Income Classification using Census Data

Statistical Analysis of income data using SAS

Abstract

In many real life instances we come across need to find income level of a person given set of variables. This is useful in policy design, to make targeted policy for certain class of people who is predicted to be below certain level of INCOME. Also useful in insurance industry to determine affordable rate of insurance if INCOME level can be determined. In this project we have tried to classify the income level using reference income of $50000. The data was found by census Bureau's current population survey, in which many different variables at various levels were recorded along with personal yearly income. We initially use chi-sq statistics to find best predictor to classify the income level, which later on was replaced by automatic stepwise selection algorithm. The logistic regression was use to classify the income level, since outcome was binary i.e. whether person is making 50000 or not. Many different model was trained on training data set which and the one which gave highest correction with good sensitivity/specificity was selected. This model was run against test data set. The result were significant the prediction accuracy of 88% was noted on both data sets with True Positive and True Negative rate was 81 and 88% respectively.

Background and Purpose of Study

The Current Population Survey (CPS), which is the Census Bureau's primary source of annual income data, collects and publishes information on household annual income based on various demographic and employment related variables. This information provides insight into income level for different categories and means for comparing them against median national income. The purpose of this study is to examine the data from survey in order to classify the income level in two categories i.e. person is making more or less than $50000 during the period 1994 and 1995.

Research Objective

The objective of this study is to identify the important factors which affect household/personal yearly income and use them to correctly classify the income level. The income level was chosen arbitrarily and kept at $50000. So goal is to classify and predict if a person is making $50000 or less based on various factors present in census data. Logistic regression was use to analyze the data with binary outcome. The analysis was carried out in SAS 9.2.

Sample Collection and Information

This data set contains weighted census data extracted from the 1994 and 1995 current population surveys conducted by the U.S. Census Bureau. The data was extracted from Census Bureau data extraction system*. The data contains 41 demographic and employment related variables. There are 299,285 records in data set and this was split into train/test in approximately 2/3, 1/3 proportions using MineSet's MIndUtil mineset-to-mlc.

Basic Statistics of data set:
Number of instances in data = 199523
Duplicate instances: 3723
Number of instances in test = 99762
Duplicate instances: 883
Number of attributes = 40 (continuous: 7 nominal: 33)

Incomes have been binned at the $50K level and it was drawn from the "total person income" field rather than the "adjusted gross income". This is our response variable.

All the variables which are part of the survey will be used to as predictor to classify the income level except “instance weight”. The instance weight indicates the number of people in the population that each record represents due to stratified sampling.

List of Predictor and Response Variables

Name	**Type**
age	continuous
class of worker	nominal
detailed industry recode	nominal
detailed occupation recode	nominal
education	nominal
wage per hour	continuous
enroll in edu inst last wk	nominal
marital stat	nominal
major industry code	nominal
major occupation code	nominal
race	nominal
hispanic origin	nominal
sex	nominal
member of a labor union	nominal
reason for unemployment	nominal
full or part time employment stat	nominal
capital gains	continuous
capital losses	continuous
dividends from stocks	continuous
tax filer stat	nominal
region of previous residence	nominal
state of previous residence	nominal
detailed household and family stat	nominal
detailed household summary in household	nominal
migration code-change in msa	nominal
migration code-change in reg	nominal
migration code-move within reg	nominal
live in this house 1 year ago	nominal
migration prev res in sunbelt	nominal
num persons worked for employer	continuous
family members under 18	nominal
country of birth father	nominal
country of birth mother	nominal
country of birth self	nominal
citizenship	nominal
own business or self employed	nominal
fill inc questionnaire for veteran's admin	nominal
veterans benefits	nominal
weeks worked in year	continuous
year	nominal

Response variable was coded using yearly income binned at +/-50000. Thus corresponding nominal response variable was 1 or 0.

All the nominal variables' categories are described in SAS code in appendix.

DATA CLEANUP AND MASSAGING

Duplicate records in the data was removed using excel. Also missing variables were coded as "?" which was not suitable for SAS so had to be renamed as "." This was done for all variable types (char and numeric). Since SAS considers "." as missing data for numeric variables and empty character for character variables, additional processing was done in SAS in DATA step where presence of "." was checked and replaced with empty character.

Also some of the variables contained lot of categories, this posed a quasi-complete separation** problem in model building. Thus to overcome this some of the variables' categories were collapsed. The education levels were collapsed from 17 categories to 12 categories, 1st to 9th grade were collapsed to one level. The class categories of workers were reduced by mapping government jobs into one category. The same procedure was use to map one or more category into one to reduced number of level for Hispanic origin and for major industry code. This was again done in DATA step before running any statistical procedures.

Since there are lot of observations, instead of replacing missing variables with mean/median they were discarded for final analysis.

Statistical Routines and Test Ran

Logistic regression was use to analyze the probability of making more or less than $50000 per year, given number of factors in the census data. Because logistic regression makes none of customary assumption regarding equal variance and normal distribution, there was no need to review here the distribution of variables.

Also to find most important factors associated with income level chi-square ranking was computed for all categorical (nominal) variables and top 15 was chosen to fit the data. In addition to data chi-square ranking for model selection automatic variable selection using stepwise procedure was used, this is to compare the models based on their classification strength.

The models were evaluated based on Hosmer-Lemeshow Goodness-of-fit test, % of correct classification and sensitivity/specificity analysis. In this paper we'll present only most significant model*.

SAMPLE DESCRIPTIVES

Mean, Median and Std deviation of all continuous variables by INCOME level presented below, followed by frequency table of significant categorical variables by INCOME level.

INCOME code (0/1)	N Obs	Variable	Label	Mean	Median	Std Dev	N
0	183912	AAGE	AGE	34.67	33.0	22.15	183912
		AHRSPAY	Wage per Hour	54.63	0	263.35	183912
		CAPGAIN	Capital Gain	146.37	0	1831.11	183912
		CAPLOSS	Capital Loss	27.48	0	231.51	183912
		DIVVAL	Dividend from Stocks	109.65	0	917.32	183912
		NOEMP	Number of employee/employer	1.85	0	2.32	183912
1	12382	AAGE	AGE	46.27	45.0	11.83	12382
		AHRSPAY	Wage per Hour	81.64	0	431.36	12382
		CAPGAIN	Capital Gain	4830.93	0	16887.63	12382
		CAPLOSS	Capital Loss	193.14	0	607.54	12382
		DIVVAL	Dividend from Stocks	1553.45	0	6998.07	12382
		NOEMP	Number of employee/employer	4.00	4.0	2.11	12382

As can be seen in above table, income level is not equally distributed, only 12382 instances for income level +50000 (level 1) where as 183912 instances for income level -50000 (level 0). Mean age for income level 0 was 35 compared to 46 for income level 1. Average number of employees per employer was around 2 for income level 0 compared to 4 for income level 1. Median wage per hour, capital gain/loss and dividend from stock was all 0 for either income level.

The next table shows distribution of INCOME level per predictor levels, only major predictor levels are shown here.

Data indicates, those in Sales and Executive Admin occupation are at INCOME higher income level. Same For MALE vs FEMALE, EDUCATION i.e. those who has higher education are at higher INCOME level then those who didn't finish school and those who has some college degree are more or less equally divided in INCOME level. Nonfiler (Tax filer) has lower INCOME level which is expected and those who filed jointly has higher INCOME level. Class of Worker shows higher INCOME level of state/local/federal government employees and self employed than their counterparts. INCOME level by race shows it is lower for American Indian/Eskimos and black. Same is true for those who are classified in industry code 0 52% are in INCOME level 0 vs 8.43% of those in Industry code 45 are on INCOME level 1. Also there is major difference in INCOME level between those who never married vs those who are married with spouse present, almost 45% of those who never married are in INCOME level 0 vs 77% of those who are married are on INCOME level 1. Member of labor union and veterans benefit also played some role on INCOME level.

