BUDT 733: Data Mining for Business
Fall 2009

Team 4 Project Assignment: Predicting the Safety of U.S Cities

Elyas Akram
Minying Chen
Phuong Le
Bavan Vargese
Fan Zhang

EXECUTIVE SUMMARY

Being in the Washington tri-state area (Maryland, Virginia and Washington, D.C), our team understands how and why home prices in different areas within the tri-state area could be very different despite them being within a 10-mile radius of each other. Of the multiple factors that contribute to these differences, we believe that the violent crime rate of a city is a big determinant of the average home prices in a specific area. The currently available online real estate services like Zillow.com and Redfin.com provide users with information like the available homes for sale, the average home prices and expected taxes to be paid, but does not incorporate an interface that can allow users to infer the "safety" of the area of a home that someone is interested in.

This leads us to the motivation for this project – to see if certain socio-economic factors influence the violent crime rates of cities in the United States, and hence can be used to predict the safety of these cities. The goal of our project is thus to classify U.S. cities as "safe" or "unsafe" based on expected violent crime rates. Initial research, domain knowledge and some rational conjecture led to a somewhat unconventional list of possible factors that could influence the violent crime rate in any specific city. These include things like the number of fast-food restaurants, the number of universities, the number of Wal-Mart stores, and the number of foreclosed homes in the city. We then obtained our source data from four different websites: 2007 crime data from the FBI website[1], number of fast food restaurants, number of Starbucks and number of Wal-Marts in each U.S. city from the POI Factory website[2], the 2007 housing data from the U.S. Department of Housing and Urban Development website[3], and the number of universities in U.S. cities from the UnivSource website[4].

The initial step in data preparation was to merge data from the four sources based on a common identity field, which in our case, is the combination of city and state names. In the data exploratory stage, we attempted to identify trends and possible correlations between variables by plotting multiple scatter plots, histograms and box plots. Exhibit B shows a few of the more significant charts. Due to the predictive goal of our project, we partitioned the data into 60% training and 40% validation sets prior to running the k-NN model, the Discriminant Analysis (DA) model and the Logistic Regression (LR) model. The k-NN model has an overall validation RMS error of 22.76% with the best k-value of 15. The DA model has an overall validation error of 10.49%, and the LR model has the lowest validation error of 5.21%. We used the default cutoff of 0.5 for all models.

Due to its lowest validation overall error out of all 3 models, and the high significance of almost all the variables in the model, we selected logistic regression as our final predictive model for this project. Despite this being a predictive exercise, we looked at the variables of the final model used to see if they make sense. The high level interpretation from the logistic regression, as can be seen in Exhibit E, is that the more number of Starbucks cafes there are in a city, the higher the odds of the city being "safe", based on the expected lower crime rates, all other variables remaining the same. In addition, as the number of mortgage-paying households increase, the odds that the city is deemed "safe" are also higher. These observations could be valuable factors in helping potential home-owners choose new homes, especially if the safety of the neighborhood is a big concern.

TECHNICAL SUMMARY

The goal of our project was predictive. For a predictive task, it is paramount that variables that can be used in the final model be available prior to utilizing the model. In our case, we want to predict whether a city is safe or not safe based on its crime rate, which, in turn, is assumed to be influenced by certain socio-economic variables which are available simply by pure observation. The socio-economic variables utilized in our models are the number of fast food restaurants such as McDonald's, Wendy's, Burger King, the number of Wal-Mart stores, the number of Starbucks cafes, and the number of for-profit, non-profit and public universities. For the

[1] http://www.fbi.gov/ucr/cius2007/data/table_08.html
[2] http://www.poi-factory.com/about
[3] http://www.huduser.org/portal/index.html
[4] http://www.univsource.com/ussc.htm

purpose of our analysis, we assume the number of fast food restaurants, number of Wal-Marts, number of Starbucks and number of universities to remain static.

Data Preparation and Exploration

After we obtained the source data from the four different websites as mentioned earlier, we spent a significant amount of time merging the data via the common identifiers – State and City names. We also had to decide which variables to keep, if there were missing values and how to deal with them, and also to remove repeated data rows. We eventually decided that certain variables such as "Police Budget as a % of City Budget" had too many missing values to be useful in our models (about 70% missing data) and therefore removed them from our analyses. We also derived new variables by dividing the initial variables which were absolute counts of the restaurants, Wal-Mart stores and universities by the population size of each city and multiplying that by 100,000 to derive the "per 100,000 of population" number. Doing this enabled us to get the variables down to a same unit where we can compare same-sized fields. We reduced the original approximately 9,400 rows of raw data to 3,456 rows of data that would be used in our data mining models.

