Application of Data Mining for Census Income Prediction

Outline

- Background and Business Goal
- Data Resource & Preparation
- Model Creation, Exploration and Model Comparison
- Conclusions & Business Applications
- Future Plans

Background and Business Goal

- **Background**

  - Although the Census Bureau has been measuring income for a half-century, it was still not quite clear what contributes the inequality of individual gross income
  - Dataset used for this analysis is an extraction from the 1994 US census data

- **Business goal**

  - Taking advantage of the US census data to predict whether individual’s gross income is greater than or less than $50K (US median income)
  - Which factors are most decisive for determining individual’s gross income?
  - How many people remain in poverty over time that might need financial assistance?

CART Vs. Logistic Regression

- Both can perform on regression
- Logistic regression cannot use missing values in predictors, but decision tree can put missing values into a group
- Logistic Outcome is categorical, result is expressed as a probability of being in either group
- CART can reduce the impact as outliers can form its own group or merge with other values

Variable Selection

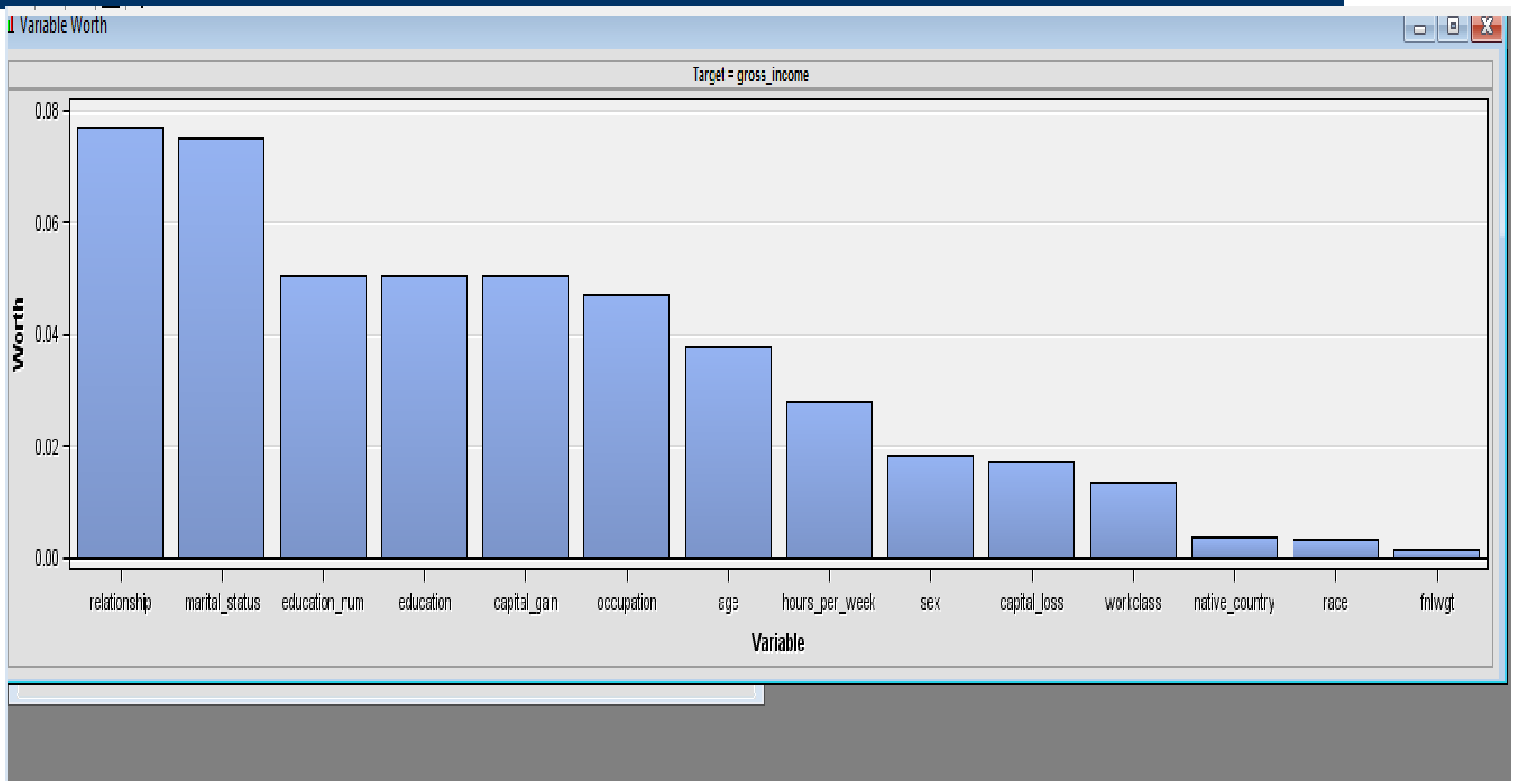
Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Informat	Label
1	age	Num	8	BEST12.	BEST32.	
11	capital_gain	Num	8			capital.gain
12	capital_loss	Num	8			capital.loss
4	education	Char	12	$12.	$12.	
9	education_num	Num	8			education.num
3	fnlwgt	Num	8	BEST12.	BEST32.	
15	gross_income	Num	8			gross.income
13	hours_per_week	Num	8			hours.per.week
10	marital_status	Char	21			marital.status
14	native_country	Char	18			native.country
5	occupation	Char	17	$17.	$17.	
7	race	Char	18	$18.	$18.	
6	relationship	Char	14	$14.	$14.	
8	sex	Char	6	$6.	$6.	
2	workclass	Char	16	$16.	$16.	

