FINAL REPORT – CENSUS

12/10/2011

IST 565 Data Mining

Chenzi Qian, Lingwen Zhang, Joshua Kitlas

FINAL REPORT – CENSUS

INTRODUCTION

The application problem we selected for the final report was to predict whether income exceeds $50K/yr based on census data.

Using a preprocessed dataset from the UC Irvine Machine Learning Repository, we set out to examine what attributes can contribute to creating a profile of wage earner either above or below $50K/year. The dataset contains 48,842 instances and has 14 different attributes. The characteristics of the attributes are both numeric and nominal. There are some missing values as this data was specifically extracted and organized for the purpose of experimentation. Throughout the process, we found ourselves discovering many other interesting profiles we did not intend to find.

The implications of our findings can have significant impact in many ways. Marketing insights can be gathered form this data, social constructs can be examined and predictive analysis of a population can be scrutinized.

From a marketing perspective, one area that we found to be particularly gripping may be of use to a financial services firm in their efforts to increase revenue. We found instances of female investors who did not fit the 'normal' profile of an investor. We identified that there were women who had little schooling (no degree other than a high school diploma), were unmarried and worked in an administrative role at an organization who have a history of making investments. Given stereotypes, this is not who is commonly identified as 'the normal' type of investor. This type of information could aid a financial services firm in creating products and targeting them specifically to this type of 'unique' investor.

From a social science perspective, understanding the characteristics that comprise a certain population (in this case individuals who make greater or less than $50K per annum) can help in determining a variety of both quantitative and qualitative studies. Characteristics can help create profiles for everything from identifying the value of an

area's school districts and delivering aid packages or identifying success characteristics of potential groups.

It is important to note that this was not the first data set we selected to use for this project. We had initially chosen a data set from the Heritage Provider Network that contained historical claims data. Despite spending several meetings working exclusively on the project, we could not come up with any compelling stories to tell about the data. After receiving guidance from Professor Yu, we chose to abandon this data set and search for a new one. There was a very important learning here for us – there is not always a way to make data useful. In this case, despite our efforts, there simply was not enough attributes to make a captivating story.

Contents

EXPERIMENT DESIGN AND RESULTS, ALGORITHMS AND TUNING PARAMETERS

Data subset 1

We discretized the age attribute to 10 bins.

Divide the data set into 2 sub data sets, one set does not have any investment (capital_gain = 0 and capital_loss = 0), the rest of the records are in the second sub data set.

SimpleKMeans

We are trying to find the people who would are likely to make an investment using SimpleKMeans. First, experiment with different k values, while keeping all other parameters at the default values. The following table shows that the higher the k value, the smaller the SSE will be. Therefore, as a measure of cluster cohesiveness, SSE has its own shortcomings.

k	SSE
3	129525
4	123028
5	122703
6	119060

Next, we set the k-means clustering to k=4 and displayStdDevs=true to obtain more details about the centroid description in the result. The running result is as following.

weka.clusterers.SimpleKMeans -V -N 4 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10

=== Model and evaluation on training set ===

kMeans

======

Number of iterations: 4

Within cluster sum of squared errors: 123028.0

Missing values globally replaced with mean/mode

Cluster centroids:

		Cluster#			
Attribute	Full Data	0	1	2	3
	(28330)	(9934)	(11280)	(3952)	(3164)
age	'(24.3-31.6]'	'(-inf-24.3]'	'(31.6-38.9]'	'(38.9-46.2]'	'(24.3-31.6]'
'(-inf-24.3]'	5308 (18%)	4130 (41%)	855 (7%)	37 (0%)	286 (9%)
'(24.3-31.6]'	5335 (18%)	1511 (15%)	1846 (16%)	419 (10%)	1559 (49%)
'(31.6-38.9]'	5233 (18%)	1059 (10%)	3216 (28%)	507 (12%)	451 (14%)
'(38.9-46.2]'	5112 (18%)	1339 (13%)	1902 (16%)	1508 (38%)	363 (11%)
'(46.2-53.5]'	3295 (11%)	797 (8%)	1520 (13%)	759 (19%)	219 (6%)
'(53.5-60.8]'	2171 (7%)	527 (5%)	1066 (9%)	430 (10%)	148 (4%)
'(60.8-68.1]'	1302 (4%)	380 (3%)	613 (5%)	220 (5%)	89 (2%)
'(68.1-75.4]'	377 (1%)	124 (1%)	178 (1%)	46 (1%)	29 (0%)
'(75.4-82.7]'	144 (0%)	47 (0%)	65 (0%)	18 (0%)	14 (0%)
'(82.7-inf)'	53 (0%)	20 (0%)	19 (0%)	8 (0%)	6 (0%)
workclass	Private	PrivatePrivatePrivatePrivate			
Self-emp-not-inc	2128 (7%)	338 (3%)	1224 (10%)	375 (9%)	191 (6%)
Private	19982 (70%)	7384 (74%)	8009 (71%)	2398 (60%)	2191 (69%)
State-gov	1134 (4%)	431 (4%)	321 (2%)	210 (5%)	172 (5%)
Federal-gov	808 (2%)	270 (2%)	253 (2%)	190 (4%)	95 (3%)
Local-gov	1772 (6%)	526 (5%)	594 (5%)	351 (8%)	301 (9%)
?	1655 (5%)	871 (8%)	525 (4%)	118 (2%)	141 (4%)
Self-emp-inc	832 (2%)	104 (1%)	345 (3%)	310 (7%)	73 (2%)
Without-pay	12 (0%)	4 (0%)	8 (0%)	0 (0%)	0 (0%)
Never-worked	7 (0%)	6 (0%)	1 (0%)	0 (0%)	0 (0%)
education	HS-grad	Some-college	HS-grad	Bachelors	Bachelors
Bachelors	4384 (15%)	248 (2%)	176 (1%)	2077 (52%)	1883 (59%)
HS-grad	9415 (33%)	2063 (20%)	7116 (63%)	45 (1%)	191 (6%)