The output variable for our data mining models is the "Safe/Unsafe" field, which is derived from the "2007 Violent crime as a % of 2007 population" field. In our final logistic regression model, a city is deemed to be "Safe" when the 2007 percentage (%) violent crime rate is less than or equal to 0.10%, and deemed to be "Unsafe" when the 2007 % violent crime rate is more than 0.10%. The success class is "1" for the "Unsafe" category, and "0" for the "Safe" category. "Violent Crime" constituted murder, non-negligent manslaughter, forcible rape, robbery and aggravated assault cases in our analyses. Exhibit A shows the list of variables.

We plotted an extensive number of graphs and charts to deduct trends and correlations between the different variables that could influence the crime rates of cities. We initially observed the lack of any particularly strong effect of any specific variable on the crime rate of U.S. cities. As shown in the left scatter plot graph in Exhibit B, where we aggregated the number of all the fast food restaurants and plotted that against the number of Wal-Mart stores to see if these variables greatly influence the crime rates. We observed that the "unsafe" cities are concentrated at the bottom left corner within the confines of the rectangle (the red dots). We can possibly imply from this that there is no direct linear relationship between the number of fast food restaurants and the number of Wal-Mart stores with the crime rates in cities. "Unsafe" cities tend to culminate in the area where the aggregated number of fast food restaurants range from 1 to 500, and the number of Wal-Mart stores range from 0 to 50 per city, but the data does not suggest that cities become more unsafe as the number of fast food restaurants and Wal-Mart stores increase beyond these numbers.

The bottom right graph in Exhibit B is interesting in that we see the highest average violent crime rates in cities where there are very few McDonalds restaurants and very few Starbucks cafes, but the lowest average violent crime rates were also observed where there are very few McDonald's restaurants and a lot of Starbucks restaurants (the orange bar in the "1 – McD Very Low" category). This observation makes practical sense when we postulate that McDonald's and Starbucks also evaluate their location options when they open new outlets, and hence they want to avoid locations with observed high crime rates in general. Cities with very few McDonalds and a lot of Starbucks tend to be more affluent and/or have a higher concentration of white collar jobs, and thus due to the very nature of these cities, have less occurrences of violent crime, on a relative scale. One important thing we gathered from the exploratory stage is that there is not necessarily a causation relationship between our input and output variables – that the number of fast food restaurants, Wal-Mart stores, Starbucks cafes and universities do not necessarily lead to the violent crime rate being at a certain level, and vice versa.

k-Nearest Neighbor (k-NN) Model

We first ran a k-NN model on the processed dataset in XLMiner. The optimal model derived after multiple iterations with different variables is shown in Exhibit C. This was the model where we achieved the lowest validation RMS error of 22.76%. There are nine input variables including the number of fast food restaurants, Wal-Mart stores, Starbucks cafes and public universities per 100,000 population, the foreclosure rate and the percentage of residential addresses that have been vacant for at least 90 days. The output variable is

Safe/Unsafe (1 if crime rate is greater than 0.0847% and 0 if otherwise). We standardized the input variables prior to running the analysis in XLMiner.

The best K-value is 15, and based on the lift chart, the k-NN model does provide a big prediction improvement over the naïve rule. Based on our historical records, there were 458 "unsafe" cities and 2998 "safe" cities, and hence the naïve rule would cause all future records to be classified as "safe". Next, we look at the decile-wise lift chart. We see that the k-NN model provides us with a 60% improvement over the naïve rule for the top 10% of records. Using this model gives us almost a 100% improvement for the top 30% records. One of the major disadvantages of the k-NN model, however, is the fact that it is extremely computation-intensive. It took us about 30 minutes to run each iteration on 3,456 records. Another disadvantage of the k-NN model, especially for an assignment where the goal is explanatory, is that it does not provide insight into why the input variables are important. Although our goal in this initiative is to predict crime rates, we were also very interested in understanding the impact of our various predictors on crime rates, insight that k-NN did not provide.