Data Preparation

	Variable Name	Type
Target	Gross Income	Binary
Predicators	Age, fnlwgt, education-num, capital-gain, capital-loss, hours-per-week,	Interval
Predicators	Workclass, education, marital-status, occupation, relationship, race, sex, native-country	Nominal

Variables Exploration



Model Creation

- Partition Data

- Two different kinds of models
  - Cart Tree
    - Use input once
    - Use input more than once
  - Logistic Regression
    - Stepwise

Data Partition

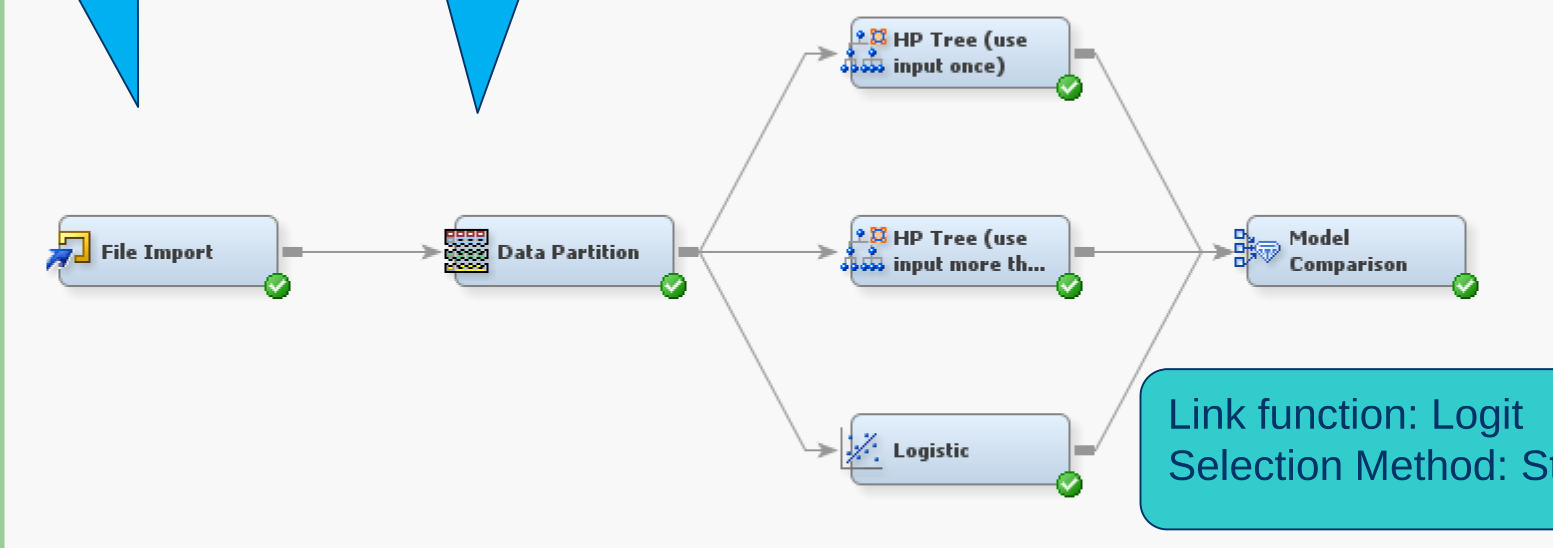
Property	Value
General	
Node ID	Part
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Output Type	Data
Partitioning Method	Default
Random Seed	12345
Data Set Allocations	
Training	70.0
Validation	30.0
Test	0.0
Report	
Interval Targets	No
Class Targets	Yes
Status	
Create Time	11/24/15 11:15 PM
Run ID	eb89a5e0-a69c-074c-a5cc-26b9fc57d21d
Last Error	
Last Status	Complete
Last Run Time	11/24/15 11:19 PM

