Factors of Income Inequality
OIM 454
By: Peilin Guo, Xin Li, Zhihang Chen

Table of Contents

Project Description

For our final project, we want to examine the factors that would determine people's income. As college students, we would always think about one question: what will our future income be? It's hard to answer that question without examining the data. There are many factors that people like to mention when talking about the income: occupution, education, sex, race, etc. One of our purposes is to find out if those exploratory variables really contribute to the determination of income by applying the data mining technology that we learned in this class, for example, logistic regression and decision trees.

In the recent years, the increase in the discussion of the gender inequality and race discrimination also brings to our attention, so it's very important for us to observe the relationship between race and gender to the income to see if the claim is true. We will dig into the data to analyze this social problem.

Data Preprocessing

Our dataset is about US Adult income, and it is from 1994 US Census Database. There are two sets of data within that database, training set and test data. Because the training set is already composed of 32560 records by itself, we will only be using the training set for the purpose of our project. There are different exploratory variables which are composed of continuous variables and category variables. For the category variables, we have income tax bracket, workclass, education, marital status, occupation, race, sex, native country. For the numerical variables, we have age, fnlwgt, education(it was already transformed to the dummy variables in the dataset), capital gain, capital loss, work hours per work. This dataset satisfies the purpose for our project, because it includes the independent variables(sex and gender) that we especially desired, and the output variable(income tax bracket). The disadvantage of this dataset is that it does not have a specific amount for the income, instead it only indicates if it is >50k or <=50k. Because of that, we can only build classification models instead of prediction equations.

Most of our variables are pretty self-explanatory, except the fnlwgt and relationship. For fnlwgt, it's the number of people the census believes the entry represents. It's hard to use and analyze this factor in our model, so we will exclude that one. Another one is relationship, it means if that person is a wife, or own-child, or husband, or not-in-family, or other-relative, or unmarried. There are few missing data in our dataset, because we have so many records, we would like to exclude those records. Other than that, our data is kind of clean.

First of all, we need to transform some categorical variables to dummy variables. The variables that we are going to transform are marital status, occupancy, relationship, race and gender. We ignore the country of origin, because it has over 30 distinct values. There is one important thing to keep in mind, for our version of Excel Data Mining, it can only handle <50 columns and <10,000 rows data for training set during the partition process. That's why we need to be careful about the dummies when partitioning, because we may create more than 50 columns, if the categorical variable has too many discrete values. We can not separate it into 70% for the training set and 30% for the validation set as we will do, because then the training set will have more than 10,000 rows. We decide to do that in R.

After obtaining the raw data set, we found that data cleaning had to be done in order to build our models. First, we removed records with missing values. Then, we created dummy variables on categorical fields to allow easier modeling. From exploratory analysis, we determined the outliers and removed them from our analysis because we think they are not representative of the population. Lastly, we created appropriate headers to the data frame for better data communication.

```
```{r}
adultdata<-read.csv("adult-training.csv")
adultdata[adultdata=="?"]<-NA
adultdata<-na.omit(adultdata)
adultdata1<-rename(adultdata,age=X39,workclass=State.gov,education=Bachelors
,educationLevel=X13,hoursPerWeek=X40,gender=Male,race=White,married=Not.in.f
amily,income=X..50K)
adultdata1<-subset(adultdata1,select =
-c(X77516,Never.married,Adm.clerical,X2174,X0))


```
```

```
```{r}
adultdata1[adultdata1 ==" ?"] <- NA
adultdata1<-na.exclude(adultdata1)
adultdata2<-adultdata1
```

```{r}
adultdata2$gender<-ifelse(adultdata2$gender== " Male",1,0)
adultdata2$workclass<-ifelse(adultdata2$workclass==" Private",1,0)
adultdata2$married<-ifelse(adultdata2$married=="
Husband"|adultdata2$married==" Wife",1,0)
adultdata2$race<-ifelse(adultdata2$race==" White",1,0)

```

Backward Elimination in R

To minimize the problems of having too many variables, such as overfitting or multi-collinearity, we used backward elimination to select our predictors. We start off by including all available predictors in our regression, and remove variables that are not significant by their p-values. We then run the regression again and repeat the elimination until when all the remaining predictors are significant. Our final model consists of age, educationLevel, maritalStatus, race, and hoursPerWeek.

```
```{r}
fit1<-glm(income~age+workclass+educationLevel+married+race+gender+hoursPerWe
ek,data = adultdata2,family = binomial())
summary(fit1)
```

Call:
glm(formula = income ~ age + workclass + educationLevel + married +
    race + gender + hoursPerWeek, family = binomial(), data = adultdata2)

Deviance Residuals:
    Min   1Q   Median   3Q   Max
-2.7664 -0.5957 -0.2694 -0.0517   3.5179

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)   -9.656350   0.149252 -64.698 < 2e-16 ***
age   0.033490   0.001439 23.278 < 2e-16 ***
workclass   0.109987   0.036722   2.995 0.00274 **
educationLevel 0.387091   0.007662 50.522 < 2e-16 ***
married   2.269905   0.043116 52.647 < 2e-16 ***
race   0.222311   0.054526   4.077 4.56e-05 ***
gender   0.110308   0.046321   2.381 0.01725 *
hoursPerWeek   0.031001   0.001500 20.665 < 2e-16 ***

```

```
```{r}
fit2<-glm(income~age+workclass+educationLevel+married+race+hoursPerWeek,data
= adultdata2,family = binomial())
summary(fit2)
```

Call:
glm(formula = income ~ age + workclass + educationLevel + married +
    race + hoursPerWeek, family = binomial(), data = adultdata2)

Deviance Residuals:
    Min   1Q   Median   3Q   Max
-2.7707 -0.5965 -0.2687 -0.0505   3.5331

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)   -9.628770   0.148760 -64.727 < 2e-16 ***
age   0.033549   0.001438 23.332 < 2e-16 ***
workclass   0.110250   0.036720   3.002 0.00268 **
educationLevel 0.386484   0.007657 50.477 < 2e-16 ***
married   2.313878   0.039152 59.100 < 2e-16 ***
race   0.228706   0.054429   4.202 2.65e-05 ***
hoursPerWeek   0.031581   0.001481 21.322 < 2e-16 ***

