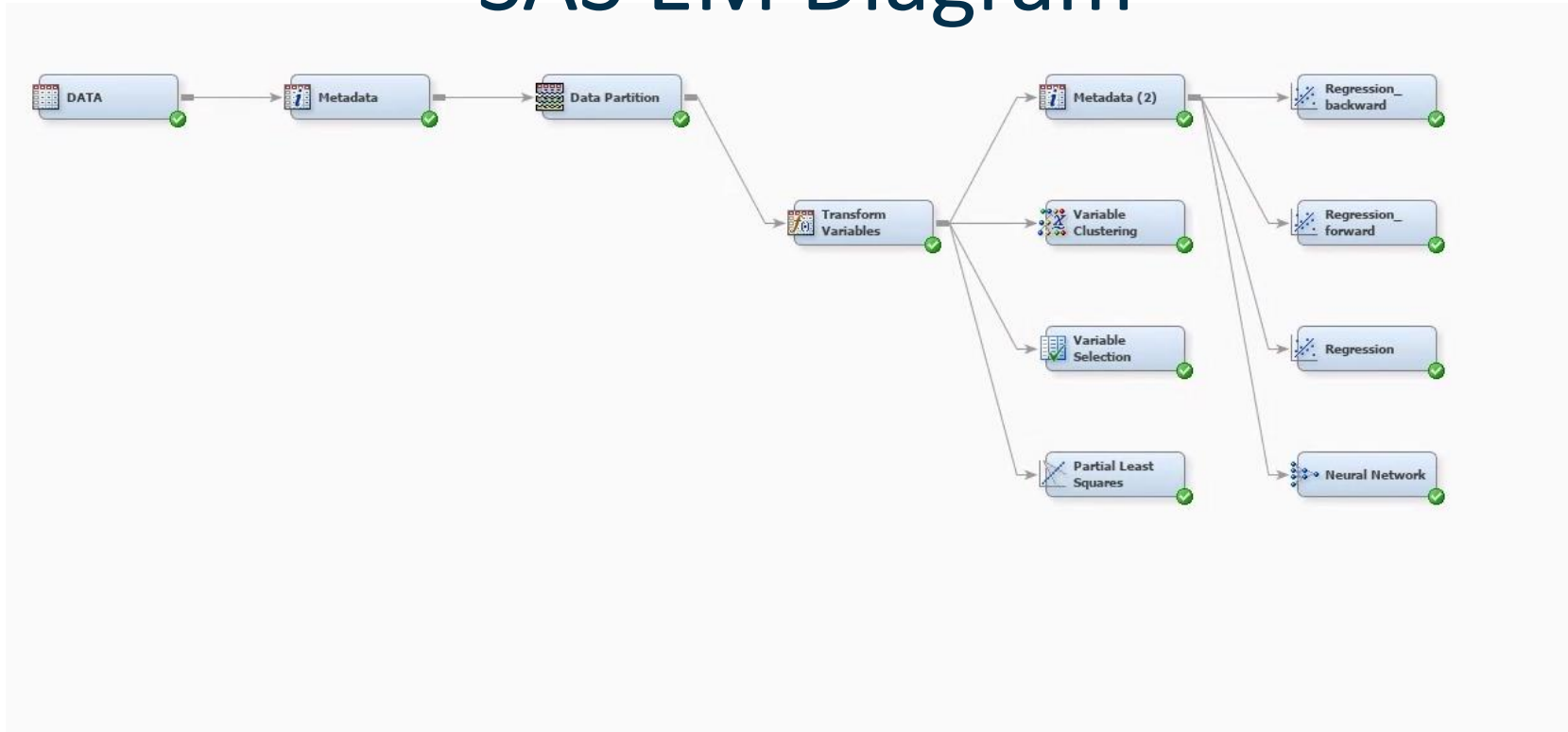


Mean earnings as a function of tuition fee





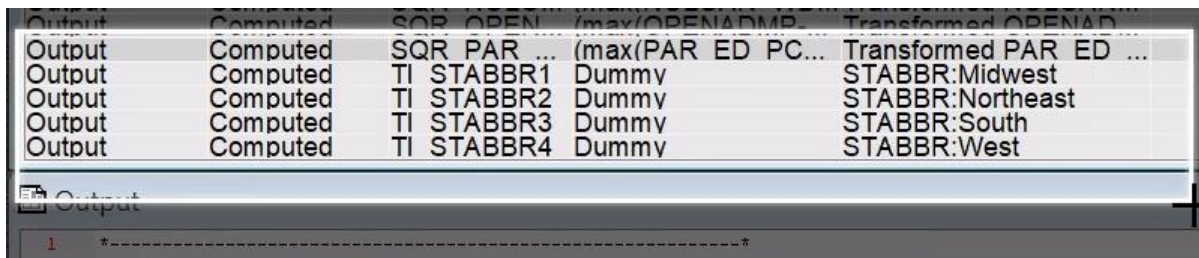
# SAS EM Diagram



# Dummies for the region Variable

The selected model is the model trained in the last step (Step 18). It consists of the following effects:

```
Intercept EXP_MAIN LOG_CIP01ASSOC LOG_CIP13BACHL LOG_CIP39ASSOC LOG_NUMBRANCH LOG_PCIP44 LOG_PCIP50 PCTPELL PWR_FIRSTGEN_DEATH_YR2_RT PWR_IND_INC_AVG PWR_LOAN_EVER PWR_PCIP52 SQR_ENRL_ORIG_YR2_RT SQR_INC_PCT_M2 SQR_NOLOAN_WDRAW_ORIG_YR3_RT SQR_WDRAW_ORIG_YR2_RT TI_STABBR1 TI_STABBR2
```



The screenshot shows a SAS Output window with a table of model effects. The table has five columns: 'Output', 'Computed', 'SQR\_OPEN', '(max(PAR ED PC...', and 'Transformed OPENAD...'. The rows list the effects: Intercept, EXP\_MAIN, LOG\_CIP01ASSOC, LOG\_CIP13BACHL, LOG\_CIP39ASSOC, LOG\_NUMBRANCH, LOG\_PCIP44, LOG\_PCIP50, PCTPELL, PWR\_FIRSTGEN\_DEATH\_YR2\_RT, PWR\_