Health Insurance Coverage: Comparing the Insured and Uninsured in the United States

Prepared by:

Jason Davis
Clarette Kim
Hari Kosaraju
Scott Wood

BUDT 733 Final Team Project
Professor Shmueli
December 11, 2007

UNIVERSITY OF
MARYLAND

ROBERT H. SMITH
SCHOOL OF BUSINESS

Executive Summary

In response to increasing healthcare costs and diminished access to healthcare in the United States, our team analyzed demographic and socioeconomic data from the Census Bureau in order to profile the health insurance coverage of U.S. residents. This data included factors such as income, education, race, work experience, gender, geographic location, etc. After analyzing the data using numerous visualization techniques and by creating numerous models, including classification trees, logistic regressions, and discriminant analyses, we identified a number of variables that distinguish persons *without* health insurance coverage from persons *with* health insurance coverage. Based on this analysis, we have developed several key recommendations designed to guide healthcare policy decisions and steer philanthropic activity in an attempt to improve nationwide access to healthcare. These key findings and recommendations include:

1. **Target Hispanic populations in Southern or Western states when designing new healthcare initiatives or refocusing existing health insurance initiatives.** This strategy is supported by our logistic regression analysis and discriminant analysis, Exhibits D and E, both of which revealed that a person of Hispanic heritage was more likely *not* to have health insurance to a statistically significant degree as compared to a non-Hispanic person, when controlling for all other factors. The effect of Hispanic heritage on increased likelihood for no insurance was larger than that of any other ethnic or racial group. In addition to the statistical evidence presented by the models, Hispanic individuals were also overrepresented in the ranks of the uninsured when the data set was visualized graphically, as seen in Exhibit A. In addition to Hispanic ethnicity, our models indicated that people living in Southern states, and to a lesser degree people living in Western states, are more likely to lack insurance. As such, we recommend that any healthcare initiative designed to target the uninsured focus on these two factors.

2. **Target the working rural poor in the United States when designing new healthcare initiatives or refocusing existing health insurance initiatives.** This strategy is strongly supported based on our data visualizations and classification tree analysis. These charts indicate that a disproportionate number of *uninsured* individuals are employed but make less than $54,652 per year in household income. The income factor is most clearly highlighted by our classification tree analysis, Exhibit C. In addition, the logistic regression and discriminant analysis models revealed that people *not* living in a metropolitan area were more likely to be uninsured. In our estimation, by targeting this profile, a significant number of people should gain the opportunity to have healthcare coverage.

3. **Target young, unmarried individuals and those who have only a high school education (or less) when designing new healthcare initiatives or refocusing existing health insurance initiatives.** This strategy is supported by our logistic regression and discriminant analysis, which identified people who are younger, who are not married, or who have achieved a high school education (or less) as having a statistically significant higher likelihood of lacking health insurance.

Technical Summary

Purpose

Our goal was to use explanatory data mining techniques to better understand the characteristics of the U.S. population that lack health insurance. Our results will be used to inform health policy and business officials as they develop strategies to address the uninsured population.

Data Description and Pre-Processing

The dataset consists of a random sample of cross-sectional data on 30,000 individuals from the 2007 U.S. Current Population Survey.[1] About 15% of the sample (4,627 of 30,000 records) are uninsured, and 85% are insured. We used domain knowledge to identify 12 key predictor variables, listed below. Almost all of the predictors (except age and income) are categorical variables. We performed several data pre-processing tasks on the data, including recoding, binning, transforming into dummy variables, and partitioning (for classification tree).

Variable	Description
Insured	Binary response variable for health insurance status: 1=Insured, 2=Uninsured
Age	Age in years (Numerical variable)
HHIncome	Household Income (Numerical variable)
WorkExp	Experienced labor force employment status: 1=Employed, 2=Not employed
Educ	Highest level of educational attainment: 31-39=High School or less; 40-43=some college to Bachelor's degree; 44-46=Master's to PhD degree
Marital	1-3= Married, 4=widowed, 5=divorced, 6=separated, 7=never married
Sex	Sex. 1=Male, 2=Female
Region	Region of U.S. where respondent is located: 1=Northeast, 2=Midwest, 3=South, 4=West
Metro	Living in metro or non-metro area: 1=Metropolitan, 2=non-metropolitan, 3=unidentified
Health	Health condition reported by respondent: 1=Excellent, 2=Very good, 3=Good, 4=Fair, 5=Poor
Hispanic	1=yes, 2=no
Citizen	Citizenship: 1=Native born U.S., 2=Foreign-born naturalized, 3=Not a U.S. Citizen
Race	For simplicity, limited to these values: 1=White, 2=Black, 4=Asian.