Discriminant Analysis Model

Once again using standardized variables, we ran a discriminant analysis on our dataset. Due to the fact that discriminant analysis does not handle outliers well, we removed the top and bottom 5% percentile of our records. This equates to about 346 records in total, and included outlier cities like New York city and Los Angeles where there are extremely high populations, high violent crime rates and high number of fast food restaurants.

Our final best iteration which has the lowest validation overall percentage error rate of 10.49% is shown in Exhibit D. The success class, 1, is once again "Unsafe", and as in the k-NN model, we deem a city to be unsafe when the violent crime rate is more than 0.0847%, which was the average violent crime rate for all the 3,456 records we have. As we can see, in this optimal DA model, the best discriminants appear to be the total number of fast food restaurants per 100,000 people and the foreclosure rate, which are similar important variables in our optimal k-NN model as mentioned in the previous section. Once again, even though our goal is predicting, we also want to be assured that our model makes sense, and in this instance, our optimal models in both the k-NN and DA analyses agree with each other in terms of significant variables. The overall error rate of 10.49% is also an improvement over the 13.25% error rate based on the naïve rule (458 / 3,456 x 100%).

Logistic Regression Model

Our team also ran the logistic regression model, which is useful for both explanatory and predictive purposes. We started running the LR with input variables based on our domain knowledge, trends seen in the charts and graphs we ran in the exploratory stage, and also inputs that have been deemed important in both the k-NN and DA models. Keeping in mind that XLMiner allows only a maximum of 30 input variables, we also made sure that variables that could be highly correlated are not included in the model (for example, we did not run the "# of xxx" and "# of xxx per 100,000 pop." variables in one same model. Once again, our output variable is Safe/Unsafe, with 1 being the success class of "Unsafe" and 0 if "Safe". After multiple iterations, we obtained the final model that is shown on the next page.

All of the input variables have very low p-values except for "# BK", "# McD" and "#Wal-Mart". We attempted dropping each of these 3 variables one by one but the overall validation error rate actually increased, and the number of false negatives (Predicted to be "Safe" but is "Unsafe" in actuality) also increased. Keeping in mind that we have a predicting task, we kept these three variables in our final model. The validation overall error rate in our optimal LR model is 5.21%, which is also shown.

The Regression Model

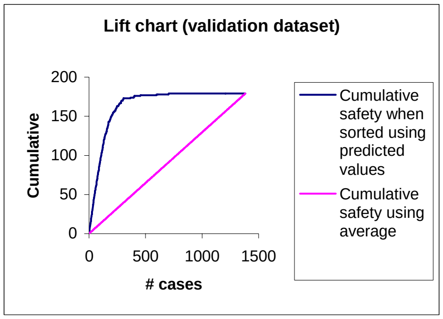
Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-8.08065605	0.49215192	0	*
2007-Population	0.0000795	0.00000919	0	1.00007951
# BK	0.14088055	0.09566267	0.14083719	1.15128708
# McD	0.06560349	0.04369678	0.13326901	1.06780326
# Starbucks	-0.18663469	0.0697379	0.00744554	0.82974678
# Walmart	0.04491873	0.07991253	0.57404876	1.0459429
# Wendys	0.26040083	0.1120597	0.02013789	1.29745007
Tot.# Universities	0.2134254	0.07446967	0.00415777	1.23791111
# Households	0.00008326	0.00002292	0.00028125	1.00008321
# Est. Mortgages	-0.00018743	0.00003195	0	0.9998126
# Est. foreclosures as a % of Est.# of Mortgages	34.21318054	4.59639931	0	7.22097E+14

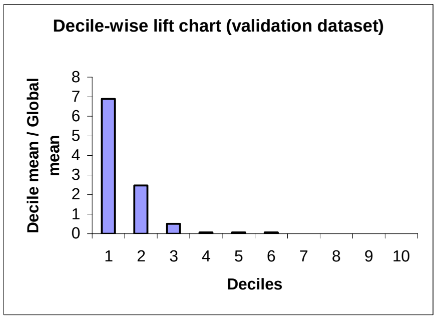
Residual df	2061
Residual Dev.	487.0836487
% Success in training data	13.46525097
# Iterations used	10
Multiple R-squared	0.70253426

Cut off Prob.Val. for Success (Updatable)	0.5

Classification Confusion Matrix		
	Predicted Class	
Actual Class	unsafe	safe
unsafe	131	48
safe	24	1179

Error Report			
Class	# Cases	# Errors	% Error
unsafe	179	48	26.82
safe	1203	24	2.00
Overall	1382	72	5.21





As can be seen from the lift chart, the LR model clearly is superior to the Naïve case, like k-NN. The Decile-wise lift chart shows that the LR model provides us with a near 70% improvement over the naïve rule for the top 10% of records, which is higher than that obtained in the k-NN model.