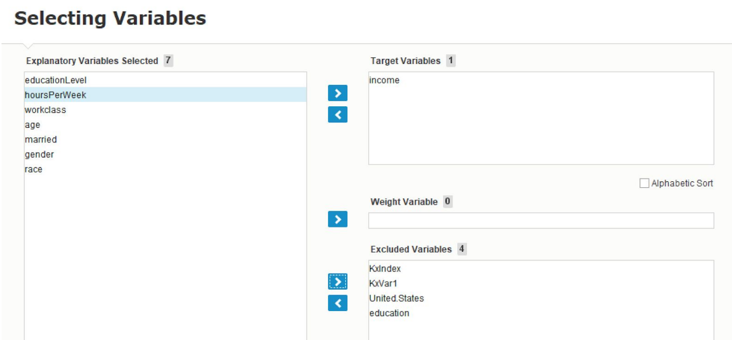
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
```{r}
fit3<-glm(income~age+educationLevel+married+race+hoursPerWeek,data =
adultdata2,family = binomial())
summary(fit3)
```

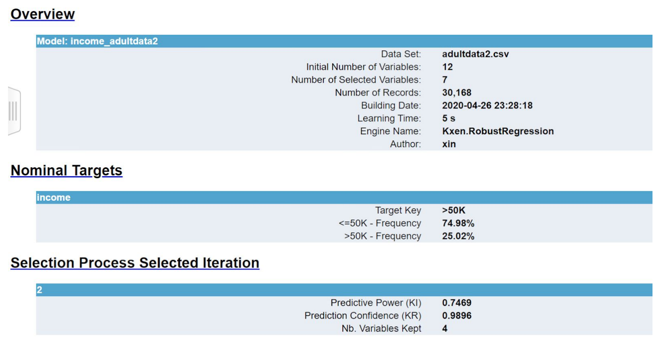
Call:
glm(formula = income ~ age + educationLevel + married + race +
    hoursPerWeek, family = binomial(), data = adultdata2)

Deviance Residuals:
    Min   1Q   Median   3Q   Max
-2.7420 -0.5982 -0.2694 -0.0514   3.4987

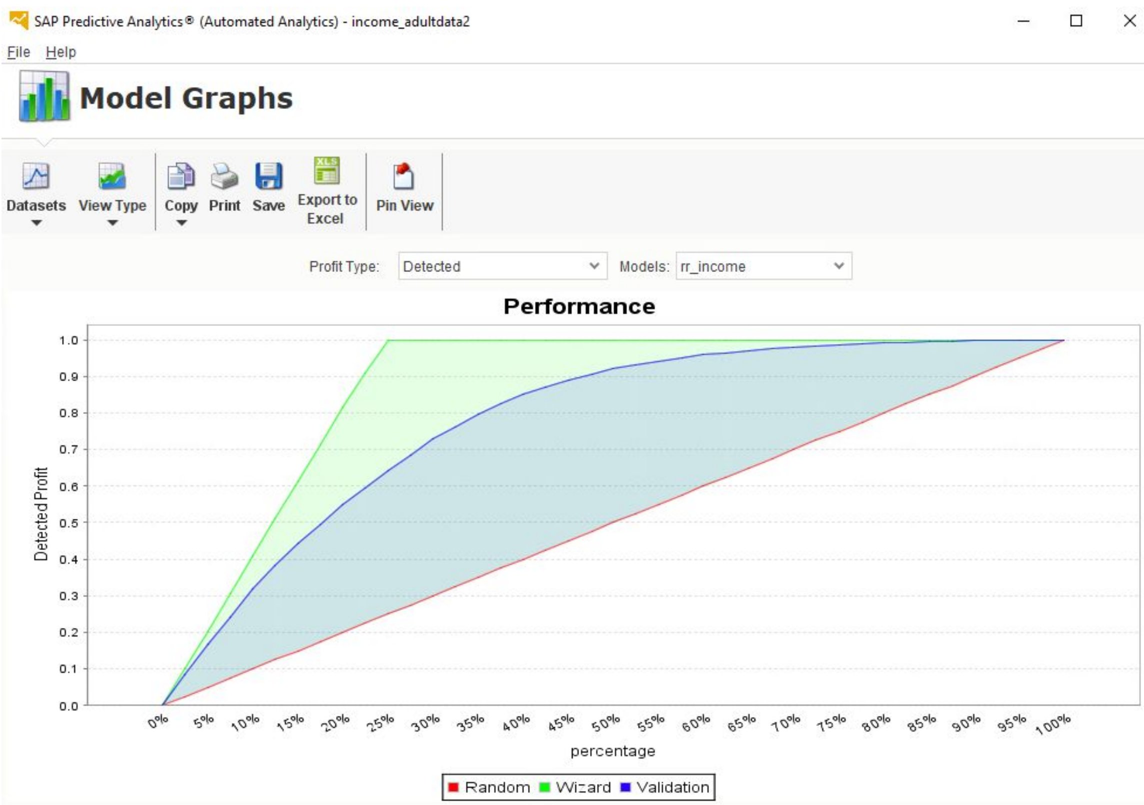
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)   -9.469504   0.138433 -68.405 < 2e-16 ***
age   0.032743   0.001412 23.192 < 2e-16 ***
educationLevel 0.383341   0.007574 50.611 < 2e-16 ***
married   2.309660   0.039122 59.038 < 2e-16 ***
race   0.231427   0.054420   4.253 2.11e-05 ***
hoursPerWeek   0.031235   0.001475 21.172 < 2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