Observations from Exploratory Analysis

Preliminary exploration of the data—including scatterplots, summary statistics, boxplots, and distributions—revealed several variables that may have explanatory power in distinguishing the insured and uninsured groups.

- **Citizenship and Hispanic status. Exhibit A** shows that Hispanic non-citizens are significantly more likely to be uninsured than any other group. At the same time, more than 50% of the uninsured population are actually non-Hispanic citizens, indicating that there are also other factors explaining one's health insurance status.
- **Age and Income.** On average, the insured tend to have higher incomes and be slightly older than the uninsured.
- **Dependence between Categorical Variables.** We also used chi-square testing to identify relationships among pairs of categorical predictors. **Exhibit B** shows selected pairs that exhibit significant dependency (but to varying degrees), based on their low p-values and high chi-squares. The variables with the strongest interdependency were Hispanic, Citizenship, and Education. Marital status, Education, and Race were also strongly interdependent.

[1] The Current Population Survey (March 2007 Supplement) is a high-quality data source published annually by the U.S. Census Bureau, based on data collected from a nationally representative survey of U.S. households.

Model Estimation and Interpretation

Classification Tree. According to our pruned classification tree (**Exhibit C**), the profile of individuals most likely to be uninsured are Hispanic non-citizens with household incomes below \$54,652.[2] The classification tree indicates that these three variables—household income, Hispanic status, and citizenship status—hold the most power in separating the insured group from the uninsured group. Although this pruned model only had a low overall error of 14.39% for the validation set, it had a much higher error (83.18%) in classifying the uninsured group in the validation set. This high classification error for the uninsured suggests that several other variables may also be useful in profiling the uninsured. For example, a Minimum Error Classification Tree was also run and revealed that three other variables also play a strong role in the profile of an uninsured person: bSouth ('1' if lives in a southern state or '0' otherwise), bEducation ('0' if has a high-school education or below, '1' if above), and Age.

Logistic Regression. Using the knowledge obtained from the visualizations and classification trees, we then performed numerous logistic regression models in order to better understand the profile of the uninsured.[3] The results of the final model (**Exhibit D**) reveals that the statistically significant factors with the largest effect in increasing one's odds of lacking health insurance in the US are: Hispanic heritage, location in the southern region of the U.S., non-US-citizenship status, and having lower household income. Other statistically significant factors include: being younger, having a lower amount of education, living in rural (rather than metropolitan) areas, not being married, being African American, being Asian, or living in a western state.

The final logistic regression was determined iteratively. We started with a model that included most of the variables that appeared to be important in the visualizations or classification trees (14 variables total). Then we systematically removed one variable at a time (i.e. the one with the highest p-value) with each new model created until all the p-values were significant. The first logistic regression model, revealed several variables that were statistically insignificant at the 5% level: employment status (WORKEXP, p = 0.81), reported health condition (HEALTH, p = 0.191), gender (SEX, p = 0.109), and whether a person is African American (bBlack p = 0.08). As a result, the final model (Exhibit D) no longer uses the variables WORKEXP, HEALTH, and SEX given the very high p-values, but it does include the variable bBlack, since its p-value was very close to 5%. Compared to the classification tree results, this model had both a lower overall error rate (12.02%) and a much lower error rate (48.27%) for the uninsured group. The Multiple R-squared of this 11-variable model was 0.3769, which was essentially equivalent to that of the original 14-variable logistic regression model (0.3774).

Discriminant Analysis. Our discriminant analysis model (**Exhibit E**) demonstrates that the most significant categorical (binary) variables in separating the insured from the uninsured are: citizenship, Hispanic status, South region, West region, and college level education. This is determined based on the difference in classification function values for insured versus not insured.