11th	1089 (3%)	641 (6%)	332 (2%)	19 (0%)	97 (3%)
Masters	1300 (4%)	323 (3%)	259 (2%)	504 (12%)	214 (6%)
9th	474 (1%)	186 (1%)	210 (1%)	14 (0%)	64 (2%)
Some-college	6533 (23%)	4542 (45%)	1343 (11%)	632 (15%)	16 (0%)
Assoc-acdm	930 (3%)	389 (3%)	272 (2%)	130 (3%)	139 (4%)
Assoc-voc	1194 (4%)	454 (4%)	426 (3%)	166 (4%)	148 (4%)
7th-8th	582 (2%)	183 (1%)	297 (2%)	22 (0%)	80 (2%)
Doctorate	284 (1%)	42 (0%)	63 (0%)	132 (3%)	47 (1%)
5th-6th	308 (1%)	104 (1%)	156 (1%)	7 (0%)	41 (1%)
10th	865 (3%)	416 (4%)	322 (2%)	31 (0%)	96 (3%)
Prof-school	363 (1%)	38 (0%)	106 (0%)	153 (3%)	66 (2%)
1st-4th	159 (0%)	53 (0%)	78 (0%)	3 (0%)	25 (0%)
Preschool	47 (0%)	21 (0%)	18 (0%)	0 (0%)	8 (0%)
12th	403 (1%)	231 (2%)	106 (0%)	17 (0%)	49 (1%)

education-num	9	10	9	13	13
1	47 (0%)	21 (0%)	18 (0%)	0 (0%)	8 (0%)
2	159 (0%)	53 (0%)	78 (0%)	3 (0%)	25 (0%)
3	308 (1%)	104 (1%)	156 (1%)	7 (0%)	41 (1%)
4	582 (2%)	183 (1%)	297 (2%)	22 (0%)	80 (2%)
5	474 (1%)	186 (1%)	210 (1%)	14 (0%)	64 (2%)
6	865 (3%)	416 (4%)	322 (2%)	31 (0%)	96 (3%)
7	1089 (3%)	641 (6%)	332 (2%)	19 (0%)	97 (3%)
8	403 (1%)	231 (2%)	106 (0%)	17 (0%)	49 (1%)
9	9415 (33%)	2063 (20%)	7116 (63%)	45 (1%)	191 (6%)
10	6533 (23%)	4542 (45%)	1343 (11%)	632 (15%)	16 (0%)
11	1194 (4%)	454 (4%)	426 (3%)	166 (4%)	148 (4%)
12	930 (3%)	389 (3%)	272 (2%)	130 (3%)	139 (4%)
13	4384 (15%)	248 (2%)	176 (1%)	2077 (52%)	1883 (59%)
14	1300 (4%)	323 (3%)	259 (2%)	504 (12%)	214 (6%)
15	363 (1%)	38 (0%)	106 (0%)	153 (3%)	66 (2%)
16	284 (1%)	42 (0%)	63 (0%)	132 (3%)	47 (1%)

marital-status	Married-civ-spouse	Never-married	Married-civ-spouse	Married-civ-spouse	Never-married
Married-civ-spouse	12199 (43%)	602 (6%)	7916 (70%)	3591 (90%)	90 (2%)
Divorced	3990 (14%)	2089 (21%)	1202 (10%)	225 (5%)	474 (14%)
Married-spouse-absent	383 (1%)	167 (1%)	126 (1%)	16 (0%)	74 (2%)