Model Exploration

Total 48,842 observations

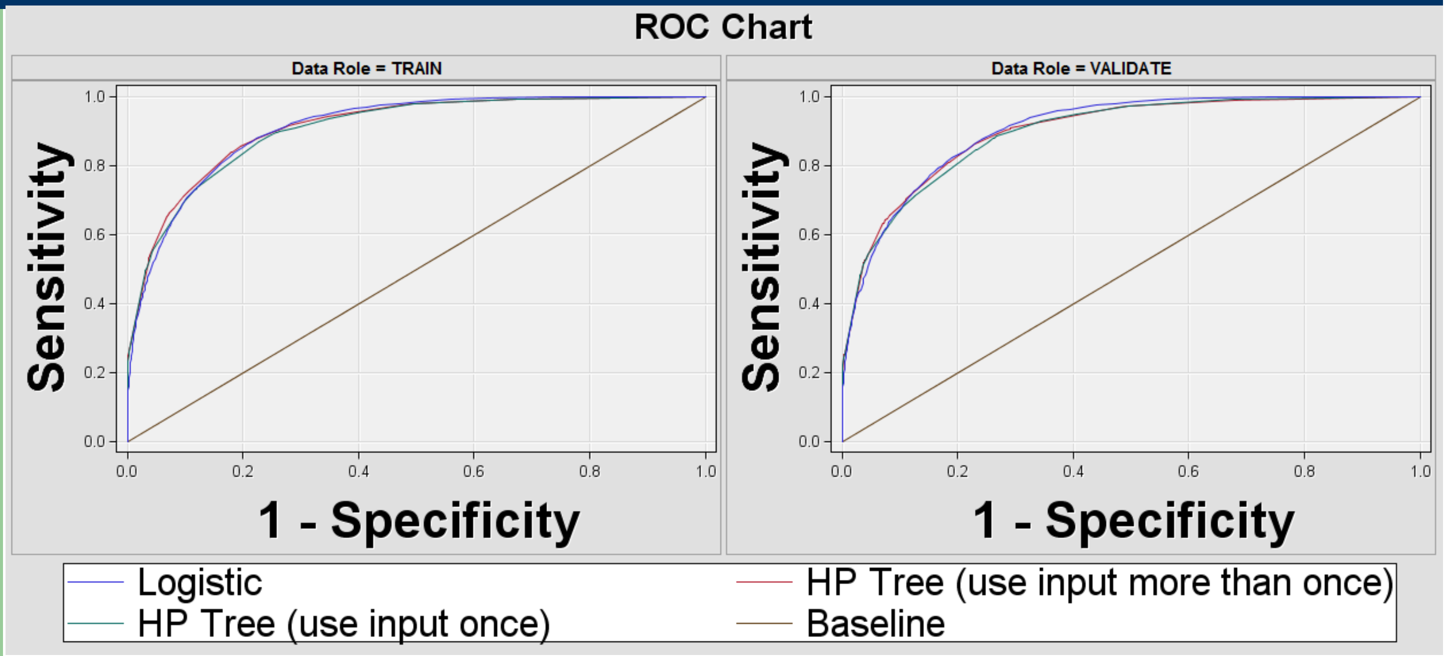
Training: 70%
Validation: 30%
Partitioning method: default

