Prediction of FamilyIncome using Logistic Regression

Takeshi Oda

2018/03/24

Summary

We conducted a brief exploratory analysis and built a regression model on house income using sample data set from "2010 American Community Survey (ACS) for New York state". We found that house type and house ownership could be better predictors for household income.

Dataset

Sample data set for house income in New York state was provided as a training set. Sample data "acs_ny.csv" was prepared based on 2010 American Community Survey (ACS) conducted by United States Census Bureau. (https://www.census.gov/programs-surveys/acs/)

Data Loading and data cleansing

Data loading

Input file "acs_ny.csv" was loaded into R data frame as a training set.
Response variable "Target" is defined as:

If Family Income >= 150,000 then Target = 1
If Family Income < 150,000 then Target = 0

Missing data handling

Next, I checked how much missing values are found in each variable and I found there is no missing value in the dataset.

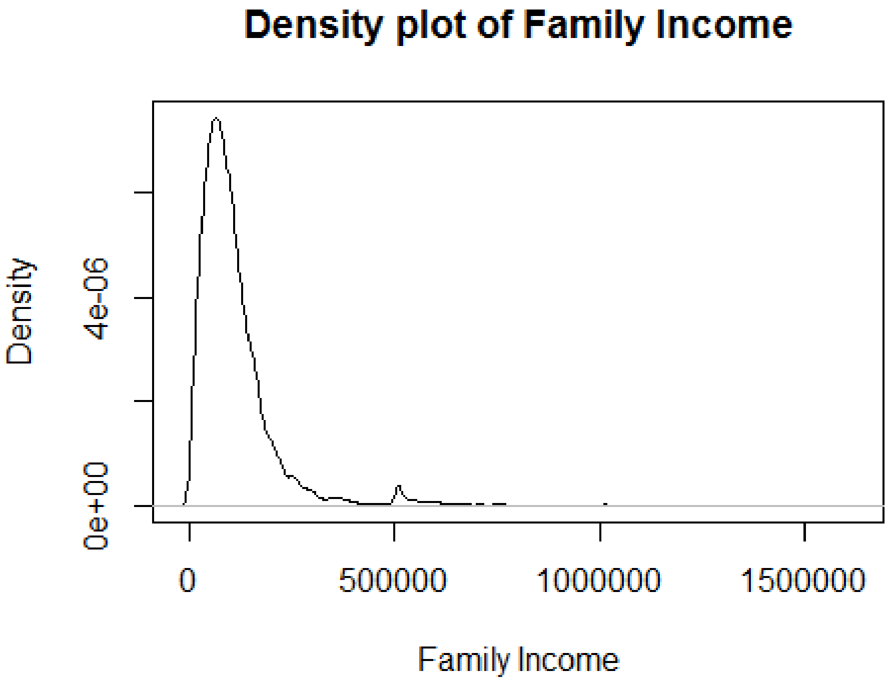
```
## Loading required package: lattice

##      Acres FamilyIncome FamilyType NumBedrooms NumChildren NumPeople
## [1,]     1            1          1           1           1         1
## [2,]     0            0          0           0           0         0
##      NumRooms NumUnits NumVehicles NumWorkers OwnRent YearBuilt HouseC
osts
## [1,]        1        1           1          1       1         1
   1
```

```
## [2,]        0        0           0          0       0         0
   0
##      ElectricBill FoodStamp HeatingFuel Insurance Language Target
## [1,]            1         1           1         1        1      1 0
## [2,]            0         0           0         0        0      0 0
```

Exploratory Data Analysis

Before building model, we conducted exploratory data analysis. First of all, I created density plot of Family Income.