Never-married	9914 (34%)	5998 (60%)	1506 (13%)	57 (1%)	2353 (74%)
Separated	944 (3%)	477 (4%)	324 (2%)	33 (0%)	110 (3%)
Married-AF-spouse	21 (0%)	8 (0%)	9 (0%)	4 (0%)	0 (0%)
Widowed	879 (3%)	593 (5%)	197 (1%)	26 (0%)	63 (1%)

occupation	Craft-repair	Adm-clerical	Craft-repair	Exec-managerial	Prof-specialty
Exec-managerial	3219 (11%)	742 (7%)	746 (6%)	1416 (35%)	315 (9%)
Handlers-cleaners	1274 (4%)	438 (4%)	684 (6%)	30 (0%)	122 (3%)
Prof-specialty	3290 (11%)	615 (6%)	595 (5%)	944 (23%)	1136 (35%)
Other-service	3122 (11%)	1779 (17%)	1027 (9%)	66 (1%)	250 (7%)
Adm-clerical	3408 (12%)	2438 (24%)	577 (5%)	192 (4%)	201 (6%)
Sales	3138 (11%)	1221 (12%)	1063 (9%)	525 (13%)	329 (10%)
Craft-repair	3593 (12%)	438 (4%)	2838 (25%)	161 (4%)	156 (4%)
Transport-moving	1416 (4%)	177 (1%)	1032 (9%)	97 (2%)	110 (3%)
Farming-fishing	890 (3%)	158 (1%)	577 (5%)	64 (1%)	91 (2%)
Machine-op-inspct	1806 (6%)	538 (5%)	1085 (9%)	72 (1%)	111 (3%)
Tech-support	795 (2%)	284 (2%)	212 (1%)	160 (4%)	139 (4%)
?	1662 (5%)	877 (8%)	526 (4%)	118 (2%)	141 (4%)
Protective-serv	570 (2%)	113 (1%)	293 (2%)	106 (2%)	58 (1%)
Armed-Forces	8 (0%)	3 (0%)	4 (0%)	0 (0%)	1 (0%)
Priv-house-serv	139 (0%)	113 (1%)	21 (0%)	1 (0%)	4 (0%)

relationship	Husband	Not-in-family	Husband	Husband	Not-in-family
Husband	10739 (37%)	114 (1%)	7329 (64%)	3258 (82%)	38 (1%)

Not-in-family	7427 (26%)	3926 (39%)	1125 (9%)	181 (4%)	2195 (69%)
Wife	1272 (4%)	439 (4%)	479 (4%)	314 (7%)	40 (1%)
Own-child	4810 (16%)	3177 (31%)	1056 (9%)	30 (0%)	547 (17%)
Unmarried	3172 (11%)	1836 (18%)	958 (8%)	147 (3%)	231 (7%)
Other-relative	910 (3%)	442 (4%)	333 (2%)	22 (0%)	113 (3%)

race	White	WhiteWhiteWhiteWhite			
White	24061 (84%)	8095 (81%)	9748 (86%)	3570 (90%)	2648 (83%)
Black	2839 (10%)	1346 (13%)	1012 (8%)	182 (4%)	299 (9%)
Asian-Pac-Islander	902 (3%)	277 (2%)	299 (2%)	172 (4%)	154 (4%)
Amer-Indian-Eskimo	280 (0%)	117 (1%)	124 (1%)	13 (0%)	26 (0%)
Other	248 (0%)	99 (0%)	97 (0%)	15 (0%)	37 (1%)

sex	Male	Female	Male	MaleMale	
Male	18551 (65%)	2893 (29%)	9951 (88%)	3501 (88%)	2206 (69%)
Female	9779 (34%)	7041 (70%)	1329 (11%)	451 (11%)	958 (30%)

capital-gain	0	0	0	0	0
0	28330 (100%)	9934 (100%)	11280 (100%)	3952 (100%)	3164 (100%)

capital-loss	0	0	0	0	0
0	28330 (100%)	9934 (100%)	11280 (100%)	3952 (100%)	3164 (100%)

hours-per-week	'(39.5-40.5]'	'(39.5-40.5]'	'(39.5-40.5]'	'(39.5-40.5]'	'(39.5-40.5]'
'(-inf-34.5]'	5141 (18%)	3112 (31%)	1298 (11%)	275 (6%)	456 (14%)
'(34.5-39.5]'	1943 (6%)	957 (9%)	583 (5%)	186 (4%)	217 (6%)
'(39.5-40.5]'	13438 (47%)	4375 (44%)	5958 (52%)	1590 (40%)	1515 (47%)
'(40.5-50.5]'	4869 (17%)	993 (9%)	2051 (18%)	1176 (29%)	649 (20%)
'(50.5-inf)'	2939 (10%)	497 (5%)	1390 (12%)	725 (18%)	327 (10%)

native-country	United-States	United-StatesUnited-StatesUnited-StatesUnited-States			
United-States	25320 (89%)	8949 (90%)	10087 (89%)	3560 (90%)	2724 (86%)
Cuba	85 (0%)	28 (0%)	31 (0%)	14 (0%)	12 (0%)
Jamaica	78 (0%)	40 (0%)	26 (0%)	4 (0%)	8 (0%)
India	79 (0%)	16 (0%)	24 (0%)	22 (0%)	17 (0%)
?	493 (1%)	130 (1%)	186 (1%)	85 (2%)	92 (2%)
Mexico	612 (2%)	198 (1%)	317 (2%)	15 (0%)	82 (2%)
South	68 (0%)	21 (0%)	30 (0%)	10 (0%)	7 (0%)
Puerto-Rico	103 (0%)	38 (0%)	41 (0%)	11 (0%)	13 (0%)
England	78 (0%)	25 (0%)	24 (0%)	17 (0%)	12 (0%)
Germany	117 (0%)	44 (0%)	31 (0%)	25 (0%)	17 (0%)