**SAP Analysis**

After the data preprocessing and data cleaning in R, we input the modified data into the SAP Analysis. We want to see which variables contribute to the income variable the most, and see if the outcome matches with our result in R. After running the model, the predictive power of our model is 0.747 which is pretty strong and there are 4 variables kept in our model.
```

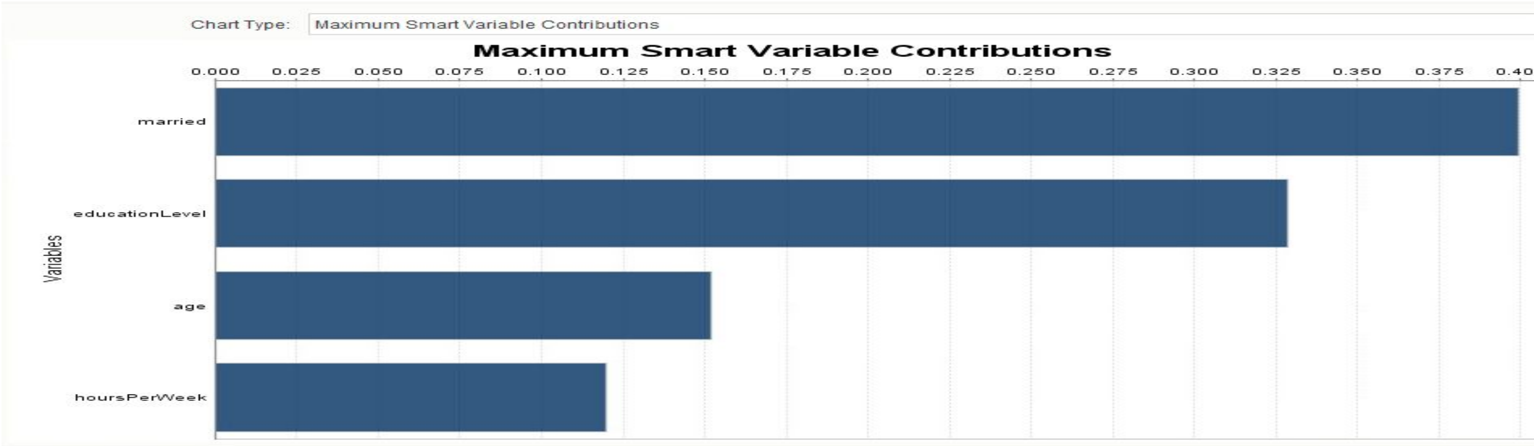




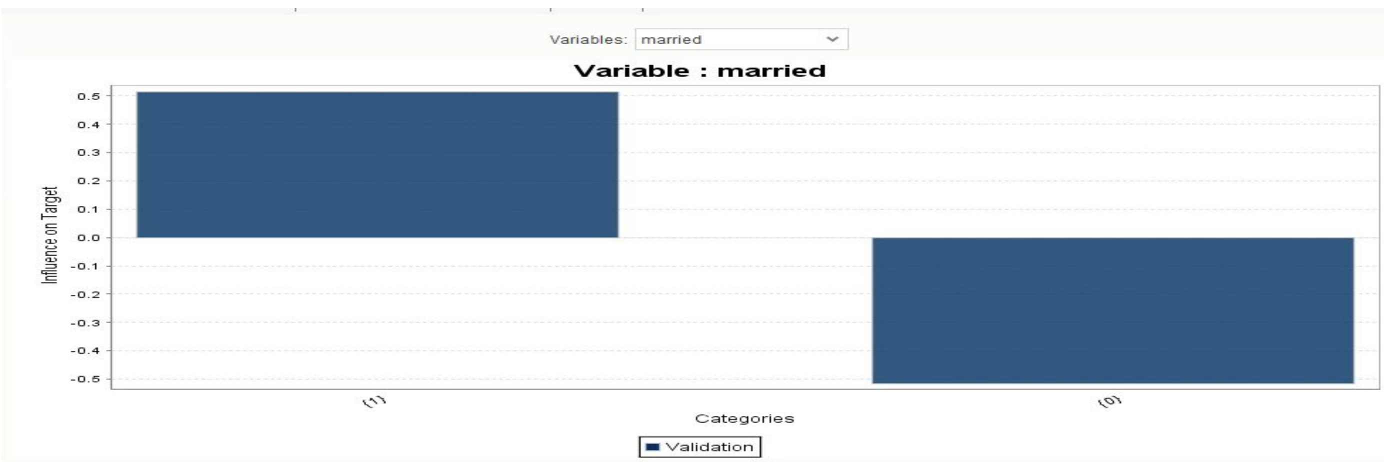
First, we take a look at the Profit Curve, which shows the quality of the model. The x-axis shows the percentage of the total dataset, and the y-axis shows the percentage of the correctly identified target. The red line indicates the random distribution of the target variable. That line is linear. The green line indicates the perfect model. All targets are identified first. The blue line is our model. The closer it is to the top left corner the better is the model. From this graph, we can see that our model is a pretty good one.



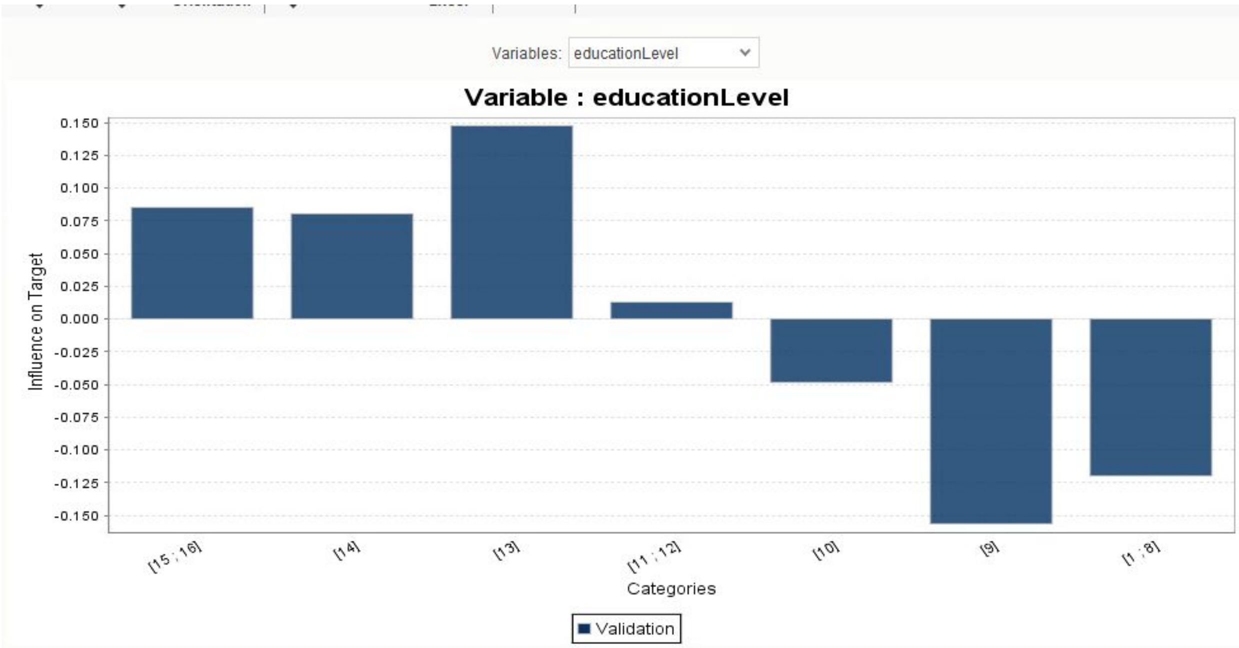
Next, we want to take a look at the Contribution by Variables in the Display section. We can see that marital status is the most important factor, followed by educationLevel, age and hoursPerWeek. This is similar to the results that we got in R, except this does not include race variables.



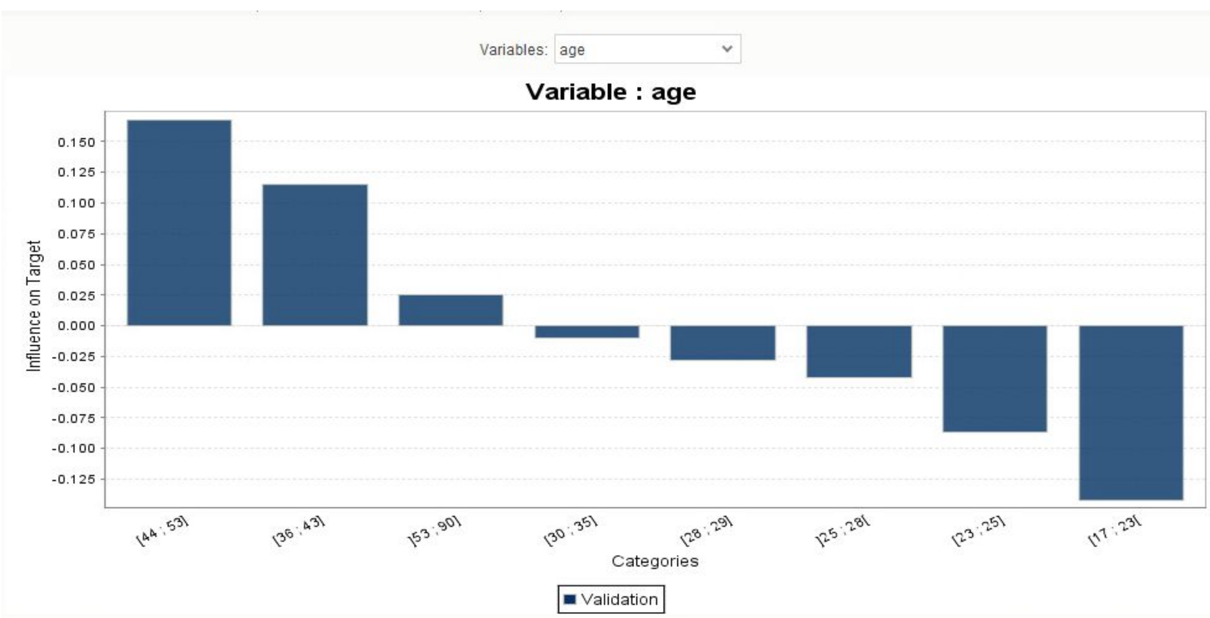
Then, we want to see how each variable contributes to our output variable - income. By observing the married variable graphs, we can see that people that are married tend to have a much higher propensity to have an income that is greater than 50k, and people who are not married tend to have a lower propensity. Another way to see this is that married people have a positive number for influence on target which means that married people have a greater likelihood of higher income(>50k), and vice versa.



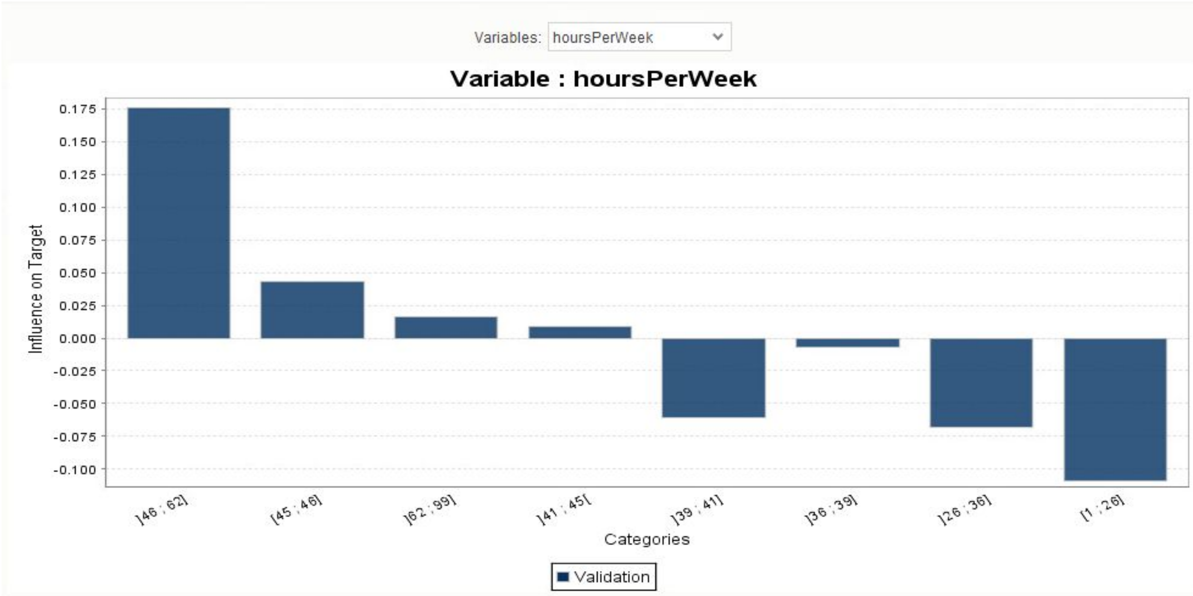
By looking at the educationLevel graph, we can see professor school:doctorate, masters, bachelors and associates have positive influences on the income, and people who have some college experiences, high-school graduates and middle school or elementary school tend to have negative influences on the income. That means, people who complete the bachelor's degree or above will have a high propensity to have income that is greater than >50k, and vice versa.



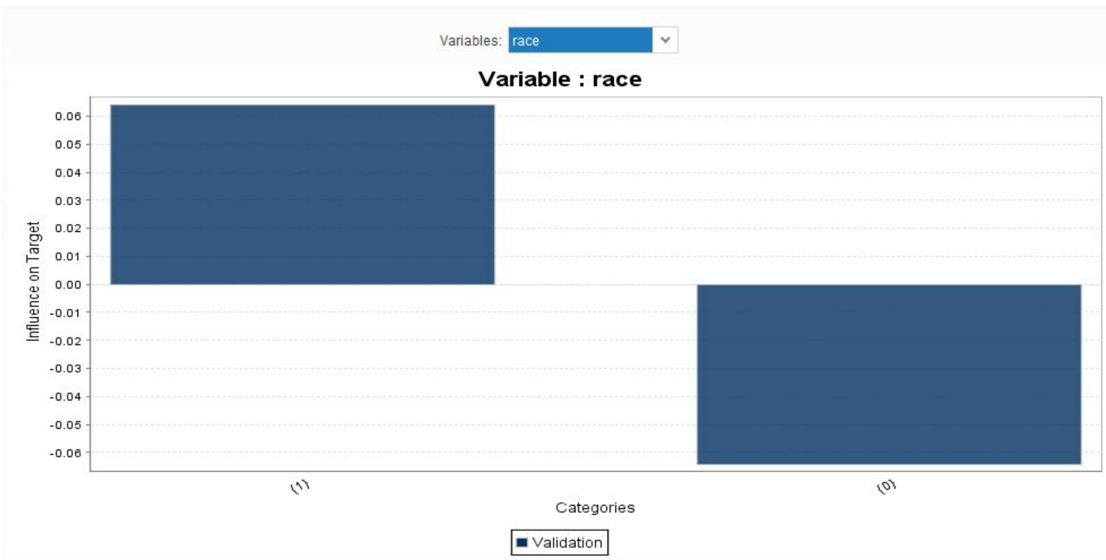
For the age variable, it is pretty clear to see that people above the age of 36 tend to have a higher propensity to have an income that is >50k, and people who are younger tend to have an income that is <=50k. This is very understable, because people who are younger have less experience, which will make them receive less income.