Variable	LEVEL	INCOME 0	INCOME 1
AMJOCC(Major Occupation Code)	Executive admin and managerial	4.84	29.02
	Sales	5.58	12.31
ASEX(GENDER)	Female	54.23	21.51
	Male	45.77	78.49
AHGA (EDUCATION)	CHILD/1-10th GRADE EDUCATION	10.83	1.18
	Some college but no degree	14.15	14.43
	Masters degree	2.45	16.46
ADTOCC(Occupation Code)	0	52.50	7.32
	2	3.23	22.78
FILESTAT(Tax filer Status)	Nonfiler	39.08	0.28
	Joint both under 65	31.82	71.50
ACLSWKR(Class Of Worker)	Federal government	1.27	4.82
	Local/State Government	5.81	10.76
	Self-employed-incorporated	1.16	9.16
ARACE(Race)	Amer Indian Aleut or Eskimo	1.19	0.40
	Black	10.69	4.36
AMJIND(Major Industry Code)	Other professional services	1.87	8.43
	Manufacturing-durable goods	4.09	12.07
AHSCOL(Enrolled in Edu Inst.)	College or university	3.07	0.20
ADTIND(Industry Code)	0	52.50	7.32
	45	1.87	8.43

AMARITL(Martial Status)	Never married	44.68	9.02
	Married-civilian spouse present	40.56	77.53
VETYN(Veterans Benefits)	2	74.94	98.13
GRINST(State of Residence)	Not in universe	92.11	95.36
AUNMEM(Member of labor union)	No	7.73	14.71

INFERENTIAL DESCRIPTION

The analysis was carried out first on training set which contains 196K+ records after removing the duplicates. First chi-square score was computed using freq procedure in SAS and top 15 predictor variables were chosen to construct first model.

Table of chi-square values of first 15 variables, which are all significant.

ADTOCC	37527.4926
AHGA	29655.9582
AMJOCC	26152.9906
ADTIND	16659.9491
AMJIND	14964.6413
ACLSWKR	13131.6849
HHDFMX	11152.8494
FILESTAT	10118.1637
HHDREL	9937.1449
AMARITL	7385.8809
GENDER	4978.4895
AWKSTAT	4932.5603
PARENT	4773.3614
VETYN	3897.1777
SEOTR	1355.6678

The second model was constructed using stepwise selection procedure to identify the prognostic factors for INCOME level. Along with selection option, significance level of 0.3 and 0.35 was specified to allow the variable into the model and level required to keep variable in the model respectively. This was done to model the response using as many variables as possible.

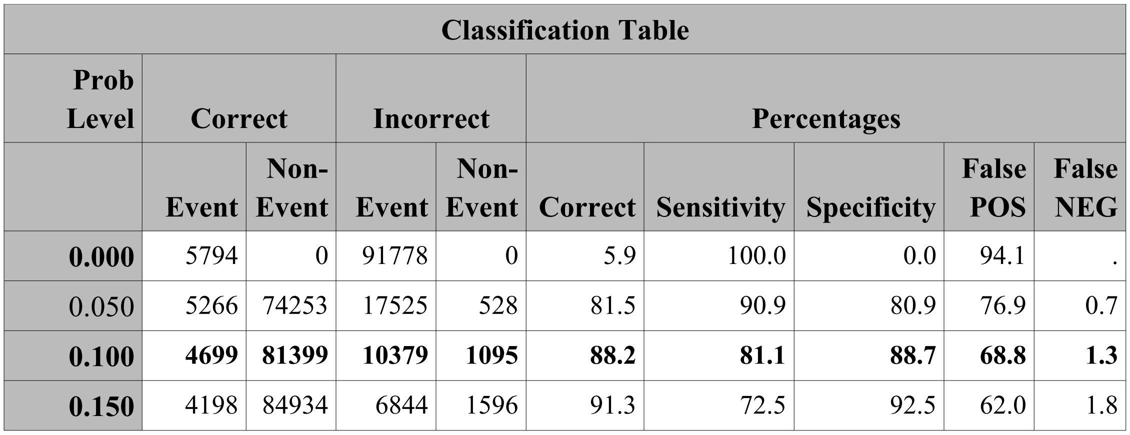
For both of these models detailed account of the variable selection process is requested by specifying the DETAILS option. The Hosmer and Lemeshow goodness-of-fit test for the final selected model is requested by specifying the LACKFIT option. The OUTEST options was use to get parameter estimates for the final selected model.

Summary of selected MODEL:

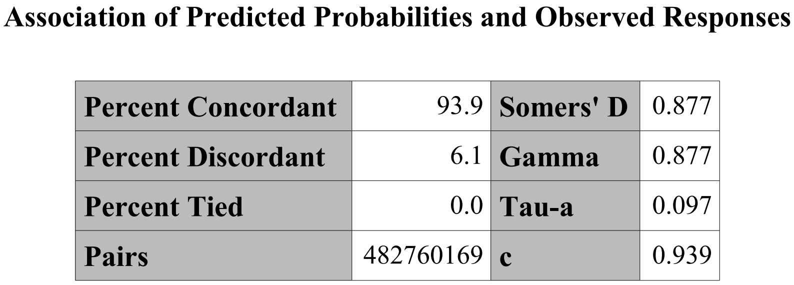
Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	**DF**	**Pr > ChiSq**
10.2423	8	0.2484

Group	**Total**	**INCOME = 1**		**INCOME = 0**	
		Observed	**Expected**	**Observed**	**Expected**
1	9797	0	0.08	9797	9796.92
2	9325	0	0.23	9325	9324.77
3	9358	1	0.62	9357	9357.38
4	9359	10	6.88	9349	9352.12
5	9358	25	27.48	9333	9330.52
6	9361	68	73.03	9293	9287.97
7	9359	165	180.00	9194	9179.00
8	9359	390	430.19	8969	8928.81
9	9359	1104	1056.26	8255	8302.74
10	8955	3716	3704.61	5239	5250.39

Hosmer and Lemeshow Goodness-of-Fit test shows model is significant at 0.05 level (p-value is 0.2484). Also there is not much difference between observed and expected frequency values for both INCOME level. The cut off probability was chosen from classification table by selecting highest % of correct, sensitivity and specificity percentages. As shown in below table, 0.1 probability level gives best values for above specified criteria. 88.2% correct classification of INCOME level with true positive rate (INCOME level 1) of 81.1% and true negative rate (INCOME level 0) of 88.7%.



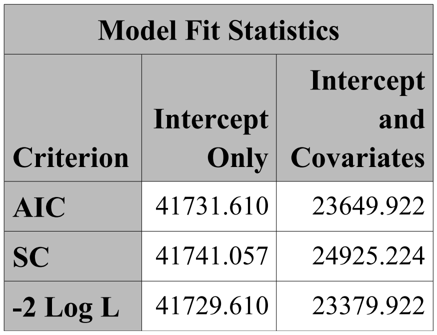
Classification Table									
Prob Level	Correct		Incorrect		Percentages				
	Event	Non-Event	Event	Non-Event	Correct	Sensitivity	Specificity	False POS	False NEG
0.000	5794	0	91778	0	5.9	100.0	0.0	94.1	.
0.050	5266	74253	17525	528	81.5	90.9	80.9	76.9	0.7
0.100	**4699**	**81399**	**10379**	**1095**	**88.2**	**81.1**	**88.7**	**68.8**	**1.3**
0.150	4198	84934	6844	1596	91.3	72.5	92.5	62.0	1.8



Association of Predicted Probabilities and Observed Responses

Percent Concordant	93.9	Somers' D	0.877
Percent Discordant	6.1	Gamma	0.877
Percent Tied	0.0	Tau-a	0.097
Pairs	482760169	c	0.939