Iran	35 (0%)	8 (0%)	12 (0%)	10 (0%)	5 (0%)
Philippines	174 (0%)	49 (0%)	57 (0%)	40 (1%)	28 (0%)
Italy	65 (0%)	20 (0%)	31 (0%)	10 (0%)	4 (0%)
Poland	53 (0%)	14 (0%)	27 (0%)	8 (0%)	4 (0%)
Columbia	55 (0%)	26 (0%)	22 (0%)	3 (0%)	4 (0%)
Cambodia	14 (0%)	1 (0%)	9 (0%)	3 (0%)	1 (0%)
Thailand	18 (0%)	6 (0%)	4 (0%)	5 (0%)	3 (0%)
Canada	103 (0%)	32 (0%)	41 (0%)	17 (0%)	13 (0%)
Ecuador	25 (0%)	9 (0%)	13 (0%)	2 (0%)	1 (0%)
Laos	17 (0%)	8 (0%)	6 (0%)	1 (0%)	2 (0%)
Haiti	42 (0%)	18 (0%)	17 (0%)	3 (0%)	4 (0%)
Portugal	35 (0%)	9 (0%)	21 (0%)	2 (0%)	3 (0%)
Dominican-Republic	67 (0%)	33 (0%)	23 (0%)	1 (0%)	10 (0%)
El-Salvador	95 (0%)	44 (0%)	31 (0%)	6 (0%)	14 (0%)
France	26 (0%)	8 (0%)	6 (0%)	7 (0%)	5 (0%)
Taiwan	44 (0%)	6 (0%)	9 (0%)	13 (0%)	16 (0%)
Honduras	12 (0%)	6 (0%)	4 (0%)	0 (0%)	2 (0%)
Guatemala	60 (0%)	29 (0%)	20 (0%)	0 (0%)	11 (0%)
Japan	51 (0%)	13 (0%)	15 (0%)	15 (0%)	8 (0%)
Yugoslavia	15 (0%)	4 (0%)	5 (0%)	6 (0%)	0 (0%)
China	64 (0%)	9 (0%)	28 (0%)	15 (0%)	12 (0%)
Peru	29 (0%)	14 (0%)	13 (0%)	1 (0%)	1 (0%)
Outlying-US(Guam-USVI-etc)	14 (0%)	6 (0%)	4 (0%)	0 (0%)	4 (0%)
Scotland	11 (0%)	5 (0%)	2 (0%)	2 (0%)	2 (0%)
Trinadad&Tobago	17 (0%)	9 (0%)	7 (0%)	0 (0%)	1 (0%)
Greece	20 (0%)	5 (0%)	10 (0%)	4 (0%)	1 (0%)
Nicaragua	30 (0%)	15 (0%)	11 (0%)	1 (0%)	3 (0%)
Vietnam	57 (0%)	28 (0%)	16 (0%)	6 (0%)	7 (0%)
Hong	19 (0%)	4 (0%)	7 (0%)	5 (0%)	3 (0%)
Ireland	21 (0%)	5 (0%)	8 (0%)	2 (0%)	6 (0%)
Hungary	9 (0%)	2 (0%)	4 (0%)	1 (0%)	2 (0%)

income	<=50K	<=50K	<=50K	>50K	<=50K
<=50K	22939 (80%)	9614 (96%)	9591 (85%)	751 (19%)	2983 (94%)
>50K	5391 (19%)	320 (3%)	1689 (14%)	3201 (80%)	181 (5%)

Clustered Instances

0 9934 (35%)
1 11280 (40%)
2 3952 (14%)
3 3164 (11%)

The centroids can be better displayed as following:

Cluster centroids:

Attribute	Full Data	Cluster# 0	1	2	3
	(28330)	(9934)	(11280)	(3952)	(3164)
age	'(24.3-31.6]'	'(-inf-24.3]'	'(31.6-38.9]'	'(38.9-46.2]'	'(24.3-31.6]'
workclass	Private	Private	Private	Private	Private
education	HS-grad	Some-college	HS-grad	Bachelors	Bachelors
education-num	9	10	9	13	13
marital-status	Married-civ-spouse	Never-married	Married-civ-spouse	Married-civ-spouse	Never-married
occupation	Craft-repair	Adm-clerical	Craft-repair	Exec-managerial	Prof-specialty
relationship	Husband	Not-in-family	Husband	Husband	Not-in-family
race	White	White	White	White	White
sex	Male	Female	Male	Male	Male
capital-gain	0	0	0	0	0
capital-loss	0	0	0	0	0
hours-per-week	'(39.5-40.5]'	'(39.5-40.5]'	'(39.5-40.5]'	'(39.5-40.5]'	'(39.5-40.5]'
native-country	United-States	United-States	United-States	United-States	United-States
income	<=50K	<=50K	<=50K	>50K	<=50K

Description of the clusters

(1) The four clusters are in different sizes. The smallest is cluster #3, which includes 3164 instances while the largest is cluster #1, which includes 11280 instances.