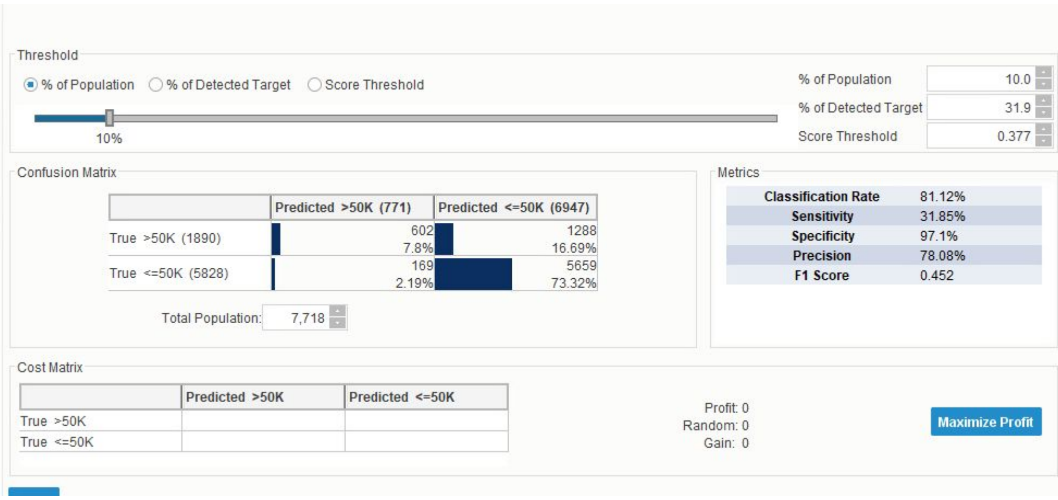
From the hoursPerWeek variable, we can see that people who work more than 40 hours per week will have a higher chance to have an income that is >50k.

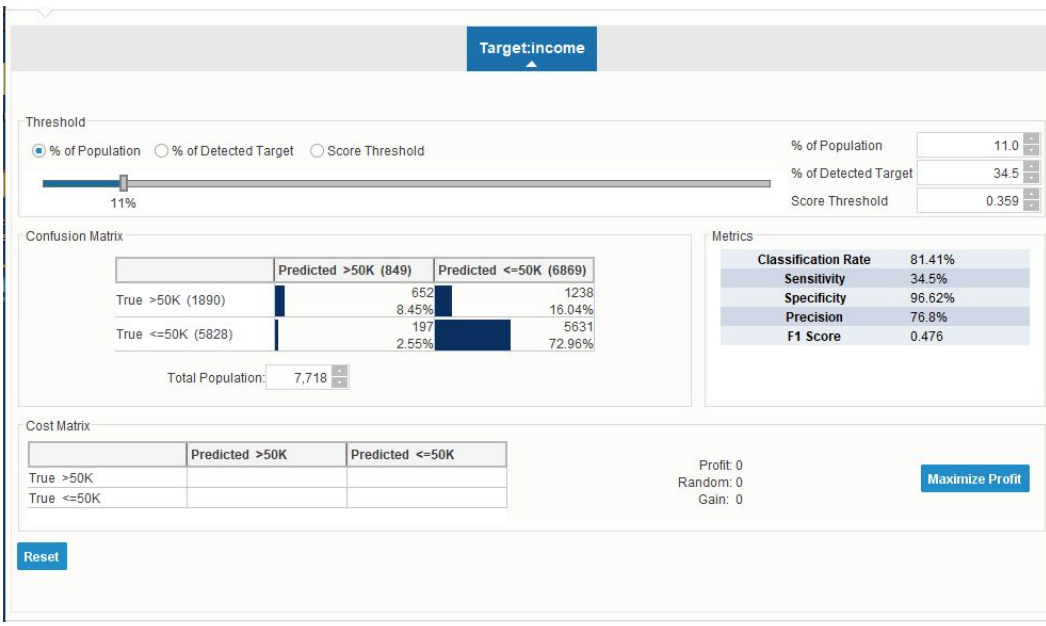


For race variables, we can see that white people have a higher propensity to have income that is greater than 50k, and other races of people have a lower propensity to be in the higher range of income.



To summarize our finding, people who are married, completed bachelor degree or higher education, work more than 40 hours per week and above 41 years old are tend to have a higher propensity to have an income that is greater than 50k, vice versa. Now, we understand the model, so we want to have a look at the value this model is adding. The confusion matrix tells us how well the model predicts the correct outcome. By setting our % of population to 10%, we can see that the % of detected targets is 31.9. This is pretty efficient, and the classification rate is 81.12%.





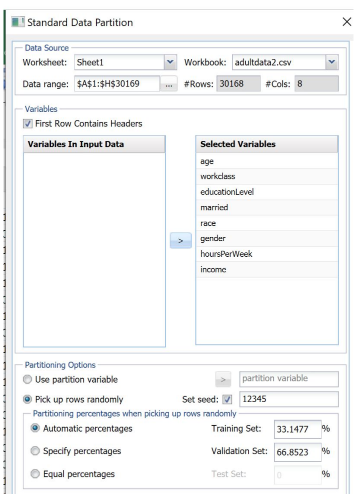
Correlation

In order to better understand the data, a correlation table between difference variables is established. The definition of correlation is a mutual relationship or connection between two or more variables. From the table, we can see that marital status has the highest correlation with income. The correlation table reveals some close relationship between predictors and we want to remove those variables to minimize the risk of multicollinearity in the model. Gender has a high correlation with married variable, so we may want to remove that variable and keep the marital status variable, since it has a higher correlation with income variable in comparison. But from the table, there are no two variables that have a very strong relationship, between all correlation numbers is less than 0.5.

	A	B	C	D	E	F	G	H	I	J
1		*age*	*workclass*	*education*	married	race	gender	hoursPer\	income	
2	age	1								
3	workclass	-0.20985	1							
4	education	0.043848	-0.16452	1						
5	married	0.313586	-0.12932	0.087199	1					
6	race	0.026961	-0.00509	0.052526	0.11248	1				
7	gender	0.081902	-0.06677	0.006077	0.439214	0.105185	1			
8	hoursPer\	0.101917	-0.09508	0.152656	0.225936	0.056293	0.231201	1		
9	income	0.242118	-0.11699	0.335374	0.449366	0.084778	0.216672	0.229545	1	
10										

Logistic Regression