IND\_INC\_AVG, PWR\_LOAN\_EVER, PWR\_PCIP52, SQR\_ENRL\_ORIG\_YR2\_RT, SQR\_INC\_PCT\_M2, SQR\_NOLOAN\_WDRAW\_ORIG\_YR3\_RT, SQR\_WDRAW\_ORIG\_YR2\_RT, TI\_STABBR1, and TI\_STABBR2. The TI\_STABBR1-4 effects are marked as 'Dummy' and their transformed values are listed as 'STABBR:Midwest', 'STABBR:Northeast', 'STABBR:South', and 'STABBR:West' respectively.

Output	Computed	SQR_OPEN	(max(PAR ED PC...	Transformed OPENAD...
Output	Computed	SQR PAR ...	(max(PAR ED PC...	Transformed PAR ED ...
Output	Computed	TI STABBR1	Dummy	STABBR:Midwest
Output	Computed	TI STABBR2	Dummy	STABBR:Northeast
Output	Computed	TI STABBR3	Dummy	STABBR:South
Output	Computed	TI STABBR4	Dummy	STABBR:West

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	18	181.626916	10.090384	264.48	<.0001
Error	5286	201.671836	0.038152		
Corrected Total	5304	383.298752			

Model Fit Statistics

R-Square	0.4739	Adj R-Sq	0.4721
AIC	-17308.0940	BIC	-17305.9853
SBC	-17183.1423	C(p)	22.8579