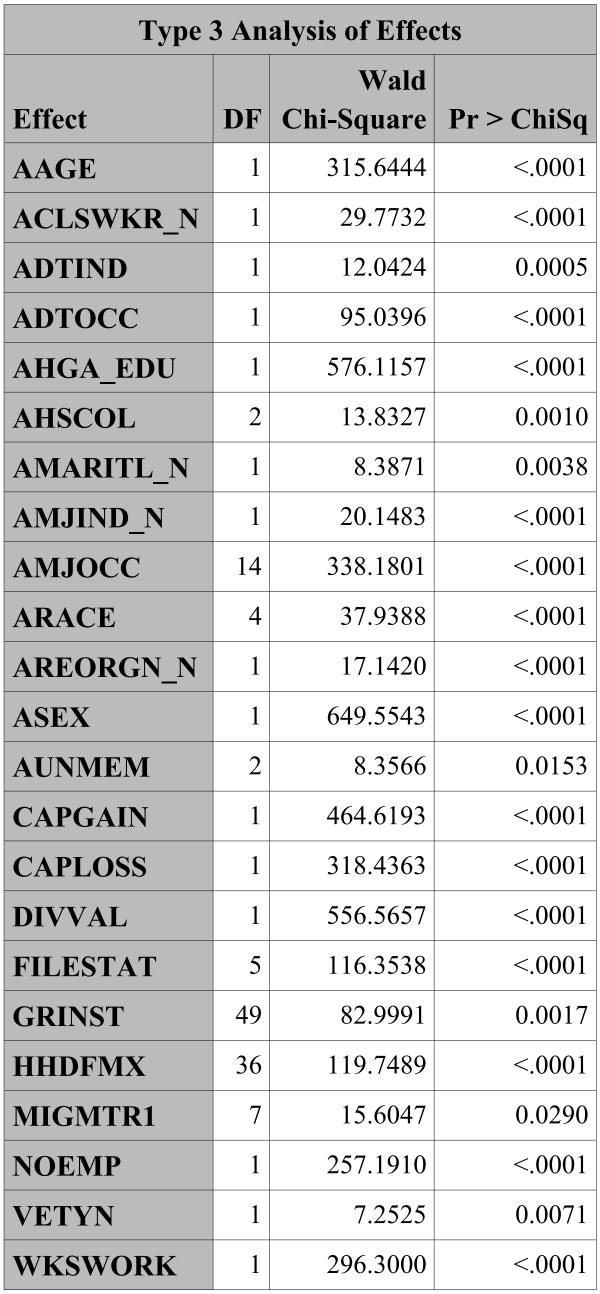
The percentage of concordant is also good at 93.9 which is much higher than 50.



Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	41731.610	23649.922
SC	41741.057	24925.224
-2 Log L	41729.610	23379.922

The model fit statistics shows with intercept only -2 LOG L value (with p-value<0.0001 and df=134) was pretty large but after adding co-variates value decrease sharply.

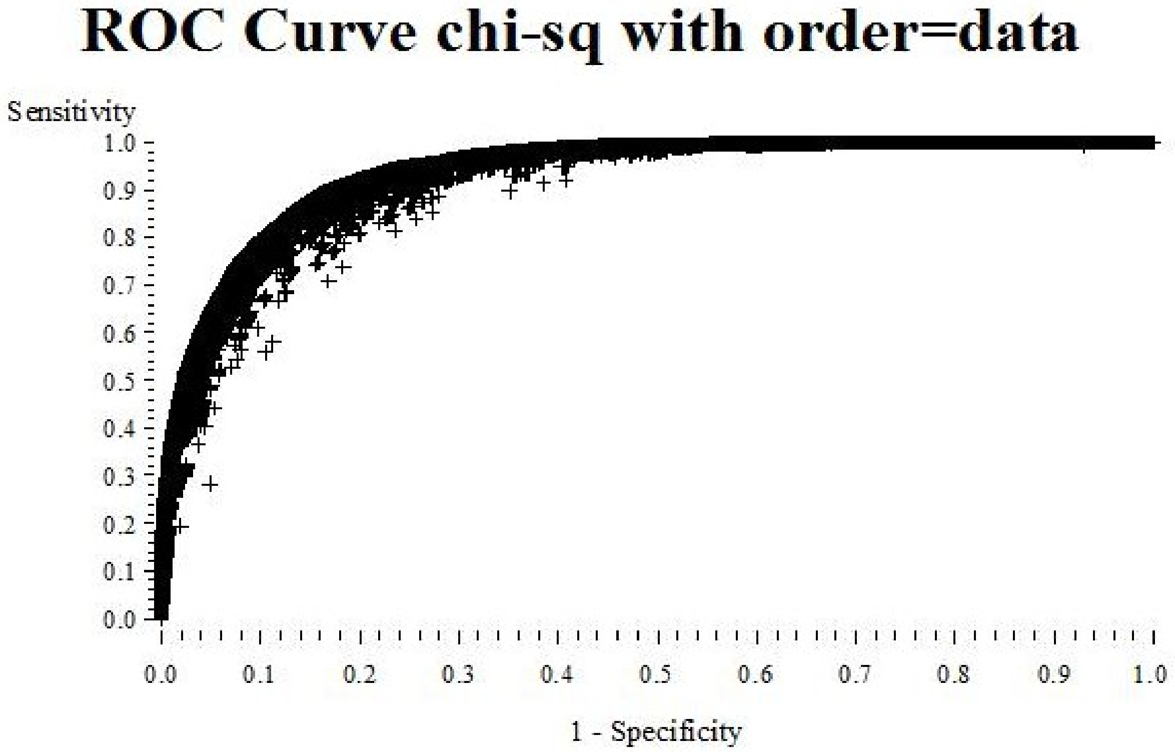
The variables selected by stepwise selection and its WALD chi-square value is:



Type 3 Analysis of Effects			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
AAGE	1	315.6444	<.0001
ACLSWKR_N	1	29.7732	<.0001
ADTIND	1	12.0424	0.0005
ADTOCC	1	95.0396	<.0001
AHGA_EDU	1	576.1157	<.0001
AHSCOL	2	13.8327	0.0010
AMARITL_N	1	8.3871	0.0038
AMJIND_N	1	20.1483	<.0001
AMJOCC	14	338.1801	<.0001
ARACE	4	37.9388	<.0001
AREORGN_N	1	17.1420	<.0001
ASEX	1	649.5543	<.0001
AUNMEM	2	8.3566	0.0153
CAPGAIN	1	464.6193	<.0001
CAPLOSS	1	318.4363	<.0001
DIVVAL	1	556.5657	<.0001
FILESTAT	5	116.3538	<.0001
GRINST	49	82.9991	0.0017
HHDFMX	36	119.7489	<.0001
MIGMTR1	7	15.6047	0.0290
NOEMP	1	257.1910	<.0001
VETYN	1	7.2525	0.0071
WKSWORK	1	296.3000	<.0001

* For detail model coefficients for all different levels please refer to addendum document.

The ROC curve shows more than 85% of area under the curve. Title is misleading, order =data is not specified in original model instead only DESCENDING option was use.



The chosen model was use to run on test data and model worked well on test data showing similar characteristics.

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
10.1942	8	0.2517

Classification Table									
Prob Level	Correct		Incorrect		Percentages				
	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.000	2840	0	46239	0	5.8	100.0	0.0	94.2	.
0.050	2571	37755	8484	269	82.2	90.5	81.7	76.7	0.7
0.100	**2307**	**41134**	**5105**	**533**	**88.5**	**81.2**	**89.0**	**68.9**	**1.3**

0.150	2080	42852	3387	760	91.6	73.2	92.7	62.0	1.7
0.200	1834	43858	2381	1006	93.1	64.6	94.9	56.5	2.2

Model was significant on test data and sensitivity-specificity and % correct also similar to training data which shows model shows similar behavior on test data as well.

Thus keeping cut-off probability level of 0.1 to classify the INCOME level of 1.

SUMMARY

We have investigated various variables and their different levels found in census data that had significant effect on INCOME level during 1995-1996. Using logistic regression we found that major occupation code, household and family stat, gender, tax file status, enrolled in educ inst last wk, migration code-change in msa, state of previous residence, member of a labor union, dividends from stocks, education, capital gain, weeks worked in year, capital loss, age, num persons worked for employer, occupation code, class of worker,Hispanic Origin, major industry code, industry code, marital status, veterans benefits. The model was used on test data which correctly classified 88.5% INCOME level with sensitivity/specificity was 81.2 and 89.0 respectively at 0.1 probability level.

Recommendation and Limitation

In order to do better analysis number of categorical values' level should be reduced to more manageable number of levels, so that model becomes easy to interpret. Also missing values should be factored into analysis by substituting it with mean/median or by combining some of the missing records with other instances.