Next, we want to build our model. Because our outcome variable has two classes, class 1 consists of people with income more than 50k, and class 0 consists of people who have income less or equal to 50k. We need to use a classification tool, which is the logistic regression function. We did a partition on all variables, except country of origin, because it has over 30 distinct values. There is one important thing to keep in mind, for our version of Excel Data Mining, it can only handle <50 columns and <10,000 rows for training set during the partition process. So, we will use automatic percentages, instead of the usual percentage we used in the homework assignments.



Inputs

Data	
Workbook	adultdata2.csv
Worksheet	Sheet1
Range	$A$1:$H$30169
# Records in the input data	30168

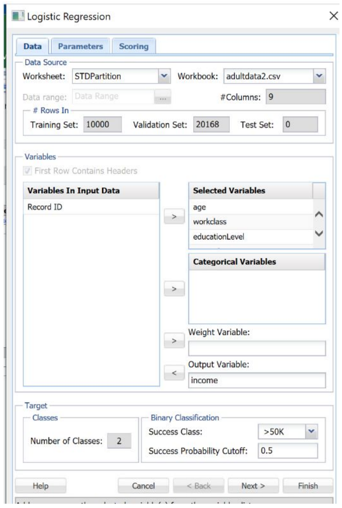
Variables								
# Selected Variables	8							
Selected Variables	age	workclass	educationL	married	race	gender	hoursPerW	income

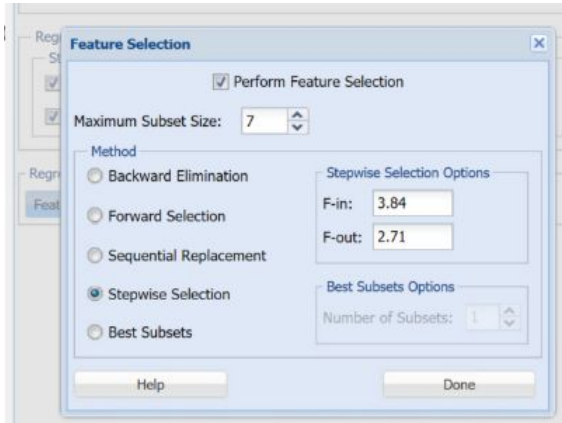
Partitioning Parameters	
Partitioning type	RANDOM
Random seed	12345
Ratio - Training	0.331477
Ratio - Validation	0.668523

Partition Summary

Partition	# Record:
Training	10000
Validation	20168

The stepwise selection method would be used, it starts with an empty model, and each step it can add or remove a variable until all the variables are significant according to p-value or some other criterias. We got six different best subset, by looking at the RSS which is the residual sum of squares, or the sum of squared deviations between the predicted probability of success and the actual value (1 or 0), it is clear that RSS for subset 6 is very close to the RSS for subset 5, which means adding last variable(race) will not significantly improve our model, but we include race variable in R, and we want to be consistent with our result, we are going to pick subset 6.





Best Subsets

Subset ID	Intercept	age	workclass	educationLevel	married	race	gender	hoursPerWeek
Subset 1	1	0	0	0	0	0	0	0
Subset 2	1	0	0	0	1	0	0	0
Subset 3	1	0	0	1	1	0	0	0
Subset 4	1	1	0	1	1	0	0	0
Subset 5	1	1	0	1	1	0	0	1
Subset 6	1	1	0	1	1	1	0	1

Best Subsets Details

Subset ID	#Coefficients	RSS	Mallows's Cp	Probability
Subset 1	1	10957.2	2244.769346	0
Subset 2	2	10053.63	1237.186423	1.3653E-249
Subset 3	3	9247.007	337.9209539	4.15862E-70
Subset 4	4	9088.501	162.8184858	6.89324E-34
Subset 5	5	8959.073	20.20478869	0.000402056
Subset 6	6	8945.874	7.457346053	0.177572904

From the classification summary of the training set, we can see that it has an accuracy rate of 82.15%. And for the validation set, the accuracy rate is 81.47%. Those two numbers are pretty high, and close to what we got in SAP Predictive Analytics. But there is one that brings our attention to, our model does not have a good performance on predicting the 1's class, which is the people who have income more than 50k. Maybe we can change the cutoff probability next time. For now, we will stick with this number, because the error rate for 0's class is low and we want to be more conservative on the prediction of income. We don't want to give out fake hope.

