Data for Diplomas
sponsored by AT&T

Insights to 90% High School Graduation Rate Using Statistical Modelling

•••

November 15, 2015
By: Izunna Okonkwo
MIT Undergraduate

Project Overview

- Modelling Methodology
- Model Analysis
- Actionable Insights
- Comparison to Benchmark Results
- Data Visualization
- Conclusion and Future Works

Methodology Overview

Selecting the variables

Data Cleaning

Boruta Algorithm to select data

Splitting the data

Split data into Training and Testing Sets

Creating the Model

Logistic Regression Model

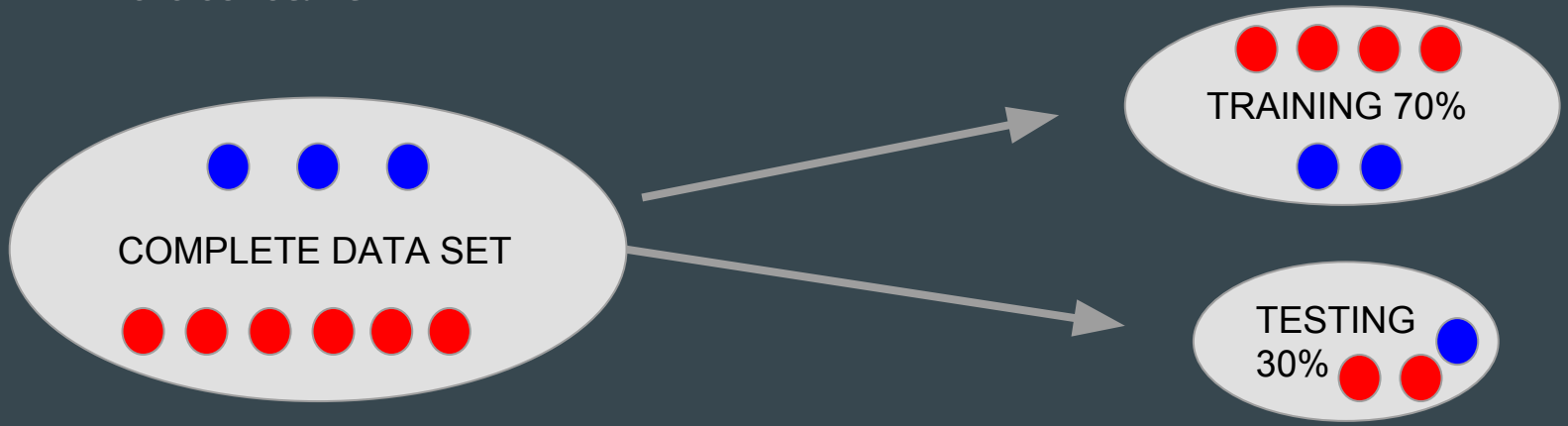
CART Model

Selecting the Variables for The Model

- Used the merged dataset which contained the rows as different schools
- Used the Local Education Agency ID (LEAID) as the Unique identifier
- Selected variables with at least 70% complete data set. (Removed NA's)
- Choose specific Algorithms using a machine learning algorithm
    - Implemented the Boruta Algorithm in R which aims to find the most statistically relevant variables for a random forest Model
    - Arrived at 94 practical variables of which I cut down to 71 that seemed reasonable
- Created a new Variable based on a graduation rate greater than or equal to 90% to be the independent variable for my Analysis
- Removed State as a pertinent Variable to remove location Bias

Creating the Training and Testing Set

- Using the caTools package in R, randomly selected 70% of Observations for Training set and 30% for Testing Set
- Proportion of Independent Variable (90% graduation rate) in Training and Testing Set were constant.