Nominal Target Criterion: Entropy
Maximum Depth: 10
HP tree1: Use input once
HP tree2: Use input more than once



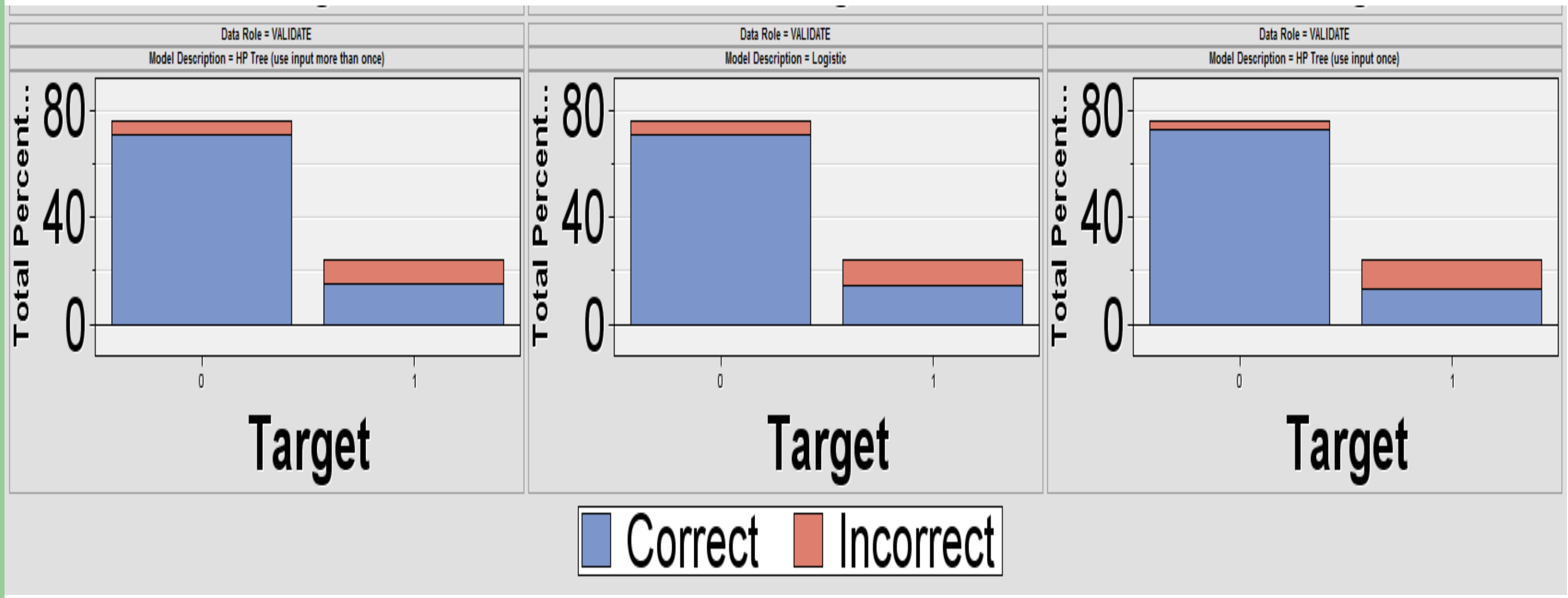
Link function: Logit
Selection Method: Stepwise