Type 3 Analysis of Effects

Effect	DF	Sum of Squares	F Value	Pr > F
EXP_MAIN	1	5.5773	146.19	<.0001
LOG_CIP01ASSOC	1	1.7340	45.45	<.0001
LOG_CIP13BACHL	1	4.8982	128.39	<.0001
LOG_CIP39ASSOC	1	1.9526	51.18	<.0001
LOG_NUMBRANCH	1	1.8570	48.67	<.0001
LOG_PCIP44	1	0.9941	26.06	<.0001
LOG_PCIP50	1	11.1396	291.98	<.0001
PCTPELL	1	7.5782	198.63	<.0001
PWR_FIRSTGEN_DEATH_YR2_RT	1	0.6531	17.12	<.0001
PWR_IND_INC_AVG	1	11.4913	301.20	<.0001
PWR_LOAN_EVER	1	39.8573	1044.70	<.0001
PWR_PCIP52	1	7.2507	190.05	<.0001
SQRT_ENRL_ORIG_YR2_RT	1	6.5225	170.96	<.0001
SQRT_INC_PCT_M2	1	0.2753	7.22	0.0072
SQRT_NOLOAN_WDRAW_ORIG_YR3_RT	1	0.2820	7.39	0.0066
SQRT_WDRAW_ORIG_YR2_RT	1	21.9883	576.33	<.0001
TI_STABBR1	1	0.8317	21.80	<.0001
TI_STABBR2	1	0.8239	21.59	<.0001

# Surrogate Modeling results

# Surrogate Modeling results

## Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	t Value	Pr >  t	
Intercept	1	0.1296	0.0540	2.40	0.0164	
EXP_MAIN	1	-0.0550	0.00455	-12.09	<.0001	
LOG_CIP01ASSOC	1	-0.1990	0.0295	-6.74	<.0001	
LOG_CIP13BACHL	1	0.2443	0.0216	11.33	<.0001	
LOG_CIP39ASSOC	1	0.2678	0.0374	7.15	<.0001	
LOG_NUMBRANCH	1	0.1988	0.0285	6.98	<.0001	
LOG_PCIP44	1	0.6064	0.1188	5.10	<.0001	
LOG_PCIP50	1	0.5338	0.0312	17.09	<.0001	
PCTPELL	1	0.2172	0.0154	14.09	<.0001	
PWR_FIRSTGEN_DEATH_YR2_RT	1	0.2964	0.0716	4.14	<.0001	
PWR_IND_INC_AVG	1	-0.7539	0.0434	-17.36	<.0001	
PWR_LOAN_EVER	1	0.4740	0.0147	32.32	<.0001	
PWR_PCIP52	1	0.1528	0.0111	13.79	<.0001	
SQRT_ENRL_ORIG_YR2_RT	1	0.3040	0.0232	13.08	<.0001	
SQRT_INC_PCT_M2	1	-0.1237	0.0461	-2.69	0.0072	
SQRT_NOLOAN_WDRAW_ORIG_YR3_RT	1	0.0845	0.0311	2.72	0.0066	
SQRT_WDRAW_ORIG_YR2_RT	1	0.6119	0.0255	24.01	<.0001	
TI_STABBR1	0	1	-0.0155	0.00331	-4.67	<.0001
TI_STABBR2	0	1	0.0164	0.00353	4.65	<.0001

	A	B
1	Variable	Estimates
2	SQRT_WDRAW_ORIG_YR2_RT	0.6119
3	LOG_PCIP44	0.6064
4	LOG_PCIP50	0.5338
5	PWR_LOAN_EVER	0.474
6	SQRT_ENRL_ORIG_YR2_RT	0.304
7	PWR_FIRSTGEN_DEATH_YR2_RT	0.2964
8	LOG_CIP39ASSOC	0.2678
9	LOG_CIP13BACHL	0.2443
10	PCTPELL	0.2172
11	LOG_NUMBRANCH	0.1988
12	PWR_PCIP52	0.1528
13	Intercept	0.1296
14	SQRT_NOLOAN_WDRAW_ORIG_YR3_RT	0.0845
15	TI_STABBR2(Northeast)	0.0164
16	TI_STABBR1(Midwest)	-0.0155
17	EXP_MAIN	-0.055
18	SQRT_INC_PCT_M2	-0.1237
19	LOG_CIP01ASSOC	-0.199
20	PWR_IND_INC_AVG	-0.7539
21		

# Final Variables selected in the Model