Creating and Testing Models: Logistic Regression

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.540e+09	9.846e+08	-1.564	0.117868	
ALL_COHORT_1112	3.780e-03	4.450e-04	8.494	< 2e-16	***
MWH_COHORT_1112	-1.033e-03	4.400e-04	-2.347	0.018910	*
CWD_COHORT_1112	-1.960e-02	2.502e-03	-7.834	4.74e-15	***
ECD_COHORT_1112	-6.430e-03	6.208e-04	-10.356	< 2e-16	***
LAND_AREA	-9.655e-04	1.745e-04	-5.534	3.14e-08	***
RURAL_POP_CEN_2010	6.816e-05	2.875e-05	2.371	0.017746	*
Hispanic_CEN_2010	4.411e-04	1.821e-04	2.421	0.015461	*
NH_White_alone_CEN_2010	4.918e-04	1.779e-04	2.765	0.005697	**
NH_White_alone_ACS_08_12	-9.077e-05	1.827e-04	-0.497	0.619318	
NH_Blk_alone_CEN_2010	5.048e-04	7.178e-04	0.703	0.481918	
NH_Blk_alone_ACS_08_12	-9.614e-05	7.615e-04	-0.126	0.899531	
NH_Blk_alone_ACSMOE_08_12	1.172e-03	1.361e-03	0.861	0.389169	
NH_AIAN_alone_CEN_2010	-1.933e-03	1.751e-03	-1.104	0.269793	
Not_HS_Grad_ACS_08_12	-1.018e-03	5.199e-04	-1.958	0.050275	.
College_ACS_08_12	-1.388e-04	3.726e-04	-0.373	0.709502	
College_ACSMOE_08_12	-6.080e-04	2.314e-03	-0.263	0.792769	
Prs_Blw_Pov_Lev_ACS_08_12	7.118e-04	4.600e-04	1.548	0.121740	
Prs_Blw_Pov_Lev_ACSMOE_08_12	-1.481e-03	1.312e-03	-1.128	0.259138	
US_Cit_Nat_ACS_08_12	6.097e-04	3.974e-04	1.534	0.125000	
MrdCple_Fmly_HHD_ACS_08_12	-4.184e-04	4.917e-04	-0.851	0.394746	
Female_No_HB_CEN_2010	-5.514e-03	1.634e-03	-3.375	0.000739	***

- Summary of Logistic Regression Model Insights (Full model Implementation can be found in the Appendix)
- Predictions with this model on the test set with a threshold of 0.5 obtains an accuracy of 70.4%
- The number of stars beside the variable signifies its importance in the logistic regression

Logistic Regression Actionable Insights - Strongly Correlated

Household Initiatives

Households with female householder, no husband present

Housing units without complete plumbing in ACS

The percentage of all ACS housing units that are in a structure that contains only that single unit

Classroom Initiatives

Total Size of Cohort

Socio-economic Initiatives

Prediction of low census mail return rate

The percentage of the 2010 Census total population that is between 5 and 17 years old

Number of economically disadvantaged students

Number of children with disabilities

Land Area of School District

Logistic Regression Actionable Insights - Medium correlation

Household Initiatives

Housing units without complete plumbing facilities in the ACS

% of ACS population aged 5 years > at the time of the interview that speak only English at home

% of all ACS housing units that are considered mobile homes

Classroom Initiatives

of white students in graduation cohort

Persons of Hispanic Origin

Non-Hispanic White only in the 2010 Census

The percentage of ACS civilians ages 16 years and over in the labor force that are employed