Currently no interaction effects was considered and no discriminant analysis was done. The data contained lot of variables and all categorical variables had many different levels, so to find interaction effects among different categories between those variables log linear analysis should be done. Also to construct the best equation to classify INCOME levels discriminant analysis should also be done.

References

An Introduction to Logistic Regression Analysis and Reporting
CHAO-YING JOANNE PENG, KUK LIDA LEE, GARY M. INGERSOLL
Indiana University-Bloomington
http://sta559s11.pbworks.com/w/file/fetch/37766848/IntroLogisticRegressionPengEducResearch.pdf

Racing Committees for Large Datasets
Eibe Frank, Geoffrey Holmes and Richard Kirkby and Mark A. Hall.
Discovery Science. 2002
http://www.cs.waikato.ac.nz/pubs/wp/2002/uow-cs-wp-2002-03.pdf

Comparing Bayesian Network Classifiers
Jie Cheng, Russell Greiner
Department of Computing Science
University of Alberta

Quasi-complete separation and possible solution
http://www.ats.ucla.edu/stat/mult_pkg/faq/general/complete_separation_logit_models.htm

UCI Machine learning data repositary.
http://archive.ics.uci.edu/ml/datasets/Census-Income+%28KDD%29

Why are my logistic results reversed?
http://www.ats.ucla.edu/stat/sas/faq/logistic_descending.htm

APPENDIX:

SAS CODE:

DATA and FORMAT section

```
proc import datafile="C:\Users\parag\SAS_Project\Final_Project\census-
income.data.csv"
      replace
      out=census_train
      dbms=csv;
      getnames=yes;
      guessingrows=32767;
run;


PROC FORMAT;

   VALUE $ACLSWKR '1' = 'Not in universe'
                 '2' = 'Federal government'
                         '3' = 'Local government'
                         '4' = 'Never worked'
                         '5' = 'Private'
                         '6' = 'Self-employed-incorporated'
                         '7' = 'Self-employed-not incorporated'
                         '8' = 'State government'
                         '9' = 'Without pay';

   VALUE $AHGA     '1' = 'Children'
                 '2' = '7th and 8th grade'
                 '3' = '9th grade'
                 '4' = '10th grade'
                         '5' = 'High school graduate'
                         '6' = '11th grade'
                         '7'  = '12th grade no diploma'
                         '8' = '5th or 6th grade'
                         '9' = 'Less than 1st grade'
                         '10' = 'Bachelors degree(BA AB BS)'
                         '11' = '1st 2nd 3rd or 4th grade'
                         '12' = 'Some college but no degree'
                         '13' = 'Masters degree(MA MS MEng MEd MSW MBA)'
                         '14' = 'Associates degree-occup /vocational'
                         '15' = 'Associates degree-academic program'
                         '16' = 'Doctorate degree(PhD EdD)'
                         '17' = 'Prof school degree (MD DDS DVM LLB JD)';
   VALUE $AHSCOL  '1' = 'Not in universe'
                 '2' = 'High school'
                 '3' = 'College or university';

   VALUE $AMARITL '1' = 'Never married'
                  '2' = 'Married-civilian spouse present'
                  '3' = 'Married-spouse absent'
                  '4' = 'Separated'
                          '5' = 'Divorced'
                          '6' = 'Widowed'
                          '7' = 'Married-A F spouse present';

   VALUE $AMJIND   '1' = 'Not in universe or children'
                 '2' = 'Entertainment'
                 '3' = 'Social services'
                 '4' = 'Agriculture'
                         '5' = 'Education'
                         '6' = 'Public administration'
                         '7' = 'Manufacturing-durable goods'
                         '8' = 'Manufacturing-nondurable goods'
                         '9' = 'Wholesale trade'
                         '10' = 'Retail trade'
                         '11' = 'Finance insurance and real estate'
                         '12' = 'Private household services'
                         '13' = 'Business and repair services'
                         '14' = 'Personal services except private HH'
                         '16' = 'Construction'
                         '17' = 'Medical except hospital'
                         '18' = 'Other professional services'
                         '19' = 'Transportation'
                         '20' = 'Utilities and sanitary services'
                         '21' = 'Mining'
                         '22' = 'Communications'
                         '23' = 'Hospital services'
                         '24' = 'Forestry and fisheries'
                         '25' = 'Armed Forces';

VALUE $AMJOCC      '1' = 'Not in universe'
                 '2' = 'Professional specialty'
                 '3' = 'Other service'
                 '4' = 'Farming forestry and fishing'
                         '5' = 'Sales'
                         '6' = 'Adm support including clerical'
                         '7' = 'Protective services'
                         '8' = 'Handlers equip cleaners etc'
                         '9' = 'Precision production craft & repair'
                         '10' = 'Technicians and related support'
                         '11' = 'Machine operators assmblrs & inspctrs'
                         '12' = 'Transportation and material moving'
                         '13' = 'Executive admin and managerial'
                         '14' = 'Private household services'
                         '16' = 'Armed Forces';

VALUE $ARACE       '1' = 'White'
                 '2' = 'Black'
                 '3' = 'Other'
                 '4' = 'Amer Indian Aleut or Eskimo'
                         '5' = 'Asian or Pacific Islander';

VALUE $AREORGN     '1' = 'Mexican (Mexicano)'
                 '2' = 'Mexican-American'
                 '3' = 'Puerto Rican'
                 '4' = 'Central or South American'
                         '5' = 'All other'
                         '6' = 'Other Spanish'
                         '7' = 'Chicano'
                         '8' = 'Cuban'
                         '9' = 'Do not know'
                         '10' = 'NA';

VALUE $ASEX        '1' = 'Female'
                 '2' = 'Male';

VALUE $AUNMEM      '1' = 'Not in universe'
                 '2' = 'No'
                 '3' = 'Yes';

VALUE $AUNTYPE     '1' = 'Not in universe'
                 '2' = 'Re-entrant'
                 '3' = 'Job loser - on layoff'
                 '4' = 'New entrant'
                         '5' = 'Job leaver'
                         '6' = 'Other job loser';

VALUE $AWKSTAT     '1' = 'Children or Armed Forces'
                 '2' = 'Full-time schedules'
                 '3' = 'Unemployed part- time'
                 '4' = 'Not in labor force'
                         '5' = 'Unemployed full-time'
                         '6' = 'PT for non-econ reasons usually FT'
                         '7' = 'PT for econ reasons usually PT'
                         '8' = 'PT for econ reasons usually FT';

VALUE $FILESTAT  '1' = 'Nonfiler'
                 '2' = 'Joint one under 65 & one 65+'
                 '3' = 'Joint both under 65'
                 '4' = 'Single'
                         '5' = 'Head of household'
                         '6' = 'Joint both 65+';

VALUE $GRINREG     '1' = 'Not in universe'
                 '2' = 'South'
                 '3' = 'Northeast'
                 '4' = 'West'
                         '5' = 'Midwest'
                         '6' = 'Abroad';

VALUE $GRINST     '1'  = 'Not in universe'
                        '2'   = 'Utah'
                        '3'   = 'Michigan'
                        '4'   = 'North Carolina'
                        '5'   = 'North Dakota'
                        '6'   = 'Virginia'
                        '7'   = 'Vermont'
                        '8'   = 'Wyoming'
                        '9'   = 'West Virginia'
                        '10'   = 'Pennsylvania'
                        '11'   = 'Abroad'
                        '12'   = 'Oregon'
                        '13'   = 'California'
                        '14'   = 'Iowa'
                        '15'   = 'Florida'
                        '16'   = 'Arkansas'
                        '17'   = 'Texas'
                        '18'   = 'South Carolina'
                        '19'   = 'Arizona'
                        '20'   = 'Indiana'
                        '21'   = 'Tennessee'
                        '22'   = 'Maine'
                        '23'   = 'Alaska'
                        '24'   = 'Ohio'
                        '25'   = 'Montana'
                        '26'   = 'Nebraska'
                        '27'   = 'Mississippi'
                        '28'   = 'District of Columbia'
                        '29'   = 'Minnesota'
                        '30'   = 'Illinois'
                        '31'   = 'Kentucky'
                        '32'   = 'Delaware'
                        '33'   = 'Colorado'
                        '34'   = 'Maryland'
                        '35'   = 'Wisconsin'
                        '36'   = 'New Hampshire'
                        '37'   = 'Nevada'
                        '38'   = 'New York'
                        '39'   = 'Georgia'
                        '40'   = 'Oklahoma'
                        '41'  = 'New Mexico'
                        '42'  = 'South Dakota'
                        '43'  = 'Missouri'
                        '44'  = 'Kansas'
                        '45'  = 'Connecticut'
                        '46'  = 'Louisiana'
                        '47'  = 'Alabama'
                        '48'  = 'Massachusetts'
                        '49'  = 'Idaho'
                        '50'  = 'New Jersey';

VALUE $HHDFMX     '1' = 'Child <18 never marr not in subfamily'
                        '2' = 'Other Rel <18 never marr child of subfamily RP'
                        '3' = 'Other Rel <18 never marr not in subfamily'
                        '4' = 'Grandchild <18 never marr child of subfamily
RP'
                        '5' = 'Grandchild <18 never marr not in subfamily'
                        '6' = 'Secondary individual'
                        '7' = 'Child under 18 of RP of unrel subfamily'
                        '8' = 'RP of unrelated subfamily'
                        '9' = 'Spouse of householder'
                        '10' = 'Householder'
                        '11' = 'Other Rel <18 never married RP of subfamily'
                        '12' = 'Grandchild <18 never marr RP of subfamily'
                        '13' = 'Child <18 never marr RP of subfamily'
                        '14' = 'Child <18 ever marr not in subfamily'
                        '15' = 'Other Rel <18 ever marr RP of subfamily'
                        '16' = 'Child <18 ever marr RP of subfamily'
                        '17' = 'Nonfamily householder'
                        '18' = 'Child <18 spouse of subfamily RP'
                        '19' = 'Other Rel <18 spouse of subfamily RP'
                        '20' = 'Other Rel <18 ever marr not in subfamily'
                        '21' = 'Grandchild <18 ever marr not in subfamily'
                        '22' = 'Child 18+ never marr Not in a subfamily'
                        '23' = 'Grandchild 18+ never marr not in subfamily'
                        '24' = 'Child 18+ ever marr RP of subfamily'
                        '25' = 'Other Rel 18+ never marr not in subfamily'
                        '26' = 'Child 18+ never marr RP of subfamily'
                        '27' = 'Other Rel 18+ ever marr RP of subfamily'
                        '28' = 'Other Rel 18+ never marr RP of subfamily'
                        '29' = 'Other Rel 18+ spouse of subfamily RP'
                        '30' = 'Other Rel 18+ ever marr not in subfamily'
                        '31' = 'Child 18+ ever marr Not in a subfamily'
                        '32' = 'Grandchild 18+ ever marr not in subfamily'
                        '33' = 'Child 18+ spouse of subfamily RP'
                        '34' = 'Spouse of RP of unrelated subfamily'
                        '35' = 'Grandchild 18+ ever marr RP of subfamily'
                        '36' = 'Grandchild 18+ never marr RP of subfamily'
                        '37' = 'Grandchild 18+ spouse of subfamily RP'
                        '38' = 'In group quarters' ;

VALUE $HHDREL      '1' = 'Child under 18 never married'
                 '2' = 'Other relative of householder'
                 '3' = 'Nonrelative of householder'
                         '4' = 'Spouse of householder'
                 '5' = 'Householder'
                         '7' = 'Child under 18 ever married'
                         '8' = 'Group Quarters- Secondary individual'
                         '9' = 'Child 18 or older';

VALUE $MIGMTRa     '1' = 'Not in universe'
                 '2' = 'Nonmover'
                 '3' = 'MSA to MSA'
                         '4' = 'NonMSA to nonMSA'
                         '5' = 'MSA to nonMSA'
                         '6' = 'NonMSA to MSA'
                         '7' = 'Abroad to MSA'
                         '8' = 'Not identifiable'
                         '9' = 'Abroad to nonMSA';

VALUE $MIGMTRb     '1' = 'Not in universe'
                 '2' = 'Nonmover'
                 '3' = 'Same county'
                 '4' = 'Different county same state'
                         '5' = 'Different state same division'
                         '6' = 'Abroad'
                         '7' = 'Different region'
                         '8' = 'Different division same region';

VALUE $MIGMTRc     '1' = 'Not in universe'
                 '2' = 'Nonmover'
                 '3' = 'Same county'
                 '4' = 'Different county same state'
                         '5' = 'Different state in West'
                         '6' = 'Abroad'
                         '7' = 'Different state in Midwest'
                         '8' = 'Different state in South'
                         '9' = 'Different state in Northeast';

VALUE $MIGSAME     '1' = 'Not in universe under 1 year old'
                 '2' = 'Yes'
                 '3' = 'No';

VALUE $MIGSUN      '1' = 'Not in universe'
                 '2' = 'Yes'
                 '3' = 'No';

VALUE $PARENT      '1' = 'Not in universe'
```

```
                 '2' = 'Both parents present'
                 '3' = 'Neither parent present'
                 '4' = 'Mother only present'
                         '5' = 'Father only present';

VALUE $PEFNTVTY   '1' = 'Mexico'
                        '2' = 'United-States'
                        '3' = 'Puerto-Rico'
                        '4' = 'Dominican-Republic'
                        '5' = 'Jamaica'
                        '6' = 'Cuba'
                        '7' = 'Portugal'
                        '8' = 'Nicaragua'
                        '9' = 'Peru'
                        '10' = 'Ecuador'
                        '11' = 'Guatemala'
                        '12' = 'Philippines'
                        '13' = 'Canada'
                        '14' = 'Columbia'
                        '15' = 'El-Salvador'
                        '16' = 'Japan'
                        '17' = 'England'
                        '18' = 'Trinadad&Tobago'
                        '19' = 'Honduras'
                        '20' = 'Germany'
                        '21' = 'Taiwan'
                        '22' = 'Outlying-U S (Guam USVI etc)'
                        '23' = 'India'
                        '24' = 'Vietnam'
                        '25' = 'China'
                        '26' = 'Hong Kong'
                        '27' = 'Cambodia'
                        '28' = 'France'
                        '29' = 'Laos'
                        '30' = 'Haiti'
                        '31' = 'South Korea'
                        '32' = 'Iran'
                        '33' = 'Greece'
                        '34' = 'Italy'
                        '35' = 'Poland'
                        '36' = 'Thailand'
                        '37' = 'Yugoslavia'
                        '38' = 'Holand-Netherlands'
                        '39' = 'Ireland'
                        '40' = 'Scotland'
                        '41' = 'Hungary'
                        '42' = 'Panama';

VALUE $PEMNTVTY '1' = 'India'
```

```
                        '2' = 'Mexico'
                        '3' = 'United-States'
                        '4' = 'Puerto-Rico'
                        '5' = 'Dominican-Republic'
                        '6' = 'England'
                        '7' = 'Honduras'
                        '8' = 'Peru'
                        '9' = 'Guatemala'
                        '10' = 'Columbia'
                        '11' = 'El-Salvador'
                        '12' = 'Philippines'
                        '13' = 'France'
                        '14' = 'Ecuador'
                        '15' = 'Nicaragua'
                        '16' = 'Cuba'
                        '17' = 'Outlying-U S (Guam USVI etc)'
                        '18' = 'Jamaica'
                        '19' = 'South Korea'
                        '20' = 'China'
                        '21' = 'Germany'
                        '22' = 'Yugoslavia'
                        '23' = 'Canada'
                        '24' = 'Vietnam'
                        '25' = 'Japan'
                        '26' = 'Cambodia'
                        '27' = 'Ireland'
                        '28' = 'Laos'
                        '29' = 'Haiti'
                        '30' = 'Portugal'
                        '31' = 'Taiwan'
                        '32' = 'Holand-Netherlands'
                        '33' = 'Greece'
                        '34' = 'Italy'
                        '35' = 'Poland'
                        '36' = 'Thailand'
                        '37' = 'Trinadad&Tobago'
                        '38' = 'Hungary'
                        '39' = 'Panama'
                        '40' = 'Hong Kong'
                        '41' = 'Scotland'
                        '42' = 'Iran';

VALUE $PENATVTY '1' = 'United-States'
                        '2' = 'Mexico'
                        '3' = 'Puerto-Rico'
                        '4' = 'Peru'
                        '5' = 'Canada'
                        '6' = 'South Korea'
                        '7' = 'India'
                        '8' = 'Japan'
                        '9' = 'Haiti'
                        '10' = 'El-Salvador'
                        '11' = 'Dominican-Republic'
                        '12' = 'Portugal'
                        '13' = 'Columbia'
                        '14' = 'England'
                        '15' = 'Thailand'
                        '16' = 'Cuba'
                        '17' = 'Laos'
                        '18' = 'Panama'
                        '19' = 'China'
                        '20' = 'Germany'
                        '21' = 'Vietnam'
                        '22' = 'Italy'
                        '23' = 'Honduras'
                        '24' = 'Outlying-U S (Guam USVI etc)'
                        '25' = 'Hungary'
                        '26' = 'Philippines'
                        '27' = 'Poland'
                        '28' = 'Ecuador'
                        '29' = 'Iran'
                        '30' = 'Guatemala'
                        '31' = 'Holand-Netherlands'
                        '32' = 'Taiwan'
                        '33' = 'Nicaragua'
                        '34' = 'France'
                        '35' = 'Jamaica'
                        '36' = 'Scotland'
                        '37' = 'Yugoslavia'
                        '38' = 'Hong Kong'
                        '39' = 'Trinadad&Tobago'
                        '40' = 'Greece'
                        '41' = 'Cambodia'
                        '42' = 'Ireland';

VALUE $PRCITSHP '1' = 'Native- Born in the United States'
                        '2' = 'Foreign born- Not a citizen of U S '
                        '3' = 'Native- Born in Puerto Rico or U S Outlying'
                        '4' = 'Native- Born abroad of American Parent(s)'
                        '5' = 'Foreign born- U S citizen by naturalization';

VALUE $VETQVA     '1' = 'Not in universe'
                        '2' = 'YES'
                        '3' = 'NO';

VALUE $INCOME     1 = '>50000'
                        0 = '<50000';
```

```
RUN;

DATA census_train;
SET census_train;
RUN;
```