Model Comparison Result



	HP tree (use input more than once)	HP tree (use input once)	Logistic regression
Train: ROC index	0.91	0.904	0.909
Validate: ROC index	0.899	0.895	0.904

Model Comparison Result (cont'd)



Validate	HP tree (use input more than once)	Logistic regression	HP tree (use input once)
Misclassification rate	**0.14276**	**0.14876**	**0.14351**

Model Comparison Result (cont'd)

Fit Statistics

Valid: Average Squared Error	Valid: Divisor for ASE	Valid: Maximum Absolute Error	Valid: Sum of Frequencies	Valid: Root Average Squared Error	Valid: Sum of Squared Errors	Valid: Frequency of Classified Cases	Valid: Misclassification Rate	Valid: Number of Wrong Classifications
0.101353	29308	1	14654	0.31836	2970.464	14654	0.14276	2092
0.103335	29308	1	14654	0.321458	3028.541	14654	0.14351	2103
0.103162	29308	0.999998	14654	0.321189	3023.483	14654	0.148765	2180

The model (HP tree (use input more than once)) has the lowest average squared error, root average squared error, sum of squared errors, misclassification rates and number of wrong classifications.

Model Comparison Summary

Best Model for Prediction

MODEL	Valid: Misclassification Rate	Valid: Root Average Squared Error	Valid: Area Under Roc
HPTree (use input more than once)	0.14276	0.31836	0.899
HPTree (use input once)	0.14351	0.321458	0.895
Logtistic Regression	0.14876	0.321189	0.904

Key Findings

Top 4 most important variables:

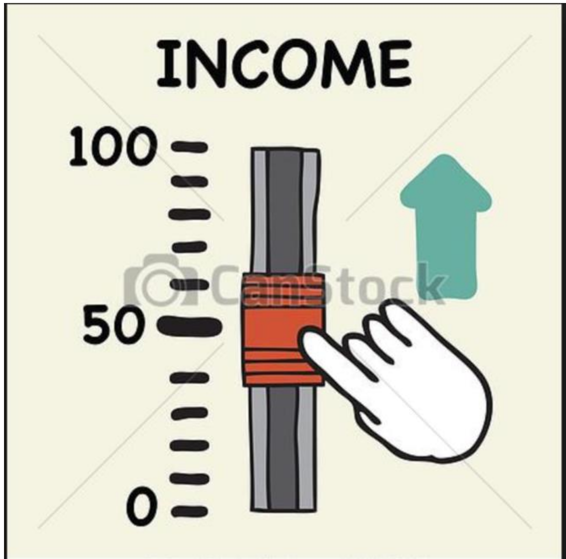
- Relationship
- Educatoin_num
- Capital_gain
- Occupation

NAME	COUNT	SSE	Import	VSSE	VImport
relationship	3	50.51376	1	32.56155	1
education_num	6	33.93573	0.671812	22.41317	0.688332
capital_gain	3	32.85254	0.650368	20.73361	0.636751
occupation	7	18.72211	0.370634	12.10741	0.371831
capital_loss	13	17.54073	0.347247	10.743	0.329929
age	16	15.74401	0.311678	8.770881	0.269363
hours_per_week	7	11.14428	0.220619	6.21724	0.190938
workclass	2	4.669095	0.092432	3.10314	0.095301
sex	1	1.868706	0.036994	1.456863	0.044742
marital_status	1	1.688369	0.033424	1.321665	0.04059
education	1	1.097155	0.02172	0.933503	0.028669
fnlwgt	1	1.851157	0.036647	0	0
native_country	0	0	0	0	0
race	0	0	0	0	0

Conclusions

- The “relationship” is most decisive to split the tree (Husband and Wife; Not-in-family, Other-relative, and own-child), however it is hard to tell if a person has a high income through only viewing “relationship” information

- The analysis confirmed (and quantified) what is considered common sense:
  - education_num, capital_gain, occupation are good for prediction (above a certain threshold)