Socio-economic Initiatives

Population living in Area outside of an Urban Area in the 2010 Census

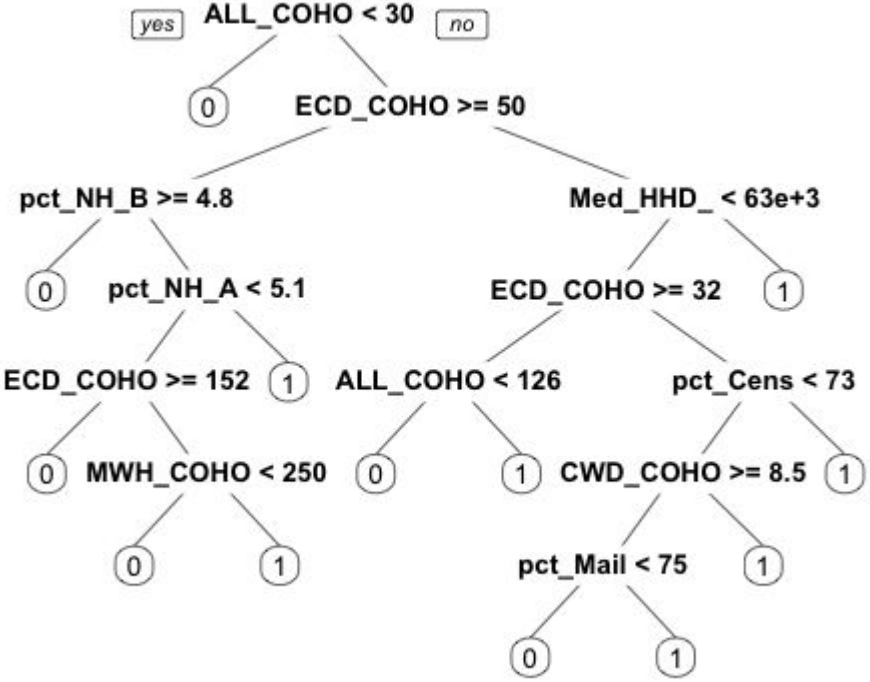
% of the 2010 Census total population that indicate no Hispanic origin and their only race as "Asian"

% of the ACS eligible population classified below the poverty level given their total family or household income within the last year, family size, and family composition

Actionable Insights - Logistics Regression

Variable	Rationale
Household Initiatives parents can lead the change	• negative correlation between single moms household • positive correlation w/ houses without complete plumbing and single unit structures
Classroom Initiatives schools can lead the change	• strong positive correlation with total size of graduation cohort • moderate correlation with %of ACS civilians ages 16 years and over in the labor force that are employed
Socio - Economic Initiatives state governments can lead the change	• negative correlation with students from economically disadvantaged backgrounds and Land Area of School District • positive correlation with Population living in Area outside of an Urban Area in the 2010 Census

Creating and Testing Models: Cart Model



- Summary of Cart Model Insights
- Predictions with this model on the test set obtains an accuracy of 74.4%
- 0 Signifies that the graduation rate is lower than 90% while 1 signifies that the Graduation rate is higher than or equal to 90%

CART Model Actionable Insights

Variable	Rationale
Size of the Cohort	• High graduation rates would be difficult to achieve if the cohor is small
Economically Disadvantaged Students (ECD)	• If the cohort is medium sized (between 30 and 125) and the number of ECD is medium (between 32 and 50) then they have low Grad Rates • If the percent of ECD in cohort is high, graduation rates are low
Percentage of African Americans (BLK)	• If the cohort is not small, the ECD is greater than 50 and the percentage of African Americans in the district is greater than 4.8%, the graduation rate is not above 90%
Percentage of American Indian or Alaska Native (AIAN)	• If the percentage of BLK in the district is smaller than 4.8% and the percentage of AIAN is greater than 5.1% then they have above 90% graduation rates.