Training: Classification Summary

Confusion Matrix

Actual\Predicted	<=50K	>50K
<=50K	7006	584
>50K	1201	1209

Error Report

Class	# Cases	# Errors	% Error
<=50K	7590	584	7.694335
>50K	2410	1201	49.83402
Overall	10000	1785	17.85

Metrics

Metric	Value
Accuracy (#correct)	8215
Accuracy (%correct)	82.15
Specificity	0.923057
Sensitivity (Recall)	0.50166
Precision	0.674289
F1 score	0.575303
Success Class	>50K
Success Probability	0.5

Validation: Classification Summary

Confusion Matrix

Actual\Predicted	<=50K	>50K
<=50K	13857	1213
>50K	2524	2574

Error Report

Class	# Cases	# Errors	% Error
<=50K	15070	1213	8.049104
>50K	5098	2524	49.50961
Overall	20168	3737	18.52935

Metrics

Metric	Value
Accuracy (#correct)	16431
Accuracy (%correct)	81.47065
Specificity	0.919509
Sensitivity (Recall)	0.504904
Precision	0.679694
F1 score	0.579403
Success Class	>50K
Success Probability	0.5

Our selected predictors are age, educationLevel, marital status, race and hoursPerWeek. The logistic equation will be (logit=-9.71+0.034age+0.40educationLevel+2.34married+0.36race+0.029hoursPerWeek). We can use this equation to model the log odds of an event as a linear function of the predictors. We can also use the confidence interval to do any future analysis. It can be used for binary responses in our case, to classify a person as class 1 (income > 50k) or class 0 (income <=50k).

Predictor Screening

Predictor	Criteria	Included
Intercept	16.38320806	TRUE
age	1607.114538	TRUE
educationLe	317.1176021	TRUE
married	46.93454698	TRUE
race	36.70857743	TRUE
hoursPerWe	4258.754748	TRUE

Tolerance fo	9.45614E-09

Coefficients

Predictor	Estimate	Confidence Interval: Lower	Confidence Interval: Upper	Odds	Standard Error	Chi2-Statistic	P-Value
Intercept	-9.712413325	-10.20033009	-9.224496557	6.05275E-05	0.248941701	1522.155451	0
age	0.033872589	0.029010929	0.038734249	1.034452798	0.002480484	186.4763867	1.869E-42
educationLe	0.395399193	0.368983576	0.421814809	1.484976865	0.013477603	860.6889745	3.45E-189
married	2.337817843	2.202867001	2.472768685	10.35860777	0.068853735	1152.832803	1.09E-252
race	0.363957233	0.167781434	0.560133031	1.43901267	0.100091532	13.22227034	0.0002766
hoursPerWe	0.028937008	0.023903796	0.03397022	1.029359751	0.002568013	126.9734613	1.883E-29

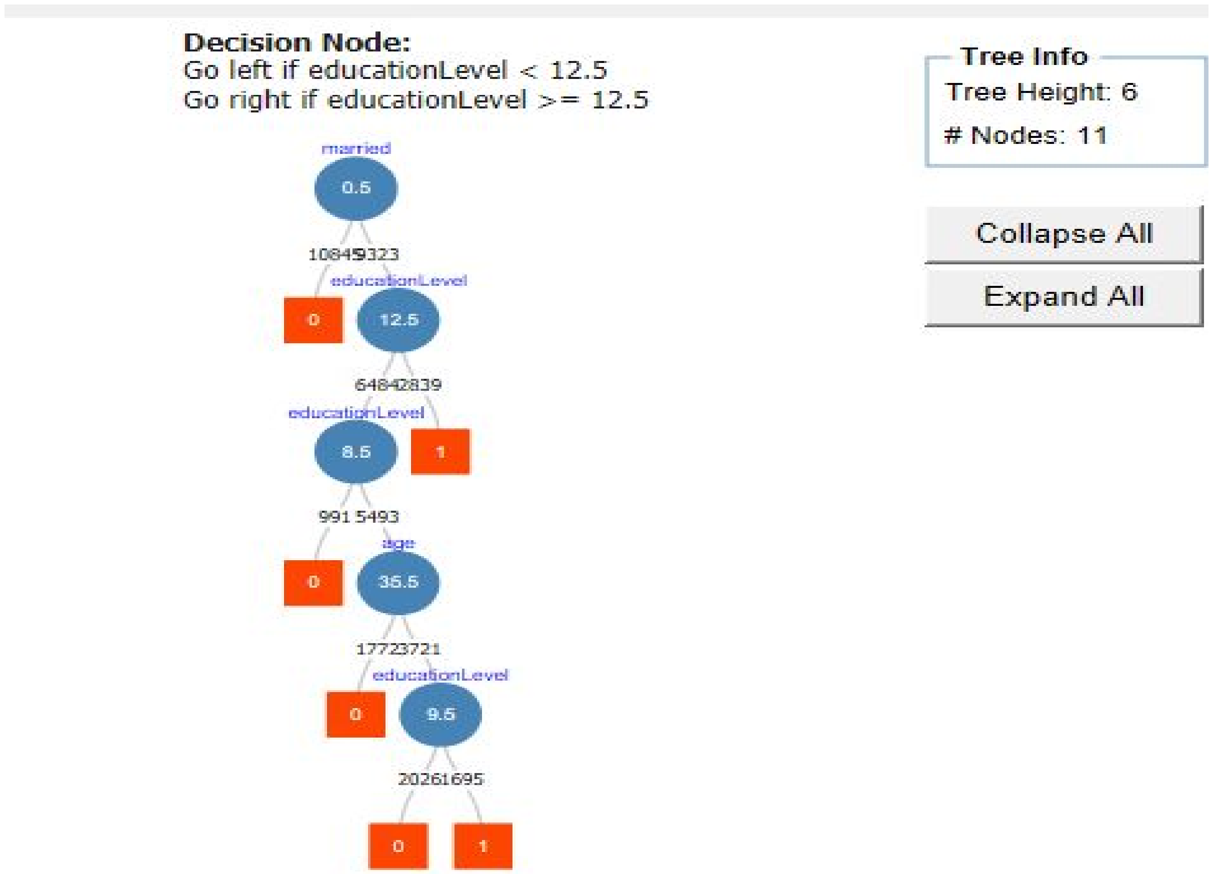
Variance-Covariance Matrix of Coefficients

Predictor	Intercept	age	educationLevel	married	race	hoursPerWeek
Intercept	0.06197197	-0.000303839	-0.002237616	-0.00453558	-0.008575933	-0.000298949
age	-0.000303839	6.1528E-06	2.80406E-06	-1.435E-05	-3.60822E-06	6.6095E-07
educationLe	-0.002237616	2.80406E-06	0.000181646	0.000206656	2.21312E-05	-3.68354E-07