(2) The centroid of Cluster#0 is characterized by the following attribute values: Most of the instances have younger than 25 age (41%), workclass are private (74%), education are some-college (45%), education-num are 10th grade (45%), marital-

status are never-married (60%), occupation are adm-clerical (24%), relationship are not-in-family (39%), race are white (81%), sex are female (70%), no investment, hours per week are 40 (44%), native-country are United States (90%) and their income are lower than 50K (96%).

The majority of members in this cluster are white, young, and female, who are not married and are doing full time administrator/clerical work. This group is less likely to make an investment.

(3) The centroid of Cluster#1 is characterized by the following attribute values: Most of the instances are age between 32 and 38 (28%), workclass are private (71%), education are high school graduate (63%), education-num are 9th grade (63%), marital-status are married-civ-spouse (70%), occupation are craft-repair (25%), relationship are husband (64%), race are white (86%), sex are male (88%), no investment, hours per week are 40 (52%), native-country are United States (89%) and their income are lower than 50K (85%).

The majority of members in this cluster are white middle-age males, who are married and are doing full time craft-repair work.This group is less likely to make an investment.

(4) The centroid of Cluster#2 is characterized by the following attribute values: Most of the instances are aged between 39 and 46 (38%), workclass are private (60%), education are bachelors (52%), education-num are 13th grade (52%), marital-status are married-civ-spouse (90%), occupation are exec-managerial (35%), relationship are husband (82%), race are white (90%), sex are male (88%), no investment, hours per week are 40 (40%), native-country are United States (90%) and their income are higher than 50K (80%).

The majority members in this cluster are white middle age male, who are married and doing full time exec-managerial work. This group is less likely to make an investment. Even their income are comparatively high, they still may not make aninvestment.

(5) The centroid of Cluster#3 is characterized by the following attribute values: Most of the instances are age between 25 and 32 (49%), workclass are private (69%), education are bachelors (59%), education-num are 13th grade (59%), marital-status are never-married (74%), occupation are prof-specialty (35%), relationship are not-in-family (69%), race are white (83%), sex are male (69%), no investment, hours per week are 40 (47%), native-country are United States (86%) and their income are lower than 50K (94%).

The majority of members in this cluster are white middle age males, who are married and doing full time prof-specialty work. This group is less likely to make an investment.

Data subset 2

(capital_gain is not 0 or capital_loss is not 0)

There are altogether 4231 instances, with 14 attributes in this data subset.

Preprocess:

1. "Discretize" age into 10 bins, and keep all the other settings as default.
2. Use "NumericToNominal" to the attribute 'education-num'.
3. "Discretize" capital_gain, capital_loss and hours-per-week into 5 bins, set 'useEqualFrequency' to true.

In this experiment, we keep all the settings the same as the first subset, rerun the process, and try to examine the characteristics of the clusters.

Decision tree

We are trying to identify people who are likely to have high income among people who make investments. The dataset we will use is the sub dataset 2. We choose to use J48 because this gives us a clear and direct result. Change the saveInstanceData to true and keep other parameter values as default. The running result is:

=== Run information ===

Scheme:weka.classifiers.trees.J48 -L -C 0.25 -M 2
Relation: census income_sub2-weka.filters.unsupervised.attribute.Remove-R3-weka.filters.unsupervised.attribute.Discretize-B10-M-1.0-R1-weka.filters.unsupervised.attribute.Discretize-F-B5-M-1.0-R10,11,12-weka.filters.unsupervised.attribute.NumericToNominal-R4
Instances:4231
Attributes:14
age
workclass
education
education-num
marital-status
occupation
relationship
race
sex
capital-gain
capital-loss
hours-per-week
native-country
income
Test mode:evaluate on training data