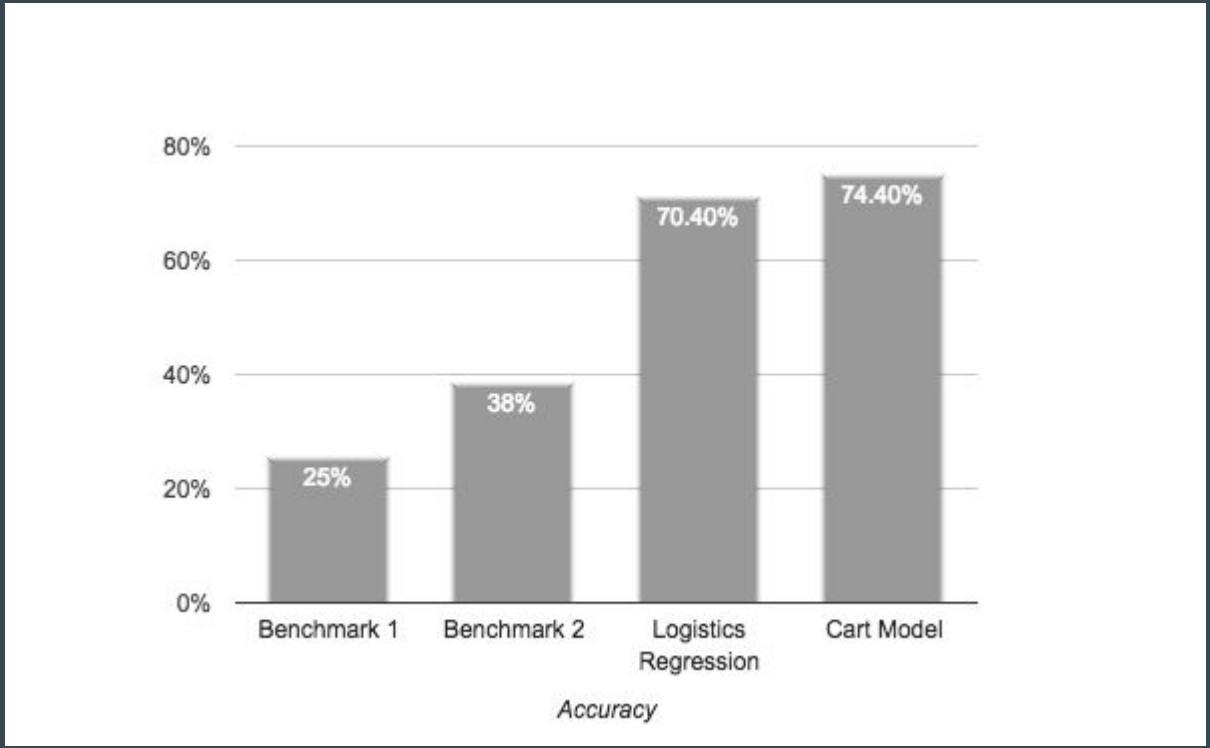
*High Graduation Rate defined as rate >=90%