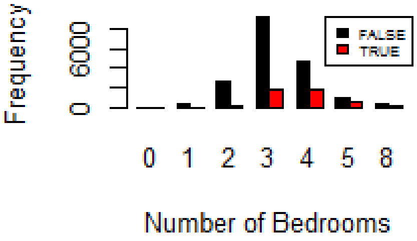
```
## Package 'sm', version 2.2-5.4: type help(sm) for summary information

## Loading required package: grid
```

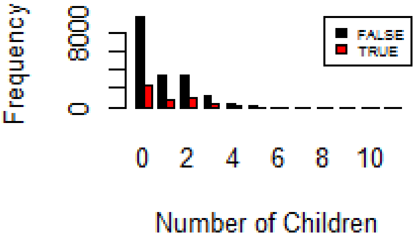


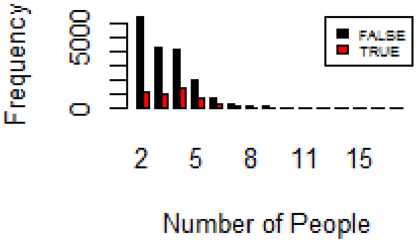
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      50   52540   87000  110300  133800 1605000

## [1] "Percentage of income higher than 150,000:  20.07"
```

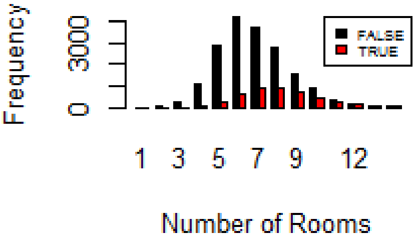
Family Income has right sided skewed distribution with median 87,000 and mean 