[2] For the pruned classification tree, the dataset was partitioned into training, validation, and test sets of 10,000 records each.

[3] For the logistic regression models, we used 10,000 records out of the 30,000 available, with no data partitioning, since our goal was explanatory.

Evaluation of Model Performance: Goodness of Fit

Since we are seeking an explanatory model, model performance is based on a measure of overall fit, such as the Multiple R-squared. A performance summary of selected models is listed below:

Model	Significant Predictors	Overall Error	Error in Classifying Uninsured	Other Performance Measures
Classification Tree	Household Income, Hispanic status, Citizenship	14.39%	83.18%	
Logistic Regression	Household Income, Hispanic status, Citizenship, Income, Age, Education, Metro, Marital Status, Black, Asian	12.02%	48.27%	Multiple R-Square: 37.69%
Discriminant Analysis	Household Income, Hispanic status, Citizenship, Income, Age, Education, Marital Status, Black, Asian, South, and West	18.09%	17.14%	

Taken together, all three models provide useful insight into factors characterizing the uninsured population. The classification tree provides results that are easily understandable and identifies the three most significant predictors. The logistic regression confirms these three predictors, shows the relative usefulness of additional predictors, and also adds interpretability using the odds values. For example, being Hispanic increases the odds of being uninsured by a factor of 3.9. Finally, the discriminant analysis also confirms these key predictors at the lowest error rate (for classifying uninsured).

Conclusions: Public Policy and Business Implications

Our technical analysis results in some interesting findings, which form the foundation for some recommendations for health policy and industry officials.

- **Target Hispanic populations in Southern or Western states when designing new healthcare initiatives or refocusing existing health insurance initiatives.** This strategy is supported by our logistic regression and discriminant analysis, which revealed that a person of Hispanic heritage was more likely not to have health insurance, when controlling for all other factors. The effect of Hispanic heritage on increased likelihood for no insurance was larger than that of any other ethnic or racial groups. Also, people who live in southern, and to a lesser degree western, states are more likely to lack insurance.
- **Target the working rural poor in the United States when designing or refocusing health insurance initiatives.** This strategy is strongly supported based on our data visualizations and classification tree analysis, which indicate that a disproportionate number of *uninsured* individuals are employed but make less than $54,652 per year in household income. In addition, the modeling revealed that those not living in a metropolitan area were more likely to be uninsured.
- **Target young, unmarried individuals and those who have only a high school education (or less) when designing or refocusing health insurance initiatives.** This strategy is supported by our logistic regression and discriminant analyses, which identified people who are younger, who are not married, or who have only achieved a high school education as having a statistically significant higher likelihood of lacking health insurance.

Appendix

Exhibit A: Health Insurance Status Distribution by Citizenship and Hispanic Indicators