CART Model Actionable Insights

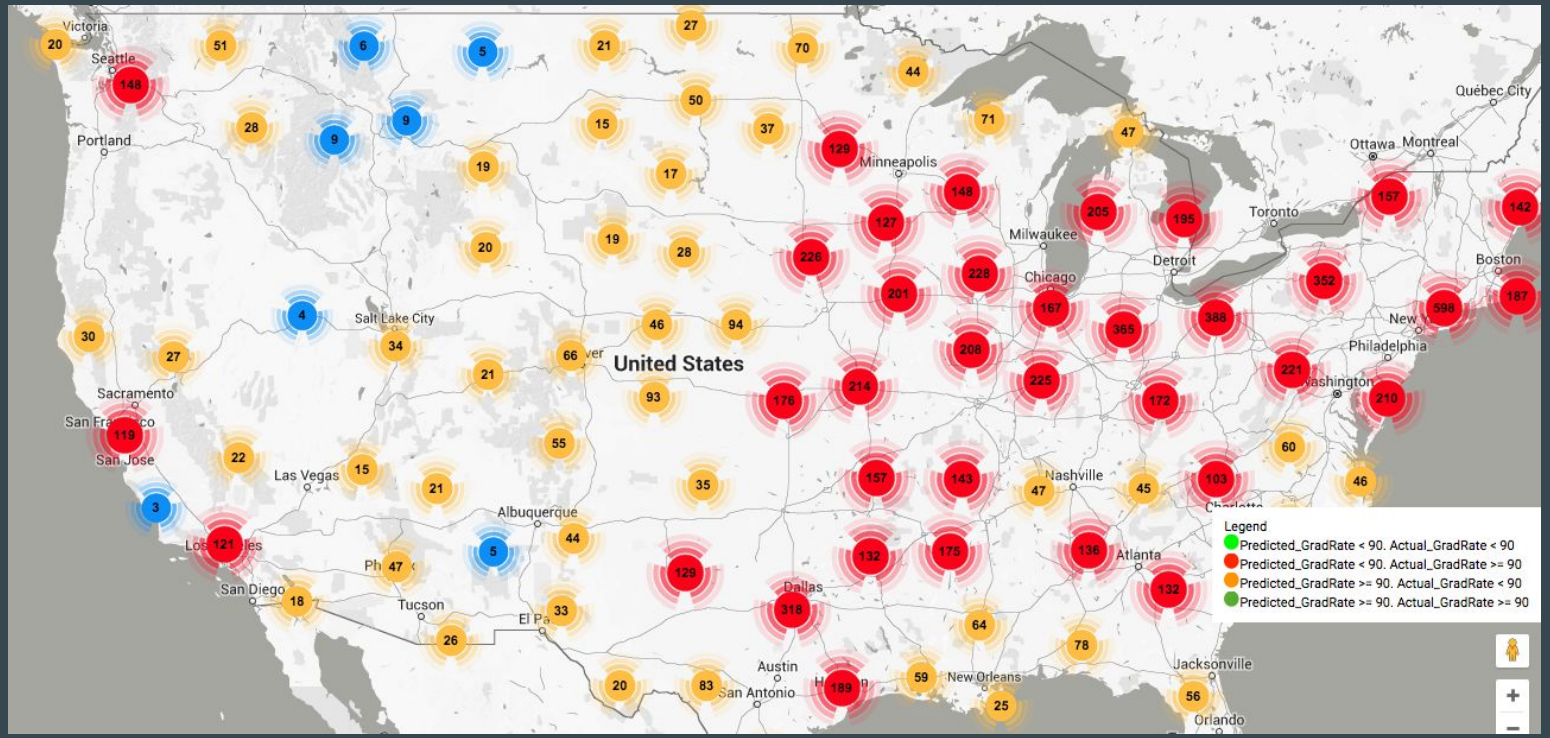
Variable	Rationale
Median Household Income (MHHD)	• If the District Median Household Income is greater than $63000 and ECD number is low, then there is going to be a high graduation rate
Number of Students with Disabilities	• If the MHHD is smaller than $63000 and the CWD is less than 8, then the School will have a high graduation rate
Percent mailback Count	• If people are moving from their addresses during the year is high, then it is likely to have smaller graduation rates

*High Graduation Rate defined as rate >=90%

Comparison with Benchmark



Data Visualization trends



http://web.mit.edu/izunna/www/ATT/map.html

Future Work and Improvements

- Creating a Random Forest Model for Predictions
- Implementing the AdaBoost Machine Learning Algorithm to combine all Models Predictive Capability
- Look into other machine learning algorithms give us higher accuracy
- incorporate more pertinent variables e.g urban planning and after school programs