married	-0.00453558	-1.435E-05	0.000206656	0.004740837	-0.000346492	-7.1354E-06
race	-0.008575933	-3.60822E-06	2.21312E-05	-0.000346492	0.010018315	-7.74984E-06
hoursPerWe	-0.000298949	6.6095E-07	-3.68354E-07	-7.1354E-06	-7.74984E-06	6.59469E-06

Our results do not reveal a strong correlation between income and gender (race), but this may due to problem in our preprocessing process. We eliminate some outliers without doing detailed data exploration, and those outliers may reveal some important information that we are not aware of. As we mention in the introduction, we want to see the effect of gender and race on the income, those two thousands data points that we eliminate may contribute some evidence to those issues.

Decision Tree

Last but not least, we will use the classification tree to validate our model. Our predictors will still be age, educationLevel, married, race and hoursPerWeek. For the simplicity, we will only look at the best pruned tree. From that tree, we can see that married, educationLevel and age are the three most important factors that determine the classification of the two classes. The classification summary of training and validation is very similar to the summary of logistic equations. It's also clear that our model more accurately predicts 0's class instead of 1's class.



If married < 0.5 classification= 0

if married >=0.5 and education level >= 12.5 classification= 1

If married >=0.5 and education level >= 12.5 and education level < 8.5 classification=0

if married >=0.5 and education level >= 12.5 and education level >=8.5 and age < 35.5 classification=0

if married >=0.5 and education level >= 12.5 and education level >=8.5 and age >35.3 and education level < 9.5 classification = 0

if married >=0.5 and education level >= 12.5 and education level >=8.5 and age >35.3 and education level < 9.5 classification = 1

Training: Classification Summary

Confusion Matrix		
Actual\Predicted	0	1
0	6836	754
1	1030	1380

Error Report			
Class	# Cases	# Errors	% Error
0	7590	754	9.934123847
1	2410	1030	42.73858921
Overall	10000	1784	17.84

Metrics	
Metric	Value
Accuracy (#correct)	8216
Accuracy (%correct)	82.16
Specificity	0.9006588
Sensitivity (Recall)	0.5726141
Precision	0.6466729
F1 score	0.6073944
Success Class	1
Success Probability	0.5

Validation: Classification Summary

Confusion Matrix		
Actual\Predicted	0	1
0	13525	1545
1	2109	2989

Error Report			
Class	# Cases	# Errors	% Error
0	15070	1545	10.2521566
1	5098	2109	41.36916438
Overall	20168	3654	18.11781039

Metrics	
Metric	Value
Accuracy (#correct)	16514
Accuracy (%correct)	81.88219
Specificity	0.8974784
Sensitivity (Recall)	0.5863084
Precision	0.6592413
F1 score	0.6206395