110,300. Families with income higher than 150,000 dollors are around 20% of all famillies.

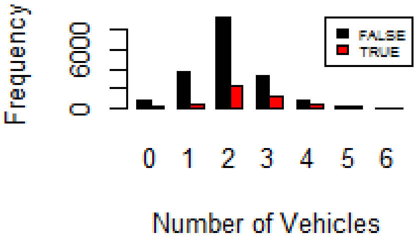
Next, we visualized relationships of each independent variable with response variable.For each independent variable, I created a simple visualization related to response variable.

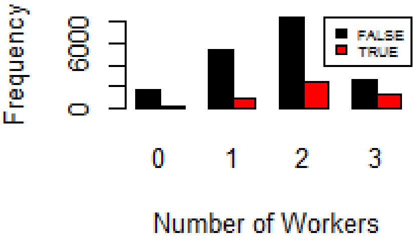

Frequency
0
6000
FALSE
TRUE
0
1
2
3
4
5
8
Number of Bedrooms

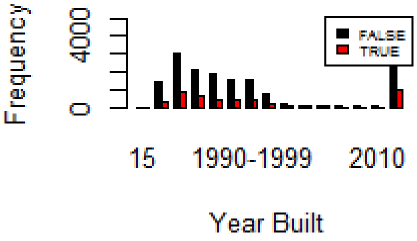

Frequency
0
8000
FALSE
TRUE
0
2
4
6
8
10
Number of Children