Business Applications

- **From a social science perspective**
  - Understanding the characteristics that comprises a certain population can help identify the value of an area’s district level and deliver aid packages
  - The government and nonprofit organizations can provide different kinds of financial assistance, such as education, housing and medications for the low-income individuals

Our model can greatly benefit for finding out which government benefits that individuals may be eligible to receive and identifying specific target population for financial assistance

Future Plans

- **Simple K Means Clusters: try to find the people who would be likely to make an investment**

- **Association Rules: use it to find what kinds of characteristics he is very likely to have if a person earns more than 50K**

- **There are still limitations in our exploration process due to the attributes of dataset. In order to analyze the investment performance, more data, such as how much they invest should be included**

Appendix-Data Resources

- This data was extracted from UCI Machine Learning Repository database https://archive.ics.uci.edu/ml/datasets/Census+Income

	age	workclass	fnlwgt	education	education−num	marital−status	occupation	relationship	race	sex	capital−gain	capital−loss	hours−per−week	native−country	label
1	37	Private	182 675	Some-college	10	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	45	United-States	>50K
2	24	Private	333 505	HS-grad	9	Married-spouse-absent	Transport-moving	Own-child	White	Male	0	0	40	Peru	<=50K
3	45	State-gov	36 032	HS-grad	9	Divorced	Protective-serv	Unmarried	Black	Female	0	0	40	United-States	<=50K
4	30	Private	202 450	HS-grad	9	Married-civ-spouse	Transport-moving	Husband	White	Male	0	0	65	United-States	>50K
5	20	Private	194 630	HS-grad	9	Married-civ-spouse	Adm-clerical	Husband	White	Male	3781	0	50	United-States	<=50K
6	17	?	170 320	11th	7	Never-married	?	Own-child	White	Female	0	0	8	United-States	<=50K
7	39	Private	255 503	Bachelors	13	Never-married	Exec-managerial	Not-in-family	White	Male	0	0	40	United-States	>50K
8	40	Private	240 124	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	40	United-States	>50K
9	46	Private	321 327	Some-college	10	Married-civ-spouse	Transport-moving	Husband	White	Male	7298	0	45	United-States	>50K
10	18	Private	109 702	Some-college	10	Never-married	Sales	Own-child	White	Female	0	0	30	United-States	<=50K
11	28	Private	125 527	Some-college	10	Never-married	Sales	Not-in-family	White	Male	0	0	50	United-States	<=50K
12	20	Private	164 219	HS-grad	9	Never-married	Handlers-cleaners	Own-child	White	Male	0	0	45	United-States	<=50K
13	36	Private	32 334	Assoc-voc	11	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	50	United-States	>50K
14	63	?	257 659	Masters	14	Never-married	?	Not-in-family	White	Female	0	0	3	United-States	<=50K
15	19	Private	358 631	HS-grad	9	Never-married	Adm-clerical	Not-in-family	White	Male	0	0	25	United-States	<=50K
16	45	Private	329 603	Doctorate	16	Married-civ-spouse	Prof-specialty	Husband	White	Male	0	0	40	Poland	>50K
17	45	Private	101 320	HS-grad	9	Divorced	Adm-clerical	Unmarried	White	Female	0	0	40	United-States	<=50K
18	19	Private	307 496	Some-college	10	Never-married	Other-service	Own-child	White	Female	0	0	23	United-States	<=50K
19	25	Private	112 847	HS-grad	9	Married-civ-spouse	Transport-moving	Own-child	Other	Male	0	0	40	United-States	<=50K
20	27	Local-gov	162 404	HS-grad	9	Never-married	Protective-serv	Not-in-family	Black	Male	2174	0	40	United-States	<=50K

Appendix- Variables definition

- Gross income (<= 50k or >50k).
- age: continuous.
- workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov,
State-gov, Without-pay, Never-worked.
- fnlwgt: continuous.
- education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
- education-num: continuous.
- marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
- occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
- relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
- race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
- sex: Female, Male.
- capital-gain: continuous.
- capital-loss: continuous.
- hours-per-week: continuous.
- native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

Appendix-CART tree diagram (use input more than once)