Let us now look at the output variables in the LR model in more detail:

Logit = -0.81+0.000080(2007 pop)+0.14(#BK)+0.07(#McD- 0.19(#Starbucks)+0.05(#Wal-Mart)+0.26(#Wendys) +0.21(Tot.# Universities)+0.000083(# Households)-0.00019(# Est.Mortgages)+34.21(#Est.foreclosures as % of Mortgages)

Looking at the odds for each of the variables, we observe that as we increase each of the variables by one unit, the odds factor that the city is deemed "unsafe" increases, except for "# Starbucks" and "# Est.Mortgages", keeping all other variables the same. As the number of Starbucks store increases by one in a certain city, the odds that that city will be deemed "unsafe" *decreases* by a factor of 1.204, keeping all other variables the same. Likewise, as the number of households that have mortgages increase by 1, the odds that that city will be deemed "unsafe" *decreases* by a factor of approximately 1, keeping all other variables the same. This conclusion actually fits well with the earlier observation in the exploratory stage that violent crime rate was the lowest in cities with a lot of Starbucks. Also, if home-ownership is high in a city, the violent crime rates are predicted to be lower.

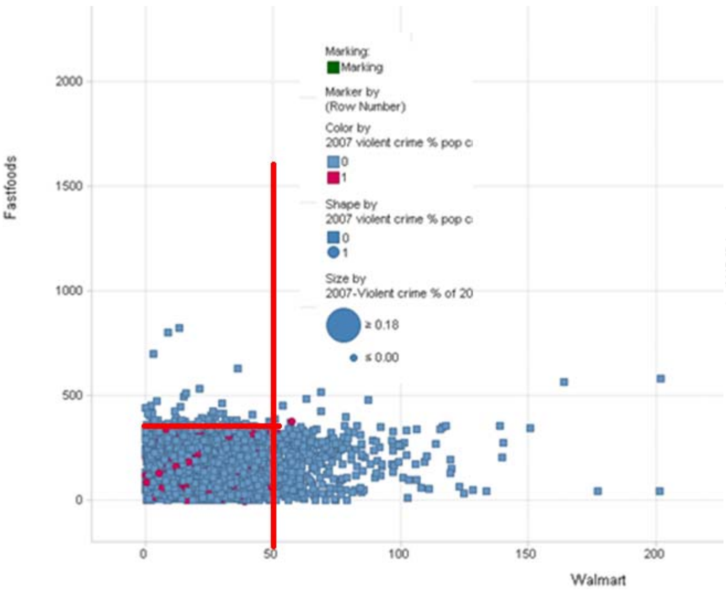
Conclusion

The logistic regression is our chosen model for the predictive task we have on hand. In addition to the fact that the LR model yielded the lowest overall validation error rates among all three models (highest accuracy), the LR model far out-performed the k-NN model based on the results as can be seen in the lift charts. The input variables in the LR model also made good rational sense and is therefore a very reasonable model. Despite this being a predictive task, in the event that someone (for example, Zillow.com) is interested in understanding the workings behind the predictive model, the variables themselves allow for easy interpretations. In addition, the LR model is much faster and efficient to run, especially over that of the k-NN model, and hence is less costly in terms of both resources and time to run and maintain. Last but not least, we believe that this is the most parsimonious and robust model when viewed in light of the fact that the amount of historical data used to refine the model will eventually get larger over time.