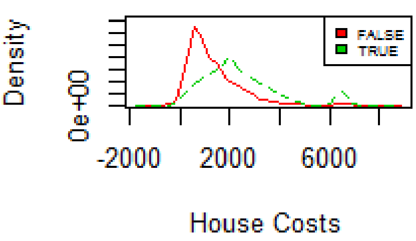

Frequency
0
5000
FALSE
TRUE
2
5
8
11
15
Number of People

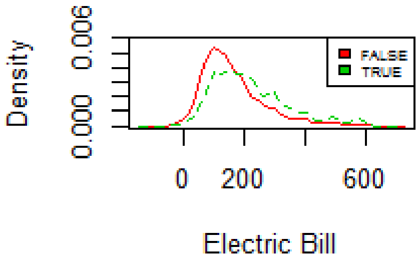

Frequency
0
3000
FALSE
TRUE
1
3
5
7
9
12
Number of Rooms

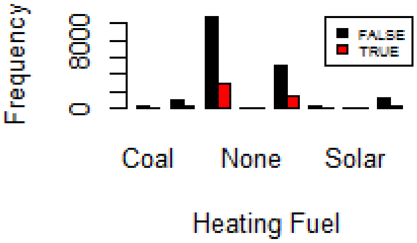

Frequency
0
6000
FALSE
TRUE
0
1
2
3
4
5
6
Number of Vehicles

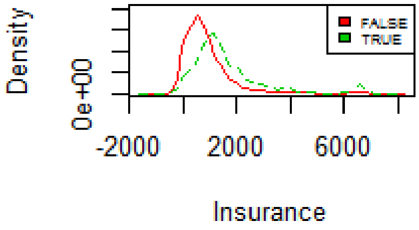

Frequency
0
6000
FALSE
TRUE
0
1
2
3
Number of Workers

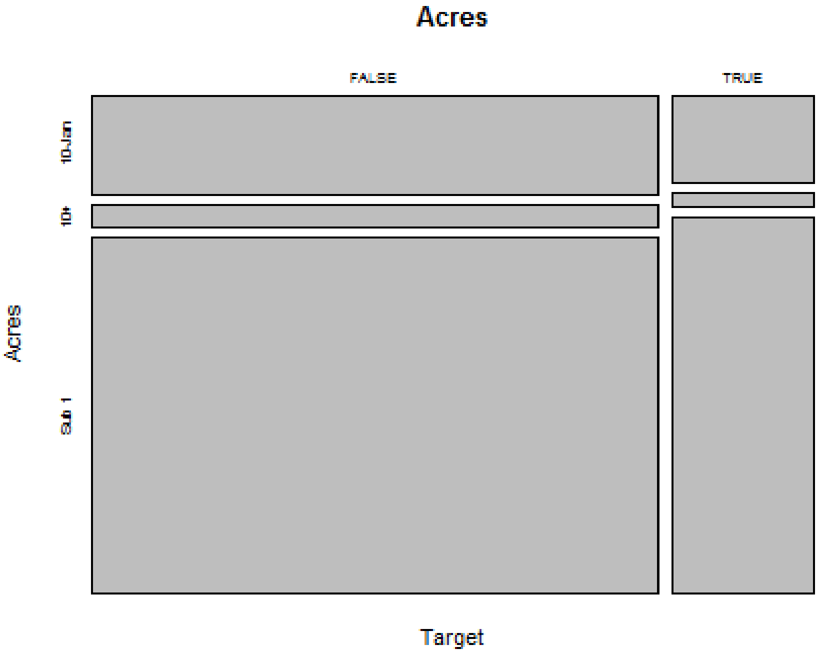

Frequency
0
4000
FALSE
TRUE
15
1990-1999
2010
Year Built

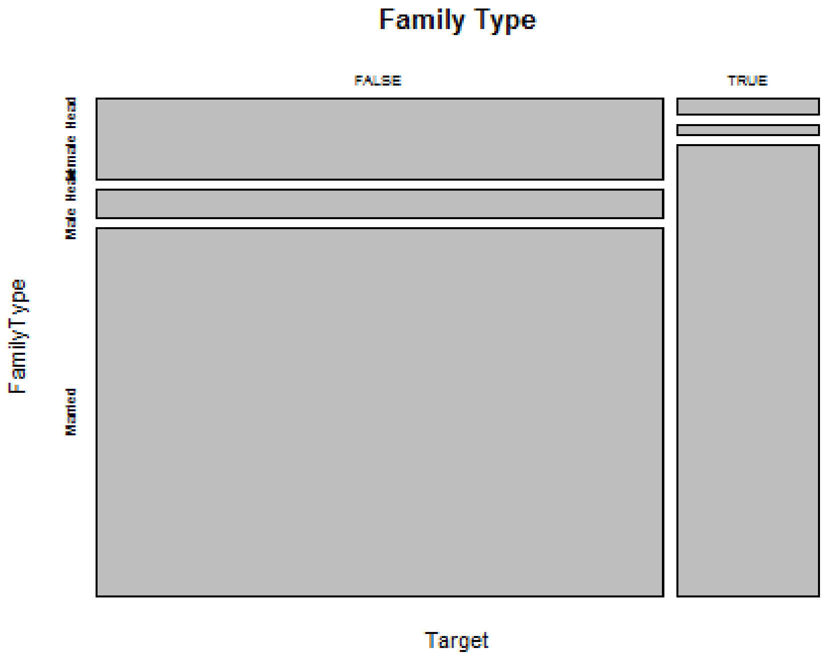

Density