=== Classifier model (full training set) ===

J48 pruned tree

---

capital-gain = '(-inf-57]'
| marital-status = Married-civ-spouse
| | capital-loss = '(-inf-77.5]': >50K (0.0)
| | capital-loss = '(77.5-1670.5]': <=50K (112.0/31.0)
| | capital-loss = '(1670.5-1894.5]': >50K (293.0/79.0)
| | capital-loss = '(1894.5-1978.5]': >50K (370.0/13.0)
| | capital-loss = '(1978.5-inf)'
| | | education = Prof-school: >50K (17.0)
| | | education = HS-grad: <=50K (52.0/8.0)
| | | education = Bachelors: >50K (35.0/8.0)
| | | education = Some-college: <=50K (28.0/10.0)
| | | education = Assoc-acdm: <=50K (5.0)
| | | education = Doctorate: >50K (7.0/1.0)
| | | education = Masters: >50K (12.0/2.0)
| | | education = Assoc-voc: <=50K (9.0/3.0)
| | | education = 10th: <=50K (2.0)
| | | education = 9th: <=50K (0.0)
| | | education = Preschool: <=50K (0.0)
| | | education = 12th: <=50K (0.0)
| | | education = 11th: <=50K (6.0)
| | | education = 7th-8th: <=50K (4.0)
| | | education = 1st-4th: <=50K (2.0)
| | | education = 5th-6th: <=50K (4.0)
| marital-status = Never-married: <=50K (314.0/43.0)
| marital-status = Divorced: <=50K (165.0/29.0)
| marital-status = Separated: <=50K (29.0/6.0)
| marital-status = Married-spouse-absent: <=50K (14.0/3.0)
| marital-status = Widowed: <=50K (39.0/9.0)

```
|   marital-status =  Married-AF-spouse:  >50K (0.0)
capital-gain = '(57-3414.5]':  <=50K (689.0/90.0)
capital-gain = '(3414.5-7073.5]':  <=50K (624.0/208.0)
capital-gain = '(7073.5-12614]':  >50K (692.0/12.0)
capital-gain = '(12614-inf)':  >50K (707.0/8.0)

Number of Leaves  :         30

Size of the tree :     34


Time taken to build model: 0.06seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances        3668               86.6935 %
Incorrectly Classified Instances       563               13.3065 %
Kappa statistic                          0.7335
Mean absolute error                      0.2024
Root mean squared error                  0.3181
Relative absolute error                 41.5217 %
Root relative squared error             64.4378 %
Total Number of Instances             4231

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
               0.82      0.069      0.942      0.82      0.877      0.919     >50K
               0.931     0.18       0.79       0.931     0.855      0.919     <=50K
Weighted Avg.    0.867     0.116      0.878     0.867     0.868      0.919

=== Confusion Matrix ===
```

```
   a    b   <-- classified as
2010  440 |   a =  >50K
 123 1658 |   b =  <=50K
```

Visualize the tree and each node, we will find that:

(1) When we visualize the node capital-gain, we find that people whose capital gain is high are likely to have higher income. However, there are people who make investments but do not make any profit may have high income too.

(2) Visualize the node marital-status, we will find people who are never married or divorced are likely to have lower income.

(3) Visualize the node capital-loss, it is hard to tell if a person has a high income through only viewing capital-loss information.

(4) Visualize the node education, we can find that if people's education level is below high-school graduate, they are more likely to have lower income.

SimpleKMeans

Parameter: numClusters4, keep other settings default

Cluster mode: Use training set

Results:

```
Number of iterations: 4
Within cluster sum of squared errors: 22296.0
Missing values globally replaced with mean/mode

Cluster centroids:
                                                Cluster#
Attribute                 Full Data                 0                 1                 2                 3
                            (4231)                (1277)            (1132)            (1114)             (708)
=============================================================================================================
age                      '(38.9-46.2]'     '(38.9-46.2]'     '(31.6-38.9]'     '(38.9-46.2]'     '(-inf-24.3]'
workclass                     Private           Private           Private           Private           Private
education                     HS-grad           Masters           HS-grad           Bachelors         HS-grad
education-num                       9                14                 9                13                 9
marital-status       Married-civ-spouse  Married-civ-spouse  Married-civ-spouse  Married-civ-spouse  Never-married
occupation             Prof-specialty    Prof-specialty       Craft-repair      Prof-specialty      Adm-clerical
relationship                  Husband           Husband           Husband           Husband           Not-in-family
race                            White             White             White             White             White
sex                              Male              Male              Male              Male            Female
capital-gain              '(-inf-57]'       '(-inf-57]'       '(57-3414.5]'     '(12614-inf)'      '(-inf-57]'
capital-loss         '(-inf-77.5]'     '(-inf-77.5]'     '(-inf-77.5]'     '(-inf-77.5]'     '(-inf-77.5]'
hours-per-week       '(39.5-40.5]'     '(39.5-40.5]'     '(39.5-40.5]'     '(39.5-40.5]'     '(39.5-40.5]'
native-country         United-States     United-States     United-States     United-States     United-States
income                         >50K              >50K              <=50K             >50K              <=50K
```

The clustering result shows that:

(1) The first three clusters are of similar size, while the fourth has fewer instances. An obvious different feature is that the attribute 'sex' is male for the first three groups while it is female for the fourth group. According to the values of each attribute, we can see that for features such as 'workclass', 'race', and 'native-country', there is no difference among the four clusters. Thus, we can conclude that these factors do not have significant influence on the clustering result.