By: Izunna Okonkwo
Civil Engineering and Management at MIT
Class of 2016

Appendix A: R Logistic Regression Model Implementation

```
glm(formula = completeGrad ~ ALL_COHORT_1112 + MWH_COHORT_1112 +
    CWD_COHORT_1112 + ECD_COHORT_1112 + LAND_AREA + RURAL_POP_CEN_2010 +
    Hispanic_CEN_2010 + NH_White_alone_CEN_2010 + NH_White_alone_ACS_08_12 +
    NH_Blk_alone_CEN_2010 + NH_Blk_alone_ACS_08_12 + NH_Blk_alone_ACSMOE_08_12 +
    NH_AIAN_alone_CEN_2010 + Not_HS_Grad_ACS_08_12 + College_ACS_08_12 +
    College_ACSMOE_08_12 + Prs_Blw_Pov_Lev_ACS_08_12 + Prs_Blw_Pov_Lev_ACSMOE_08_12 +
    US_Cit_Nat_ACS_08_12 + MrdCple_Fmly_HHD_ACS_08_12 + Female_No_HB_CEN_2010 +
    Med_HHD_Inc_ACS_08_12 + Tot_Vacant_Units_CEN_2010 + Tot_Vacant_Units_ACS_08_12 +
    Mobile_Homes_ACS_08_12 + No_Plumb_ACS_08_12 + FRST_FRMS_CEN_2010 +
    Mail_Return_Rate_CEN_2010 + Low_Response_Score + pct_Pop_5_17_CEN_2010 +
    pct_Pop_18_24_CEN_2010 + pct_Hispanic_ACS_08_12 + pct_NH_White_alone_CEN_2010 +
    pct_NH_White_alone_ACS_08_12 + pct_NH_Blk_alone_CEN_2010 +
    pct_NH_Blk_alone_ACS_08_12 + pct_NH_Blk_alone_ACSMOE_08_12 +
    pct_NH_AIAN_alone_CEN_2010 + pct_NH_Asian_alone_CEN_2010 +
    pct_Pop_5yrs_Over_ACSMOE_08_12 + pct_Age5p_Only_Eng_ACS_08_12 +
    pct_Age5p_Scandinav_ACSMOE_08_12 + pct_Pop_25yrs_Over_ACSMOE_08_12 +
    pct_Not_HS_Grad_ACS_08_12 + pct_Not_HS_Grad_ACSMOE_08_12 +
    pct_College_ACS_08_12 + pct_College_ACSMOE_08_12 + pct_Prs_Blw_Pov_Lev_ACS_08_12 +
    pct_Prs_Blw_Pov_Lev_ACSMOE_08_12 + pct_Civ_emp_16p_ACS_08_12 +
    pct_Civ_unemp_16p_ACSMOE_08_12 + pct_Pop_1yr_Over_ACSMOE_08_12 +
    pct_Diff_HU_1yr_Ago_ACSMOE_08_12 + pct_Born_US_ACS_08_12 +
    pct_MrdCple_HHD_ACS_08_12 + pct_MrdCple_HHD_ACSMOE_08_12 +
    pct_Not_MrdCple_HHD_ACS_08_12 + pct_Female_No_HB_CEN_2010 +
    pct_Female_No_HB_ACS_08_12 + pct_HHD_PPL_Und_18_CEN_2010 +
    pct_Tot_Occp_Units_CEN_2010 + pct_Vacant_Units_CEN_2010 +
    pct_Vacant_Units_ACSMOE_08_12 + pct_Single_Unit_ACS_08_12 +
    pct_MLT_U2_9_STRC_ACSMOE_08_12 + pct_Mobile_Homes_ACS_08_12 +
    pct_Mobile_Homes_ACSMOE_08_12 + pct_No_Plumb_ACS_08_12 +
    pct_Census_Mail_Returns_CEN_2010 + pct_Mailback_Count_CEN_2010 +
    pct_FRST_FRMS_CEN_2010, family = binomial, data = gDataTrain)
```

Appendix B-1: Summary of R Logistic Model

Variable	Estimate	Std. Error	z value	Pr(>|z|)	Importance
ALL_COHORT_1112	3.78E-03	4.45E-04	8.494	< 2e-16	***
MWH_COHORT_1112	-1.03E-03	4.40E-04	-2.347	0.01891	*
CWD_COHORT_1112	-1.96E-02	2.50E-03	-7.834	4.74E-15	***
ECD_COHORT_1112	-6.43E-03	6.21E-04	-10.356	< 2e-16	***
LAND_AREA	-9.66E-04	1.75E-04	-5.534	3.14E-08	***
RURAL_POP_CEN_2010	6.82E-05	2.88E-05	2.371	0.017746	*
Hispanic_CEN_2010	4.41E-04	1.82E-04	2.421	0.015461	*
NH_White_alone_CEN_2010	4.92E-04	1.78E-04	2.765	0.005697	**
NH_White_alone_ACS_08_12	-9.08E-05	1.83E-04	-0.497	0.619318	