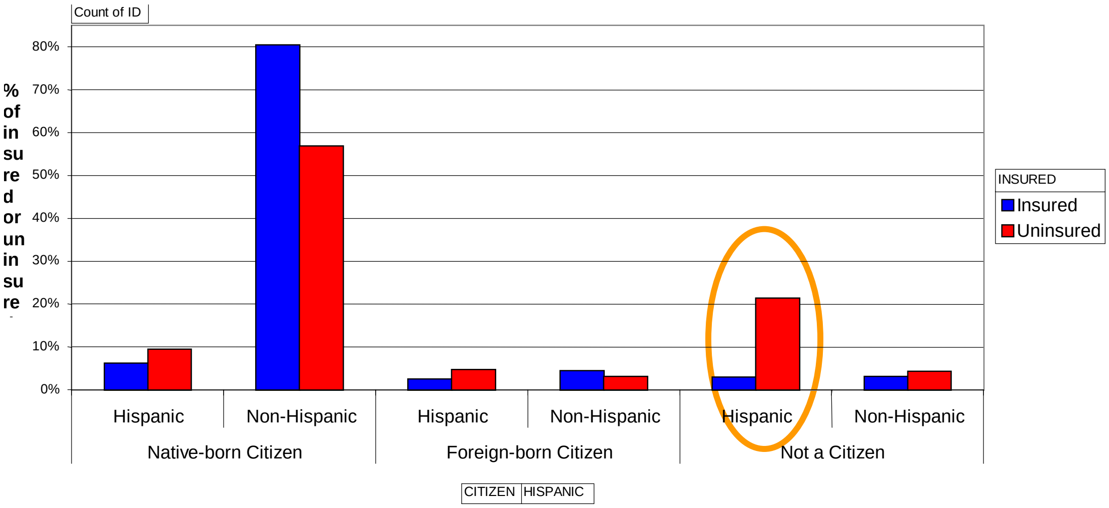
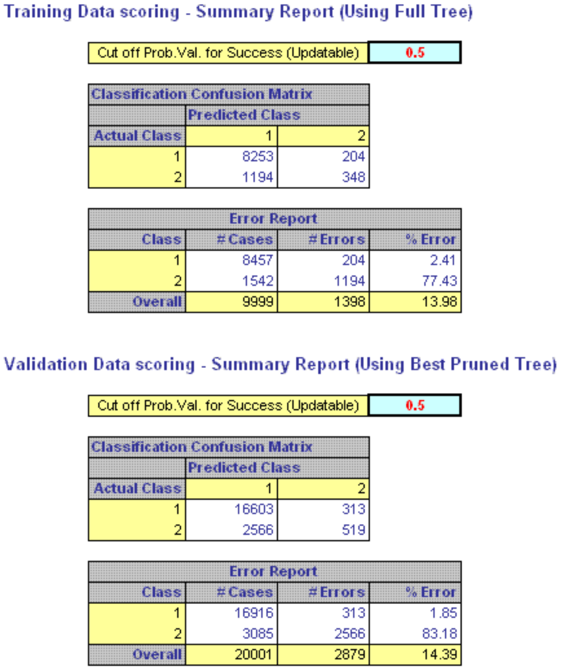


Exhibit B: Dependence between Categorical Variables

Categorical variable pair		p-value	Chi2-stat	Df
Hispanic	Citizen	0.00E+000	6971.96	2
Educ	Citizen	0.00E+000	5038.83	30
Hispanic	Educ	0.00E+000	4620.36	15
Marital	Race	0.00E+000	1867.56	12
Educ	Marital	2.02E-295	1700.18	90
Educ	Health	5.17E-247	1370.49	60
Hispanic	Region	2.15E-282	1303.85	3
Educ	Race	3.80E-141	762.78	30

Exhibit C: Pruned Classification Tree Results



Training Data scoring - Summary Report (Using Full Tree)

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

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	2
1	8253	204
2	1194	348

Error Report			
Class	# Cases	# Errors	% Error
1	8457	204	2.41
2	1542	1194	77.43
Overall	9999	1398	13.98

Validation Data scoring - Summary Report (Using Best Pruned Tree)

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

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	2
1	16603	313
2	2566	519

Error Report			
Class	# Cases	# Errors	% Error
1	16916	313	1.85
2	3085	2566	83.18
Overall	20001	2879	14.39