(2) The centroid of Cluster#0 is characterized by the following attribute values: middle-age group (38.9 – 46.2), most of them are males with master's degree, work as professional-specialty with an average income over 50K. The majority of this cluster makes investment but has a low capital- gain (-inf-57) and low capital-loss (-inf-77.5).

(3) The centroid of Cluster#1 is characterized by the following attribute values: middle- age group (31.8 – 38.9) but younger than Cluster#1, most of them are

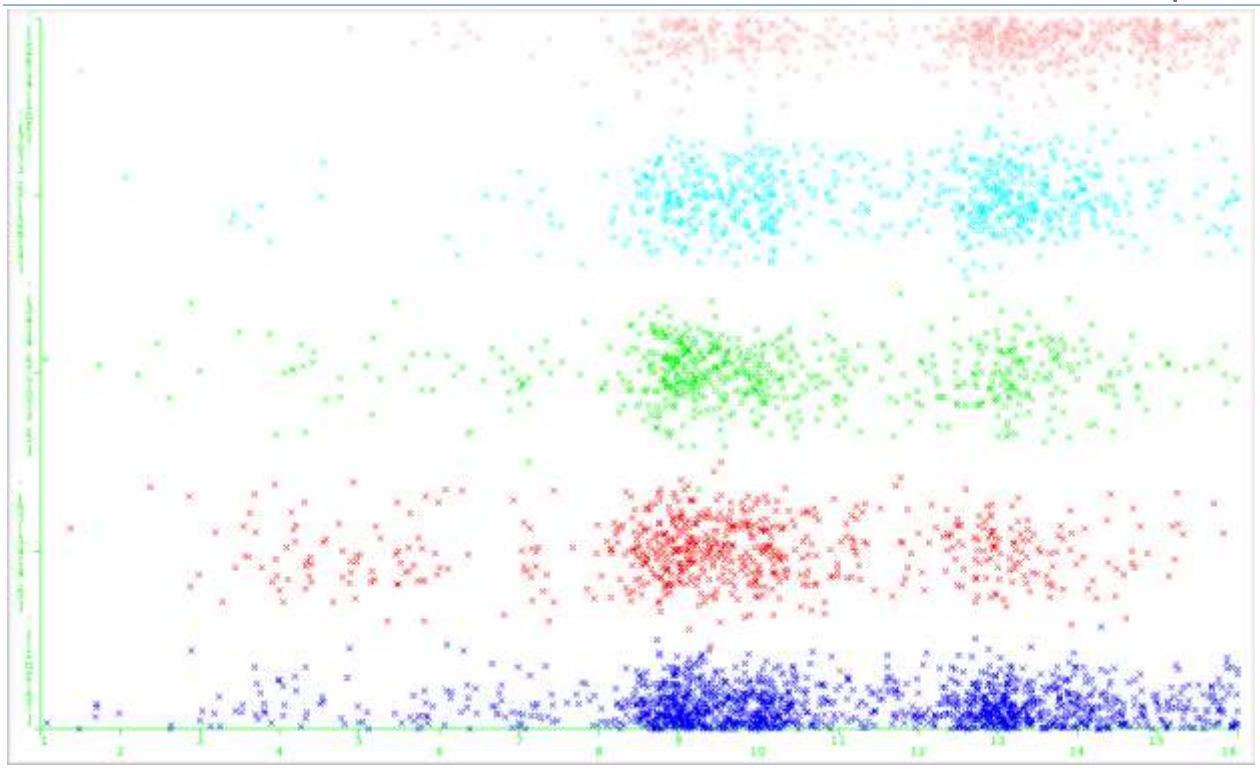
males with an education level of High school Graduate, mostly work as craft-repair with an average income less than 50K. The majority of this cluster makes investment, and has capital- gain larger than Cluster#1 (57-3414.5) and a low capital-loss (-inf-77.5).
Compared to cluster#0, members of cluster#1 tend to be younger, with a lower education level and lower annual income, but the majority members can have more capital- gain than the Cluster#1.

(4) The centroid of Cluster#2 is characterized by the following attribute values: middle- age group (38.9 – 46.2), most of them are males with a Bachelor's degree, mostly work as Professional-Specialty with an average income over 50K. The majority of this cluster makes investment, and has the largest capital- gain (12614-inf) and a low capital-loss (-inf-77.5).
Compared to cluster#0 and #1, the features of cluster#2 are closer to cluster#0, and the only distinct difference is the education level. The majority of the instances own Bachelor's degree, which is lower than cluster#0, but they are more likely to have better return on investment.

(5) The centroid of Cluster#3 is characterized by the following attribute values: young women (-inf-24.3) with high school graduate degree, working as Administration-clerical with an average income less than 50K. The majority of this cluster makes investment but has both low capital- gain (-inf-57) and low capital-loss (-inf-77.5).
Compared to the other clusters, the majority members of cluster#3 are young females with a relatively low education level. The major occupation is also different, but they still make investment.