Exhibit A: Variable Descriptions

Variable	DataType	Description
City	String	U.S. city name
State	String	U.S. state name
2007-Population	Numerical	Total 2007 population size of the U.S. city
2007-Violent crime % of 2007 Pop	Numerical	Total violent crime (inclusive of the # of murder, non-negligent manslaughter, forcible rape, robbery & aggravated assault cases) as a % of 2007 population
Safe/Unsafe	Categorical	Final Output Variable - "Safe" if 2007 % Violent crime rate more than or equal to 0.10%; "Unsafe" if less than 0.10%
2007-Property crime % of 2007 Pop	Numerical	# of property crime cases as a percentage of 2007 population
2007-Burglary % of 2007 Pop	Numerical	# of burglary cases as a percentage of 2007 population
2007-Larceny-theft % of 2007 Pop	Numerical	# of larceny theft cases as a % of 2007 population
2007-Motor vehicle theft % of 2007 Pop	Numerical	# of motor vehicle theft cases as a % of 2007 population
# BK	Numerical	# of Burger King restaurants in the city
# of BK per 100,000 pop	Numerical	# of Burger King restaurants per 100,000 population
mcd_bin	Categorical	5 bins based on the# of McD per 100,000 population: 5 - If # of McD per population>60; 4 if # of McD per population >=45; 3 if # of McD per population >=30; 2 if # of McD per population >=15; 1 if # of McD per population <15
starbuck_bin	Categorical	5 bins based on the# of Starbuck per 100,000 population: 5 - If # of Starbuck per population>60; 4 if # of Starbuck per population >=45; 3 if # of Starbuck per population >=30; 2 if # of Starbuck per population >=15; 1 if # of Starbuck per population <15
# McD	Numerical	# of McDonalds restaurants in the city
# McD per 100,000 pop	Numerical	# of McDonalds restaurants per 100,000 population
# StarBucks	Numerical	# of Starbucks cafes in the city
# of StarBucks per 100,000 pop	Numerical	# of Starbucks cafes per 100,000 population
# Walmart	Numerical	# of Walmart stores in the city
# Walmart per 100,000 pop	Numerical	# of Walmart stores per 100,000 population
# Wendys	Numerical	# of Wendys restaurants in the city
# Wendys per 100,000 pop	Numerical	# of Wendys restaurants per 100,000 population
Tot. # FastFd Restaurants	Numerical	Total # of fast-food restaurants(# of McDonald's + # of Burger King's + # of Wendy's)
Tot. # FastFd Restaurants per 100,000 pop	Numerical	Total # of fast-food restaurants(# of McDonald's + # of Burger King's + # of Wendy's) per 100,000 population
# ForProfitPrivateUniversity	Numerical	# of For-profit private universities in the city
# ForProfitPrivate Uni per 100,000 pop	Numerical	# of For-profit private universities in the city oer 100,000 population
# NonProfitPrivateUniversity	Numerical	# of Non-profit private universities in the city
# NonProfitPrivate Uni per 100,000 pop	Numerical	# of Non-profit private universities in the city per 100,000 population
# PublicUniversity	Numerical	# of Public universities in the city
# Public Uni per 100,000 pop	Numerical	# of Public universities in the city per 100,000 population
Tot. # Universities	Numerical	Total # of For-profit, non-profit, and public universities in the city
Tot. # Universities per 100,000 pop	Numerical	Total # of For-profit, non-profit, and public universities in the city per 100,000 population
# HouseHolds	Numerical	Total # of households in the city
# Est. Mortgages	Numerical	Total # of estimated mortgage-paying households
# Est.Foreclosures as a % of est. # mortgages	Numerical	Total # of foreclosed households as a % of total number of mortgage-paying households
Fore_bin	Categorical	5 bins based on the # of foreclosed households as a % of total number of mortgage-paying households : 5 - If % is >9%; 4 if % is between 9% and 7%; 3 if % is between 4.9% and 7%; 2 if % is between 3% and 4.9%; 1 if % is =< 3%
# Tot. Residential Add.	Numerical	Total # of residential addresses
Tot. 90D VacResidential Add.as % of Tot. Residential Add.	Numerical	Total # of vacant for at least 90 days residential addresses as a % of tota # ofl residential addresses