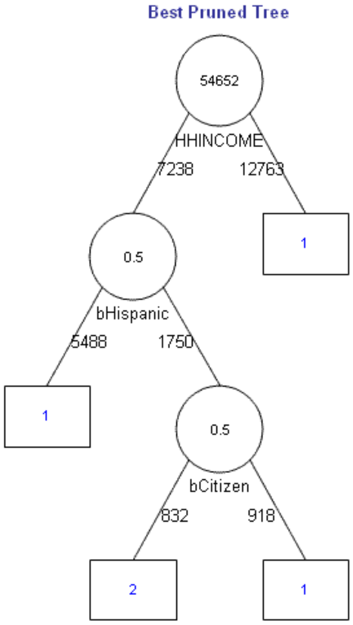
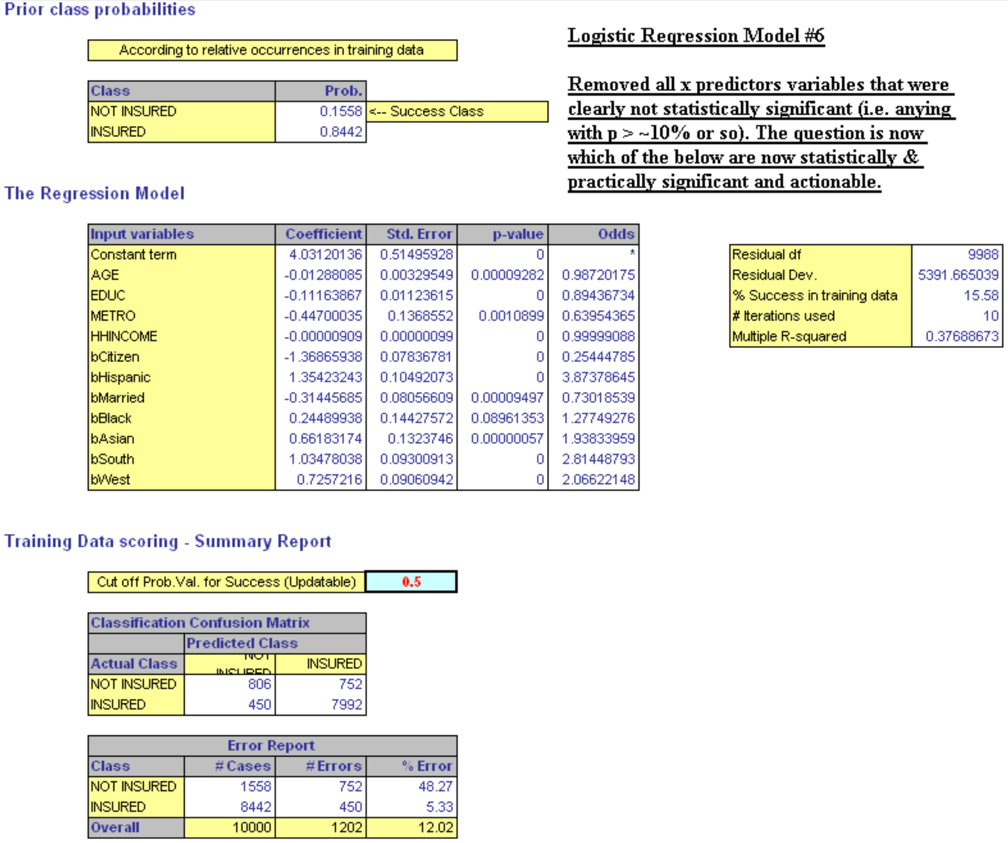


Exhibit D: Logistic Regression with Insignificant Variables Removed



Prior class probabilities

According to relative occurrences in training data

Class	Prob.	
NOT INSURED	0.1558	<-- Success Class
INSURED	0.8442	

Logistic Regression Model #6

Removed all x predictors variables that were clearly not statistically significant (i.e. anying with p > ~10% or so). The question is now which of the below are now statistically & practically significant and actionable.

The Regression Model

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	4.03120136	0.51495928	0	*
AGE	-0.01288085	0.00329549	0.00009282	0.98720175
EDUC	-0.11163867	0.01123615	0	0.89436734
METRO	-0.44700035	0.1368552	0.0010899	0.63954365
HHINCOME	-0.00000909	0.00000099	0	0.99999088
bCitizen	-1.36865938	0.07836781	0	0.25444785
bHispanic	1.35423243	0.10492073	0	3.87378645
bMarried	-0.31445685	0.08056609	0.00009497	0.73018539
bBlack	0.24489938	0.14427572	0.08961353	1.27749276
bAsian	0.66183174	0.1323746	0.00000057	1.93833959
bSouth	1.03478038	0.09300913	0	2.81448793
bWest	0.7257216	0.09060942	0	2.06622148

Residual df	9988
Residual Dev.	5391.665039
% Success in training data	15.58
# Iterations used	10
Multiple R-squared	0.37688673