Appendix B-2: Summary of R Logistic Model

Variable	Estimate	Std. Error	z value	Pr(>|z|)	Importance
NH_Blk_alone_CEN_2010	5.05E-04	7.18E-04	0.703	0.481918	
NH_Blk_alone_ACS_08_12	-9.61E-05	7.62E-04	-0.126	0.899531	
NH_Blk_alone_ACSMOE_08_12	1.17E-03	1.36E-03	0.861	0.389169	
NH_AIAN_alone_CEN_2010	-1.93E-03	1.75E-03	-1.104	0.269793	
Not_HS_Grad_ACS_08_12	-1.02E-03	5.20E-04	-1.958	0.050275	.
College_ACS_08_12	-1.39E-04	3.73E-04	-0.373	0.709502	
College_ACSMOE_08_12	-6.08E-04	2.31E-03	-0.263	0.792769	
Prs_Blw_Pov_Lev_ACS_08_12	7.12E-04	4.60E-04	1.548	0.12174	
Prs_Blw_Pov_Lev_ACSMOE_08_12	-1.48E-03	1.31E-03	-1.128	0.259138	

Appendix B-3: Summary of R Logistic Model

Variable	Estimate	Std. Error	z value	Pr(>|z|)	Importance
US_Cit_Nat_ACS_08_12	6.10E-04	3.97E-04	1.534	0.125	
MrdCple_Fmly_HHD_ACS_08_12	-4.18E-04	4.92E-04	-0.851	0.394746	
Female_No_HB_CEN_2010	-5.51E-03	1.63E-03	-3.375	0.000739	***
Med_HHD_Inc_ACS_08_12	1.73E-06	3.96E-06	0.437	0.661985	
Tot_Vacant_Units_CEN_2010	-1.25E-05	5.21E-04	-0.024	0.980821	
Tot_Vacant_Units_ACS_08_12	-3.10E-04	4.45E-04	-0.697	0.485761	
Mobile_Homes_ACS_08_12	2.52E-04	4.99E-04	0.505	0.613499	
No_Plumb_ACS_08_12	3.03E-03	1.34E-03	2.251	0.024368	*
FRST_FRMS_CEN_2010	2.45E-04	3.20E-04	0.765	0.444259	

Appendix B-4: Summary of R Logistic Model

Variable	Estimate	Std. Error	z value	Pr(>|z|)	Importance
Low_Response_Score	-1.57E-01	3.14E-02	-4.999	5.76E-07	***
pct_Pop_5_17_CEN_2010	9.25E-02	2.67E-02	3.467	0.000525	***
pct_Pop_18_24_CEN_2010	2.01E-02	1.15E-02	1.747	0.080686	.
pct_Hispanic_ACS_08_12	2.25E-02	1.79E-02	1.255	0.20936	
pct_NH_White_alone_CEN_2010	-7.93E-03	1.51E-02	-0.526	0.598664	
pct_NH_White_alone_ACS_08_12	2.82E-03	1.67E-02	0.169	0.866023	
pct_NH_Blk_alone_CEN_2010	-1.92E-02	3.49E-02	-0.549	0.582679	
pct_NH_Blk_alone_ACS_08_12	-7.54E-03	3.64E-02	-0.207	0.835734	
pct_NH_Blk_alone_ACSMOE_08_12	-6.89E-02	5.12E-02	-1.346	0.178412	