Exhibit B: Exploratory Charts



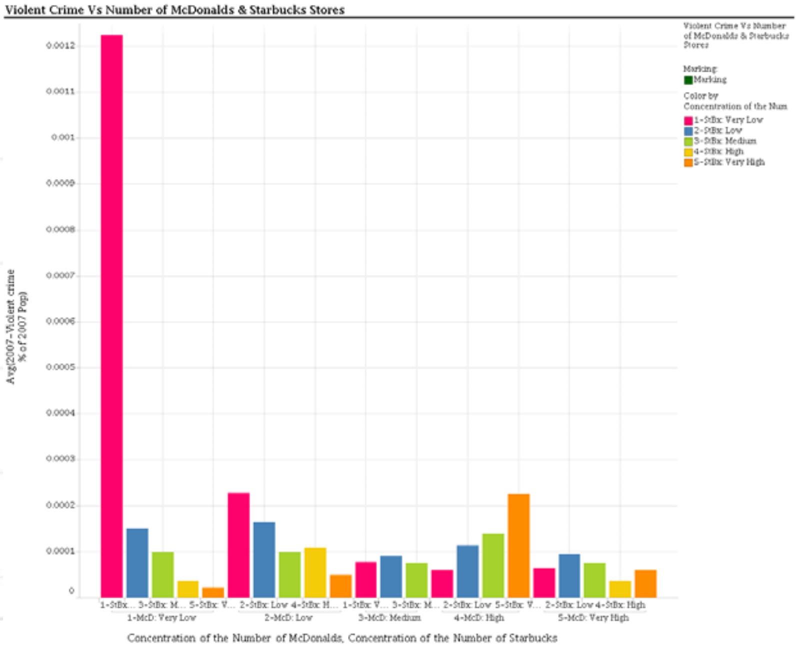
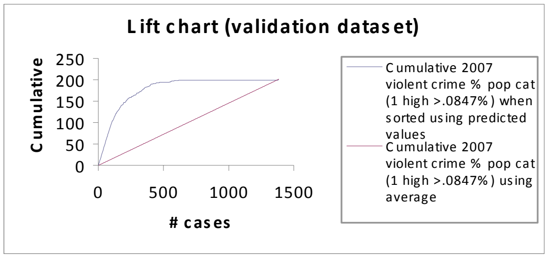


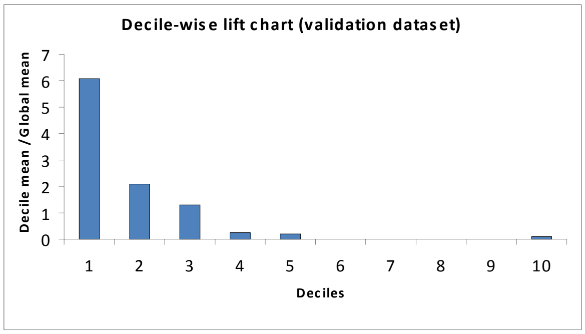
Exhibit C: k-NN XLMiner Output

Variables									
# Input Variables	9								
Input variables	2007-Population	# of BK per 100,000 pop	# McD per 100,000 pop	# of StarBucks per 100,000 pop	# of Walmart per 100,000 pop	# of Wendys per 100,000 pop	# of Public Unis per 100,000 pop	# Est.Foreclosures as a % of est. # mortgages	Tot. 90D VacResidential Add.as % of Tot. Residential Add.
Output variable	2007 violent crime % pop cat (1 high >.0847%)								

Validation error log for different k

Value of k	Training RMS Error	Validation RMS Error	
1	0	0.306702739	
2	0	0.272343332	
3	0	0.255812021	
4	0	0.244035302	
5	0	0.237604162	
6	0	0.237632813	
7	0	0.235723763	
8	0	0.232005874	
9	0	0.230339586	
10	0	0.229602763	
11	0	0.228596124	
12	0	0.227948988	
13	0	0.228181732	
14	0	0.227868784	
15	0	0.227569677	<--- Best k
16	0	0.228463137	
17	0	0.228463608	
18	0	0.228062102	
19	0	0.228097648	
20	0	0.228043071	





Validation Data scoring - Summary Report (for k=15)

Total sum of squared errors	RMS Error	Average Error
71.57095769	0.227569677	0.001263034

Exhibit D: Discriminant Analysis XLMiner Output

Classification Function

Variables	Classification Function	
	1	0
Constant	-1.94311512	-0.72962779
# Est. Foreclosures as a % of est. # mortgages	0.70414591	-0.1114331
Tot. 90D VacResidential Add.as % of Tot.Residential Add.	-0.14286315	-0.00923953
Tot. # FastFd Restaurants per 100,000 pop.	1.5478282	-0.26120687
Tot. # Universities per 100,000 pop.	0.22839473	-0.04210817

Validation Data scoring - Summary Report

Cut off Prob.Val. for Success (Updatable)	0.5

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	124	94
0	51	1113

Error Report			
Class	# Cases	# Errors	% Error
1	218	94	43.12
0	1164	51	4.38
Overall	1382	145	10.49