DATA CLEANUP CODE

```
                              DATA census_train;
      SET census_train;
      LABEL AAGE        = "AGE"
            ACLSWKR     = "Class of Worker"
            ADTIND      = "Industry Code"
            ADTOCC      = "Occupation Code"
            AHGA        = "Education"
            AHRSPAY     = "Wage per Hour"
            AHSCOL      = "Enrolled in Edu Inst."
            AMARITL     = "Martial Status"
            AMJIND      = "Major Industry Code"
            AMJOCC      = "Major Occupation Code"
            ARACE       = "Race"
            AREORGN     = "hispanic Origin"
            ASEX        = "Sex"
            AUNMEM      = "Member of labor union"
            AUNTYPE     = "Reason of unemployment"
            AWKSTAT     = "Full/Part time"
            CAPGAIN     = "Capital Gain"
            CAPLOSS     = "Capital Loss"
            DIVVAL      = "Dividend from Stocks"
            FILESTAT    = "Tax filer Status"
            GRINREG     = "Region of Residence"
            GRINST      = "State of Residence"
            HHDFMX      = "Household/family status"
            HHDREL      = "Household summary"
            MARSUPWT    = "Instance weight"
            MIGMTR1     = "Migration code-change (MSA)"
            MIGMTR3     = "Migration code-change (Region)"
            MIGMTR4     = "Migration code-move (Region)"
            MIGSAME     = "Lived in household last year"
            MIGSUN      = "Migration Prev. Res. in Sunbelt"
            NOEMP       = "Number of employee/employer"
            PARENT      = "Family Members Under 18"
            PEFNTVTY    = "Country of Birth Father"
            PEMNTVTY    = "Country of Birth Mother"
            PENATVTY    = "Country of Birth self"
            PRCITSHP    = "Citizenship"
            SEOTR       = "Owner/Self-Employed"
            VETQVA      = "Fill Questionnaire for Veteran's Admin"
            VETYN       = "Veterans Benefits"
```

```
             WKSWORK    = "Weeks Worked in Year"
             AGI        = "Adjusted Gross Income (-/+50000)"
             INCOME     = "INCOME code (0/1)";

IF (AGI eq '50000+.') THEN INCOME = 1;
ELSE IF (AGI eq '-50000') THEN INCOME = 0;
ELSE INCOME = .;



FORMAT      ACLSWKR     $ACLSWKR.
            AHGA        $AHGA.
            AHSCOL      $AHSCOL.
            AMARITL     $AMARITL.
            AMJIND      $AMJIND.
            AMJOCC      $AMJOCC.
            ARACE       $ARACE.
            AREORGN     $AREORGN.
            ASEX        $ASEX.
            AUNMEM      $AUNMEM.
            AUNTYPE     $AUNTYPE.
            AWKSTAT     $AWKSTAT.
            FILESTAT    $FILESTAT.
            GRINREG     $GRINREG.
            GRINST      $GRINST.
            HHDFMX      $HHDFMX.
            HHDREL      $HHDREL.
            MIGMTR1     $MIGMTRa.
            MIGMTR3     $MIGMTRb.
            MIGMTR4     $MIGMTRc.
            MIGSAME     $MIGSAME.
            MIGSUN      $MIGSUN.
            PARENT      $PARENT.
            PEFNTVTY    $PEFNTVTY.
            PEMNTVTY    $PEMNTVTY.
            PENATVTY    $PENATVTY.
            PRCITSHP    $PRCITSHP.
            VETQVA      $VETQVA.;

/*
IF AAGE eq . THEN AAGE = 35;
IF WKSWORK eq . THEN WKSWORK = 23;
IF ADTOCC eq . THEN ADTOCC = 12;
IF ADTIND eq . THEN ADTIND = 15;

IF AHRSPAY eq . THEN AHRSPAY = 56;
IF CAPGAIN eq . THEN CAPGAIN = 0;
IF CAPLOSS eq . THEN CAPLOSS = 0;
IF DIVVAL eq . THEN DIVVAL = 0;
IF NOEMP eq . THEN NOEMP = 1;
IF SEOTR eq . THEN SEOTR = 0;
IF VETYN eq . THEN VETYN = 1;
*/

IF STRIP (ACLSWKR) eq '.' THEN ACLSWKR='';
IF STRIP (AHSCOL) eq '.' THEN AHSCOL='';
IF STRIP (AMJOCC) eq '.' THEN AMJOCC='';
IF STRIP (AUNMEM) eq '.' THEN AUNMEM='';
IF STRIP (AUNTYPE) eq '.' THEN AUNTYPE='';
IF STRIP (GRINREG) eq '.' THEN GRINREG='';
IF STRIP (GRINST) eq '.' THEN GRINST='';
IF STRIP (MIGMTR1) eq '.' THEN MIGMTR1='';
IF STRIP (MIGMTR3) eq '.' THEN MIGMTR3='';
IF STRIP (MIGMTR4) eq '.' THEN MIGMTR4='';
IF STRIP (MIGSUN) eq '.' THEN MIGSUN='';
IF STRIP (PARENT) eq '.' THEN PARENT='';
IF STRIP (VETQVA) eq '.' THEN VETQVA='';
IF STRIP (AMJIND) eq '.' THEN AMJIND='';
IF STRIP (MIGSAME) eq '.' THEN MIGSAME='';

IF STRIP (ARACE) eq '.' THEN ARACE='';
IF STRIP (ASEX) eq '.' THEN ASEX='';
IF STRIP (AWKSTAT) eq '.' THEN AWKSTAT='';
IF STRIP (FILESTAT) eq '.' THEN FILESTAT='';
IF STRIP (HHDFMX) eq '.' THEN HHDFMX='';
IF STRIP (HHDREL) eq '.' THEN HHDREL='';
IF STRIP (PARENT) eq '.' THEN PARENT='';
IF STRIP (PEFNTVTY) eq '.' THEN PEFNTVTY='';
IF STRIP (PEMNTVTY) eq '.' THEN PEMNTVTY='';

IF STRIP (PENATVTY) eq '.' THEN PENATVTY='';
IF STRIP (PRCITSHP) eq '.' THEN PRCITSHP='';

IF AHGA eq '7th and 8th grade' OR AHGA eq '9th grade' OR AHGA eq '5th or 6th
grade' OR AHGA eq 'Less than 1st grade' OR AHGA eq '1st 2nd 3rd or 4th grade'
THEN AHGA_EDU = 2;
ELSE IF AHGA eq 'Children' THEN AHGA_EDU = 1;
ELSE IF AHGA eq '10th grade' OR AHGA eq '11th grade' THEN AHGA_EDU = 3;
ELSE IF AHGA eq '12th grade no diploma' THEN AHGA_EDU = 4;
ELSE IF AHGA eq 'High school graduate' THEN AHGA_EDU = 5;
ELSE IF AHGA eq 'Bachelors degree(BA AB BS)' THEN AHGA_EDU = 6;
ELSE IF AHGA eq 'Some college but no degree' THEN AHGA_EDU = 7;
ELSE IF AHGA eq 'Masters degree(MA MS MEng MEd MSW MBA)' THEN AHGA_EDU = 8;
ELSE IF AHGA eq 'Associates degree-occup /vocational' THEN AHGA_EDU = 9;
ELSE IF AHGA eq 'Associates degree-academic program' THEN AHGA_EDU = 10;
ELSE IF AHGA eq 'Doctorate degree(PhD EdD)' THEN AHGA_EDU = 11;
ELSE IF AHGA eq 'Prof school degree (MD DDS DVM LLB JD)' THEN AHGA_EDU = 12;
ELSE AHGA_EDU=.;
```

```
IF ACLSWKR eq 'Local government' OR ACLSWKR eq 'State government' THEN
ACLSWKR_N=3;
ELSE IF ACLSWKR eq 'Not in universe' THEN ACLSWKR_N=1;
ELSE IF ACLSWKR eq 'Federal government' THEN ACLSWKR_N=2;
ELSE IF ACLSWKR eq 'Never worked' THEN ACLSWKR_N=4;
ELSE IF ACLSWKR eq 'Private' THEN ACLSWKR_N=5;
ELSE IF ACLSWKR eq 'Self-employed-incorporated' THEN ACLSWKR_N=6;
ELSE IF ACLSWKR eq 'Self-employed-not incorporated' THEN ACLSWKR_N=7;
ELSE IF ACLSWKR eq 'Without pay' THEN ACLSWKR_N=8;
ELSE ACLSWKR_N=.;

IF AMARITL eq 'Married-A F spouse present' THEN AMARITL_N = 2;
ELSE IF AMARITL eq 'Never married' THEN AMARITL_N = 1;
ELSE IF AMARITL eq 'Married-civilian spouse present' THEN AMARITL_N = 2;
ELSE IF AMARITL eq 'Married-spouse absent' THEN AMARITL_N = 3;
ELSE IF AMARITL eq 'Separated' THEN AMARITL_N = 4;
ELSE IF AMARITL eq 'Divorced' THEN AMARITL_N = 5;
ELSE IF AMARITL eq 'Widowed' THEN AMARITL_N = 6;
ELSE AMARITL_N = .;

/*HISPANIC ORIGIN*/
IF AREORGN eq 'NA' OR AREORGN eq 'Do not know' THEN AREORGN_N=9;
ELSE IF AREORGN eq 'Mexican (Mexicano)' THEN AREORGN_N=1;
ELSE IF AREORGN eq 'Mexican-American' THEN AREORGN_N=2;
ELSE IF AREORGN eq 'Puerto Rican' THEN AREORGN_N=3;
ELSE IF AREORGN eq 'Central or South American' THEN AREORGN_N=4;
ELSE IF AREORGN eq 'All other' THEN AREORGN_N=5;
ELSE IF AREORGN eq 'Other Spanish' THEN AREORGN_N=6;
ELSE IF AREORGN eq 'Chicano' THEN AREORGN_N=7;
ELSE IF AREORGN eq 'Cuban' THEN AREORGN_N=8;
ELSE AREORGN_N=.;

IF AMJIND eq 'Not in universe or children' THEN AMJIND_N=1;
ELSE IF AMJIND eq 'Entertainment' THEN AMJIND_N=2;
ELSE IF AMJIND eq 'Social services' THEN AMJIND_N=3;
ELSE IF AMJIND eq 'Agriculture' THEN AMJIND_N=4;
ELSE IF AMJIND eq 'Education' THEN AMJIND_N=5;
ELSE IF AMJIND eq 'Public administration' THEN AMJIND_N=6;
ELSE IF AMJIND eq 'Manufacturing-durable goods' OR AMJIND eq 'Manufacturing-
nondurable goods' THEN AMJIND_N=7;
ELSE IF AMJIND eq 'Wholesale trade' OR AMJIND eq 'Retail trade' THEN
AMJIND_N=8;
ELSE IF AMJIND eq 'Finance insurance and real estate' THEN AMJIND_N=9;
ELSE IF AMJIND eq 'Private household services' THEN AMJIND_N=10;
ELSE IF AMJIND eq 'Business and repair services' THEN AMJIND_N=11;
ELSE IF AMJIND eq 'Personal services except private HH' THEN AMJIND_N=12;
ELSE IF AMJIND eq 'Construction' THEN AMJIND_N=13;
ELSE IF AMJIND eq 'Medical except hospital' THEN AMJIND_N=14;
ELSE IF AMJIND eq 'Other professional services' THEN AMJIND_N=15;
ELSE IF AMJIND eq 'Transportation' THEN AMJIND_N=16;
ELSE IF AMJIND eq 'Utilities and sanitary services' THEN AMJIND_N=17;
ELSE IF AMJIND eq 'Mining' THEN AMJIND_N=18;
ELSE IF AMJIND eq 'Communications' THEN AMJIND_N=19;
ELSE IF AMJIND eq 'Hospital services' THEN AMJIND_N=20;
ELSE IF AMJIND eq 'Mexican-American' THEN AMJIND_N=21;
ELSE IF AMJIND eq 'Forestry and fisheries' THEN AMJIND_N=22;
ELSE IF AMJIND eq 'Armed Forces' THEN AMJIND_N=23;
ELSE AMJIND_N=.;

RUN;
```