Appendix B-5: Summary of R Logistic Model

Variable	Estimate	Std. Error	z value	Pr(>|z|)	Importance
pct_NH_Asian_alone_CEN_2010	8.32E-02	2.54E-02	3.28	0.00104	**
pct_Pop_5yrs_Over_ACSMOE_08_12	5.62E-03	7.37E-03	0.763	0.445594	
pct_Age5p_Only_Eng_ACS_08_12	2.38E-02	9.10E-03	2.616	0.008886	**
pct_Age5p_Scandinav_ACSMOE_08_12	-5.97E-01	3.17E-01	-1.883	0.059733	.
pct_Pop_25yrs_Over_ACSMOE_08_12	1.82E-03	1.12E-02	0.162	0.870964	
pct_Not_HS_Grad_ACS_08_12	3.05E-02	1.72E-02	1.773	0.076217	.
pct_Not_HS_Grad_ACSMOE_08_12	3.01E-02	3.14E-02	0.958	0.338127	
pct_College_ACS_08_12	1.96E-02	1.08E-02	1.825	0.068004	.
pct_College_ACSMOE_08_12	-1.28E-02	6.30E-02	-0.203	0.839113	

Appendix B-6: Summary of R Logistic Model

Variable	Estimate	Std. Error	z value	Pr(>|z|)	Importance
pct_Prs_Blw_Pov_Lev_ACSMOE_08_12	1.10E-01	5.39E-02	2.04	0.041319	*
pct_Civ_emp_16p_ACS_08_12	3.69E-02	1.42E-02	2.589	0.009622	**
pct_Civ_unemp_16p_ACSMOE_08_12	3.91E-02	3.57E-02	1.096	0.273166	
pct_Pop_1yr_Over_ACSMOE_08_12	-2.55E-04	6.12E-03	-0.042	0.966834	
pct_Diff_HU_1yr_Ago_ACSMOE_08_12	-9.54E-02	5.86E-02	-1.628	0.103529	
pct_Born_US_ACS_08_12	1.21E-02	1.21E-02	1.001	0.316609	
pct_MrdCple_HHD_ACS_08_12	1.54E+07	9.85E+06	1.564	0.117868	
pct_MrdCple_HHD_ACSMOE_08_12	4.12E-02	2.44E-02	1.691	0.090911	.
pct_Not_MrdCple_HHD_ACS_08_12	1.54E+07	9.85E+06	1.564	0.117868	

Appendix B-7: Summary of R Logistic Model

Variable	Estimate	Std. Error	z value	Pr(>|z|)	Importance
pct_Female_No_HB_ACS_08_12	1.52E-02	1.05E-02	1.442	0.149405	
pct_HHD_PPL_Und_18_CEN_2010	-1.70E-03	1.57E-02	-0.108	0.913813	
pct_Tot_Occp_Units_CEN_2010	1.94E+01	7.97E+01	0.244	0.807403	
pct_Vacant_Units_CEN_2010	1.95E+01	7.97E+01	0.244	0.80729	
pct_Vacant_Units_ACSMOE_08_12	5.37E-02	2.70E-02	1.989	0.04668	*
pct_Single_Unit_ACS_08_12	-1.74E-02	5.12E-03	-3.407	0.000658	***
pct_MLT_U2_9_STRC_ACSMOE_08_12	-4.75E-03	2.47E-02	-0.192	0.847779	
pct_Mobile_Homes_ACS_08_12	-2.85E-02	1.33E-02	-2.14	0.032328	*
pct_Mobile_Homes_ACSMOE_08_12	-7.99E-03	3.23E-02	-0.247	0.804758	

Appendix B-8: Summary of R Logistic Model

Variable	Estimate	Std. Error	z value	Pr(>|z|)	Importance
pct_No_Plumb_ACS_08_12	-4.60E-02	2.72E-02	-1.688	0.091397	.
pct_Census_Mail_Returns_CEN_2010	4.24E-02	4.27E-02	0.995	0.319947	
pct_Mailback_Count_CEN_2010	-2.14E-03	3.34E-02	-0.064	0.948889	
pct_FRST_FRMS_CEN_2010	-2.83E-02	1.69E-02	-1.675	0.093863	.