Based on the analysis of the clustering result, we found that education level and occupation have significant impact on the clusters. We then use 'visualize cluster assignments' to further examine the relationship between some key attribute.

X: education -num (Nom), Y: capital -gain (Nom), Color: capital -gain(Nom)



From the above chart, we can see that most of the people who make investments have an education level higher than '8', so most of them, at the least, graduate from high school. However, this fact does not necessarily mean that the higher the education level is the better return they can get.

From this chart, we can see that most of the spots are centralized on 9 (HS-Grad), 10 (Some Collage) and 13 (Bachelor's degree). For the first three values area, more spots gather near 9 and 10 than above 13, for the fourth value (7073.5 – 12614), the spots distribute around the two areas, and for the group, which have the largest return on investment, and more instances gather over 13 that are over bachelor's degree than 9 and 10. From the above analysis, we can roughly conclude that, the higher the education people have, the more likely they can have better return on investment.

Comparison

In order to better compare the characteristics of people who are more likely to make an investment with returns with people who are less likely to do so, we compare the clustering results of the two subsets to do the further examine.

1. **Age group** - The first phenomenon we found is that the age groups of the first subset distribute among the four clusters while in the second subset, the majority of them are ofmiddle age.

2. **Clusters with similar features**

Attribute	Sub2 --C3 (708)	Sub1 --C0 (9934)	Sub2 --C2 (1114)	Sub1 --C2 (3952)
age	'(-inf-24.3]'	'(-inf-24.3]'	'(38.9-46.2]'	'(38.9-46.2]'
workclass	Private	Private	Private	Private
education	HS-grad	Some-college	Bachelors	Bachelors
education-num	9	10	13	13
marital-status	Never-married	Never-married	Married-civ-spouse	Married-civ-spouse
occupation	Adm-clerical	Adm-clerical	Prof-specialty	Exec-managerial
relationship	Not-in-family	Not-in-family	Husband	Husband
race	White	White	White	White
sex	Female	Female	Male	Male
capital-gain	'(-inf-57]'	0	'(12614-inf)'	0
capital-loss	'(-inf-77.5]'	0	'(-inf-77.5]'	0
hours-per-week	'(39.5-40.5]'	'(39.5-40.5]'	'(39.5-40.5]'	'(39.5-40.5]'
native-country	United-States	United-States	United-States	United-States
income	<=50K	<=50K	>50K	>50K

We picked two clusters from each subset to make a comparison separately. The two clusters from the left section show young women with different education levels might have different investment decisions. Young women with an education level of high school seem to make an investment, but if the education level is some-college, they will choose not to make an investment. In this comparison group, the education level is the factor that determines the result. In the right section, the only different feature is occupation, in other words, middle-age males with professional-specialty positions are more likely to not just make investments, but also make investments that have the best return out of any other group. Nevertheless, most of the middle-age males with exec-managerial position will not choose to make investment.

RESULT AND IMPLICATIONS

In summary, our clustering analysis shows that the four clusters in two subsets both distinguish from each other based on the combination of age, education level, occupation, and income. However, for other attributes used in the dataset, they do not significantly affect the clustering results. Particularly, middle-age males with education level higher than that of a high school graduate are more likely to make investments and have returns on those investments. It is probably safe to assume that the higher education level they have, the more likely they can make wise decisions, and have better returns on their investments. In addition, young women with admin-clerical occupations will make an investment, but both their capital gainsand capitallosses are lower. Additionally, the majority of the instances from this dataset have low capitallosses (-inf-77.5).

However, there are still limitations in our exploration process due to the attributes of our dataset. From the current data we have, we can only examine the characteristics of people with different capital gains and capital losses, but simply given these two attributes, we still cannot conclude how well they do in the investment process. In order to analyze the investment performance for different groups of people and give recommendations on people's investment decisions, more data, such as how much they invest or the percent of the capital gains they receive, should be included.

GROUP ROLES AND RESPONSIBILITIES

At one time or another, each of us shared the general roles and responsibilities for executing this project. We executed the project with a team mindset.a more specific listing of roles and responsibilities is listed in the below Participant Matrix.

Participant Matrix

Role	Responsibilities	Participant(s)
Project Sponsor	• Ultimate decision-maker • Provide project oversight and guidance • Review/approve some project elements	Bei Yu
Chief Data Scientist	• Managed machine learning approaches • Utilized data modeling and best practices • Lead data review meetings • Matched analysis with real world applications	Chenzi Qian
Chief Statisticia n	• Formulated and executed strategies for obtaining and handling the data • Built probabilistic models • Lead innovation	Lingwen Zhang
Project Manager	• Directed/lead team members toward project objectives • Handled problem resolution • Aided with analysis • Edited final paper	Joshua Kitlas