MEAN/FREQ CODE:

```
/*Find number of missing data for all variables*/
proc means data=census_train N nmiss;
run;

Ods graphics on;
Ods rtf file = 'census_mean_freq_1.rtf';
/*check means of continuos variables by income class*/
proc means data = census_train mean median stddev N maxdec=2;
      TITLE 'MEANS of continuos variables';
  var AAGE AHRSPAY CAPGAIN CAPLOSS DIVVAL NOEMP;
  class INCOME;
run;

PROC SORT DATA=census_train;
      BY INCOME;
RUN;

/*check chisq values of each major categorical variable by income*/
PROC FREQ data = census_train COMPRESS ORDER=FORMATTED;
      TITLE 'CHI-SQ of (major occupation code) AMJOCC';
      tables AMJOCC*INCOME /nocum norow nopercent chisq nofreq;
RUN;

PROC FREQ data = census_train COMPRESS ORDER=FORMATTED;
      TITLE 'CHI-SQ of GENDER';
      tables ASEX*INCOME /nocum norow nopercent chisq nofreq;
RUN;

PROC FREQ data = census_train COMPRESS ORDER=FORMATTED;
      TITLE 'CHI-SQ of Education';
      tables AHGA_EDU*INCOME /nocum norow nopercent chisq nofreq;
RUN;

PROC FREQ data = census_train COMPRESS ORDER=FORMATTED;
      TITLE 'CHI-SQ of Occupation code';
      tables ADTOCC*INCOME /nocum norow nopercent chisq nofreq;
RUN;


PROC FREQ data = census_train COMPRESS ORDER=FORMATTED;
      TITLE 'CHI-SQ of Tax File Status';
      tables FILESTAT*INCOME /nocum norow nopercent chisq nofreq ;
RUN;

PROC FREQ data = census_train COMPRESS ORDER=FORMATTED;
      TITLE 'CHI-SQ of worker class';
      tables ACLSWKR_N*INCOME /nocum norow nopercent chisq nofreq;
RUN;


PROC FREQ data = census_train COMPRESS ORDER=FORMATTED;
      TITLE 'CHI-SQ of RACE';
      tables ARACE*INCOME /nocum norow nopercent chisq nofreq ;
RUN;


PROC FREQ data = census_train COMPRESS ORDER=FORMATTED;
      TITLE 'CHI-SQ of Hispanic Origin';
      tables AREORGN*INCOME /nocum norow nopercent chisq nofreq ;
RUN;


PROC FREQ data = census_train COMPRESS ORDER=FORMATTED;
      TITLE 'CHI-SQ of AMJIND';
      tables AMJIND*INCOME /nocum norow nopercent chisq nofreq;
RUN;


PROC FREQ data = census_train COMPRESS ORDER=FORMATTED;
      TITLE 'CHI-SQ of AHSCOL';
      tables AHSCOL*INCOME /nocum norow nopercent chisq nofreq;
RUN;


PROC FREQ data = census_train COMPRESS ORDER=FORMATTED;
      TITLE 'CHI-SQ of Industry code';
      tables ADTIND*INCOME /nocum norow nopercent chisq nofreq;
RUN;


PROC FREQ data = census_train COMPRESS ORDER=FORMATTED;
      TITLE 'CHI-SQ of MIGMTR1';
      tables MIGMTR1*INCOME /nocum norow nopercent chisq nofreq ;
RUN;
```

```
PROC FREQ data = census_train COMPRESS ORDER=FORMATTED;
      TITLE 'CHI-SQ of Marital Status';
      tables AMARITL*INCOME /nocum norow nopercent chisq nofreq;
RUN;


PROC FREQ data = census_train COMPRESS ORDER=FORMATTED;
      TITLE 'CHI-SQ of Veteran benefit';
      tables VETYN*INCOME /nocum norow nopercent chisq nofreq ;
RUN;


PROC FREQ data = census_train COMPRESS ORDER=FORMATTED;
      TITLE 'CHI-SQ of (state of previous residence) GRINST';
      tables GRINST*INCOME /nocum norow nopercent chisq nofreq ;
RUN;


PROC FREQ data = census_train COMPRESS ORDER=FORMATTED;
      TITLE 'CHI-SQ of Union Member';
      tables AUNMEM*INCOME /nocum norow nopercent chisq nofreq ;
RUN;
```