Success Class	1
Success Probability	0.5

Conclusion

This group project proved to be a challenge for all of us but also a great way to apply the concepts and knowledge we learned throughout the semester. We were able to understand the materials much more clearly and use the techniques for data clean, process, regression, SAP, correlation and R in this project. Our goal is to examine the factors that would determine people's income. First we clean the data by transforming categorical variables to dummy variables for efficient results for the target variables. We transformed marital status, occupancy, relationship, race and gender from categorical variables to dummy and ignore the country of origin because it has over 30 distinct values. We want to maintain enough data while being processable for further analysis. After data preprocess and clean we input modified data into SAP analysis. The result we got from the profit curve was pretty good. The marital status was the most important factor and then education Level,age and hoursPerWeek. Based on the result married people tend to have higher propensity with income greater than 50k. On the other hand, people who are not married have lower propensity. For the educationLevel chart we found out the professor school:doctorate,masters, bachelors and associates have positive on income vs who have highschool, middle school and elementary school have negative on their income. People who are over 36 years old have more than 50k income while youngs have less than 50k. People who worked full time of more than 40 hours per week would have a higher chance to achieve more than 50k of income. People who are white have higher propensity to have greater income. We then use correlation, logistic regression and decision trees to have further understanding of the data. The result is that marital status has the highest correlation with income. For the decision tree married, educationLevel and age are the three most important factors. One crucial takeaway from this project was the importance of data clean and preprocess in order to create a useful model with simplicity of understanding when presenting them. We spend hours modifying the data to make it run properly and with expected results.

Additionally, this project provides a great opportunity for real business experience that we would be expecting in the future. The comparison between different analysis and techniques we learned was fully applied to this project, yet great lesson and valuable assets. In the future, we may want to apply new data set into this function by using the SAP Predictive Analytics to get a solid conclusion on the issues that we care about, since this data set is from 1994.