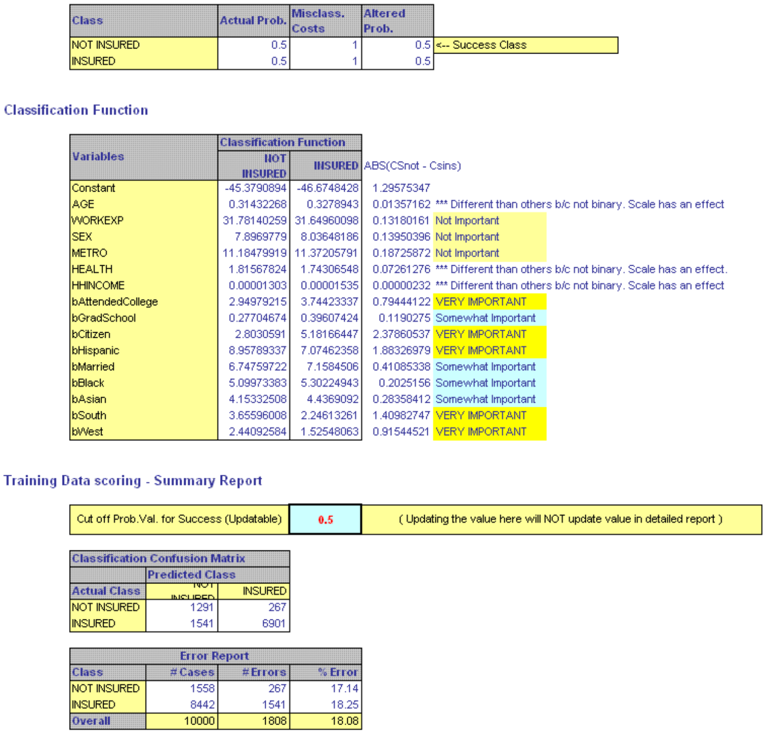
Training Data scoring - Summary Report

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

Classification Confusion Matrix		
	Predicted Class	
Actual Class	NOT INSURED	INSURED
NOT INSURED	806	752
INSURED	450	7992

Error Report			
Class	# Cases	# Errors	% Error
NOT INSURED	1558	752	48.27
INSURED	8442	450	5.33
Overall	10000	1202	12.02

Exhibit E: Discriminant Analysis



Class	Actual Prob.	Misclass. Costs	Altered Prob.	
NOT INSURED	0.5	1	0.5	<-- Success Class
INSURED	0.5	1	0.5	

Classification Function

Variables	Classification Function			
	NOT INSURED	INSURED	ABS(CSnot - Csins)	
Constant	-45.3790894	-46.6748428	1.29575347	
AGE	0.31432268	0.3278943	0.01357162	*** Different than others b/c not binary. Scale has an effect
WORKEXP	31.78140259	31.64960098	0.13180161	Not Important
SEX	7.8969779	8.03648186	0.13950396	Not Important
METRO	11.18479919	11.37205791	0.18725872	Not Important
HEALTH	1.81567824	1.74306548	0.07261276	*** Different than others b/c not binary. Scale has an effect.
HHINCOME	0.00001303	0.00001535	0.00000232	*** Different than others b/c not binary. Scale has an effect
bAttendedCollege	2.94979215	3.74423337	0.79444122	VERY IMPORTANT
bGradSchool	0.27704674	0.39607424	0.1190275	Somewhat Important
bCitizen	2.8030591	5.18166447	2.37860537	VERY IMPORTANT
bHispanic	8.95789337	7.07462358	1.88326979	VERY IMPORTANT
bMarried	6.74759722	7.1584506	0.41085338	Somewhat Important
bBlack	5.09973383	5.30224943	0.2025156	Somewhat Important
bAsian	4.15332508	4.4369092	0.28358412	Somewhat Important
bSouth	3.65596008	2.24613261	1.40982747	VERY IMPORTANT
bWest	2.44092584	1.52548063	0.91544521	VERY IMPORTANT

Training Data scoring - Summary Report

Cut off Prob.Val. for Success (Updatable)	0.5	(Updating the value here will NOT update value in detailed report)

Classification Confusion Matrix		
	Predicted Class	
Actual Class	NOT INSURED	INSURED
NOT INSURED	1291	267
INSURED	1541	6901

Error Report			
Class	# Cases	# Errors	% Error
NOT INSURED	1558	267	17.14
INSURED	8442	1541	18.25
Overall	10000	1808	18.08