LOGISTIC REGRESSSION/ROC CURVE CODE:

```
/*All char chi-sq ranking*/
PROC    LOGISTIC       data=    census_train_no_missing    OUTEST=estimates_chisq
DESCENDING;  /*this is the main initial model */
      TITLE 'Model selection using CHI_SQ';
      CLASS AMJOCC HHDFMX FILESTAT HHDREL ASEX AWKSTAT PARENT;
        MODEL   INCOME  =  ADTOCC  AHGA_EDU  AMJOCC  ADTIND  AMJIND_N   ACLSWKR_N
HHDFMX FILESTAT HHDREL AMARITL_N ASEX AWKSTAT PARENT VETYN SEOTR /
        MAXITER=200
Ctable pprob = ( 0 to 1 by 0.1)
Lackfit
details
Risklimits
OUTROC=ROC_chisq1;
Run;

/*All char MOST SIGNIFICANT KEPT THIS AND RUN ON TEST DATA*/
PROC LOGISTIC   data= census_train OUTEST=estimates_chisq DESCENDING;   /*this
is the main initial model */
      TITLE 'Model selection using STEPWISE';
      CLASS AHSCOL AMJOCC ARACE ASEX AUNMEM AUNTYPE AWKSTAT FILESTAT GRINREG
GRINST HHDFMX HHDREL MIGMTR1 MIGMTR3 MIGMTR4 MIGSAME MIGSUN PARENT PEFNTVTY
PEMNTVTY PENATVTY PRCITSHP VETQVA;
      MODEL  INCOME = AAGE ACLSWKR_N ADTIND ADTOCC    AHGA_EDU  AHRSPAY  AHSCOL
      AMARITL_N AMJIND_N      AMJOCC ARACE AREORGN_N ASEX    AUNMEM      AUNTYPE
```

```
AWKSTAT   CAPGAIN   CAPLOSS   DIVVAL   FILESTAT   GRINREG   GRINST   HHDFMX   HHDREL
      MIGMTR1  MIGMTR3  MIGMTR4  MIGSAME  MIGSUN  NOEMP  PARENT  PEFNTVTY  PEMNTVTY
PENATVTY PRCITSHP SEOTR VETQVA VETYN WKSWORK /
        MAXITER=200
        SELECTION = STEPWISE
        slentry=0.10
        slstay=0.15
Ctable pprob = ( 0 to 0.2 by 0.05)
Lackfit
details
Risklimits
OUTROC=ROC_chisq;
Run;

PROC LOGISTIC  data= census_train OUTEST=estimates_chisq DESCENDING;  /*this
is the main initial model */
      TITLE 'Model selection using STEPWISE';
      CLASS AHSCOL AMJOCC ARACE ASEX AUNMEM AUNTYPE AWKSTAT FILESTAT GRINREG
GRINST HHDFMX HHDREL MIGMTR1 MIGMTR3 MIGMTR4 MIGSAME MIGSUN PARENT PEFNTVTY
PEMNTVTY PENATVTY PRCITSHP VETQVA;
      MODEL  INCOME = AAGE ACLSWKR_N ADTIND ADTOCC    AHGA_EDU  AHRSPAY  AHSCOL
      AMARITL_N AMJIND_N      AMJOCC ARACE AREORGN_N ASEX    AUNMEM      AUNTYPE
AWKSTAT   CAPGAIN   CAPLOSS   DIVVAL   FILESTAT   GRINREG   GRINST   HHDFMX   HHDREL
      MIGMTR1  MIGMTR3  MIGMTR4  MIGSAME  MIGSUN  NOEMP  PARENT  PEFNTVTY  PEMNTVTY
PENATVTY PRCITSHP SEOTR VETQVA VETYN WKSWORK /
        MAXITER=200
        SELECTION = FORWARD
        slentry=0.10
        slstay=0.15
Ctable pprob = ( 0 to 1 by 0.05)
Lackfit
details
Risklimits
OUTROC=ROC_chisq;
Run;

PROC GPLOT DATA=ROC_chisq;
   TITLE 'ROC Curve chi-sq with order=data';
   PLOT _SENSIT_ * _1MSPEC_ = 'o';
   LABEL _SENSIT_ = 'Sensitivity'
         _1MSPEC_ = '1 - Specificity';
RUN;
```

FINAL MODEL ON TEST SET

```
PROC LOGISTIC  data= census_test OUTEST=estimates_chisq DESCENDING;  /*this 
is the final initial model */
      TITLE 'Model on test data';
        CLASS AMJOCC HHDFMX ASEX FILESTAT ARACE AHSCOL MIGMTR1 GRINST AUNMEM;
      MODEL  INCOME = AMJOCC HHDFMX ASEX FILESTAT ARACE AHSCOL MIGMTR1 GRINST 
AUNMEM DIVVAL AHGA_EDU CAPGAIN WKSWORK CAPLOSS AAGE NOEMP ADTOCC ACLSWKR_N 
AREORGN_N AMJIND_N ADTIND AMARITL_N VETYN /
        MAXITER=200
Ctable pprob = ( 0 to 0.2 by 0.05)
Lackfit
details
Risklimits
OUTROC=ROC_chisq_test;
Run;

PROC GPLOT DATA=ROC_chisq_test;
   TITLE 'ROC Curve chi-sq with order=data';
   PLOT _SENSIT_ * _1MSPEC_ = 'o';
   LABEL _SENSIT_ = 'Sensitivity'
         _1MSPEC_ = '1 - Specificity';
RUN;
```