0e+00
FALSE
TRUE
-2000
2000
6000
House Costs

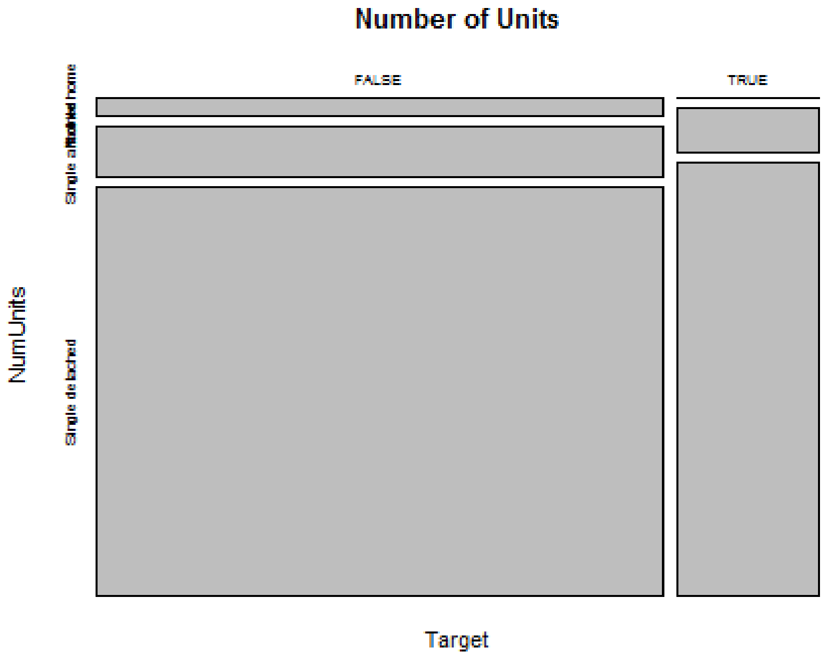

Density
0.000
0.006
FALSE
TRUE
0
200
600
Electric Bill

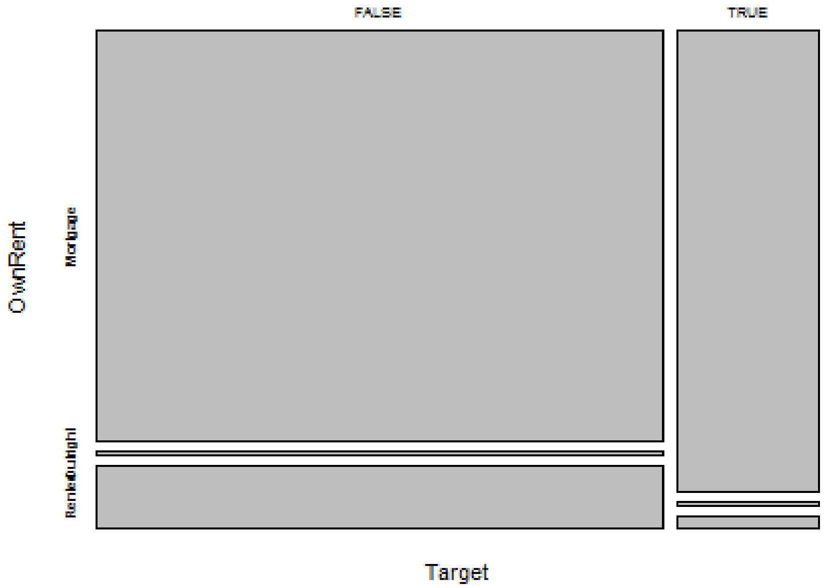

Frequency
0
8000
FALSE
TRUE
Coal
None
Solar
Heating Fuel

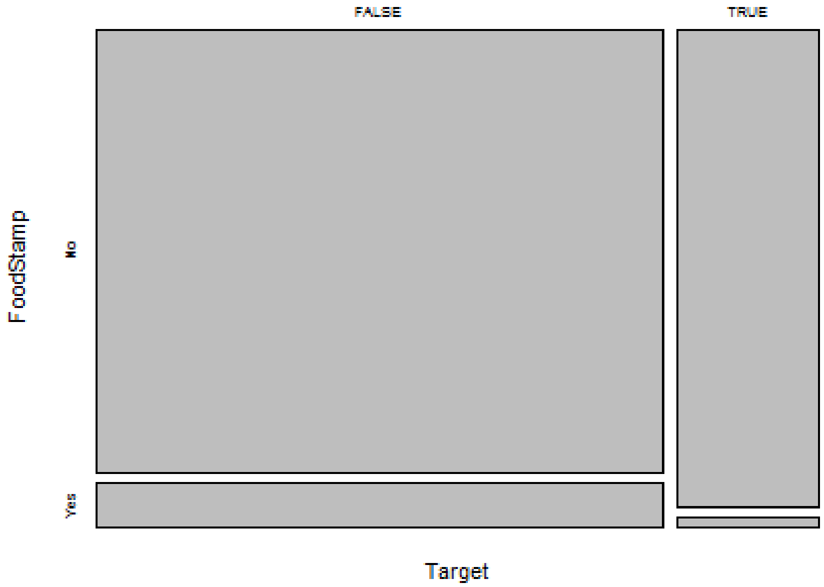

Density
0e+00
FALSE
TRUE
-2000
2000
6000
Insurance

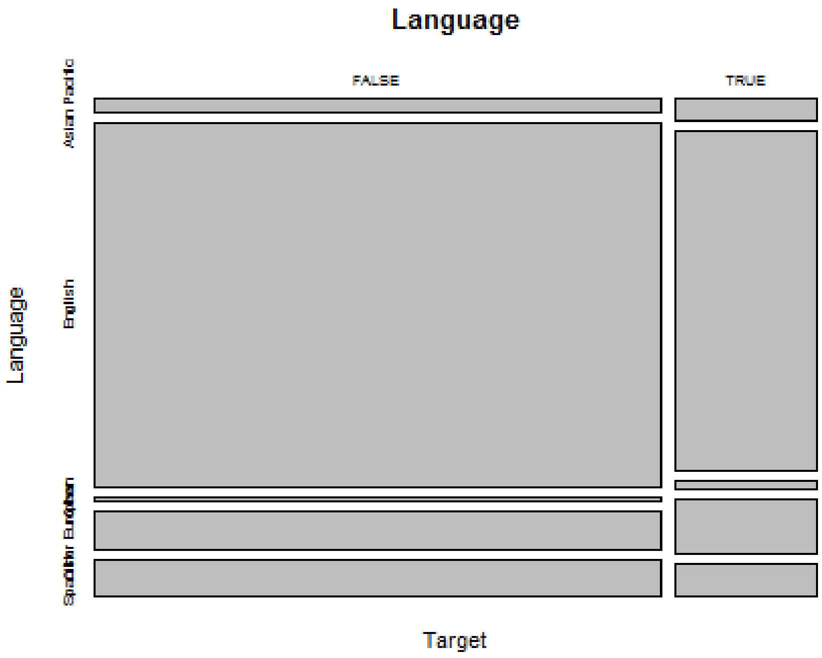

Acres
FALSE
TRUE
Acres
10-Jan
10+
Sub 1
Target


Family Type
FALSE
TRUE
FamilyType
Female Head
Male Head
Married
Target


Number of Units
FALSE
TRUE
NumUnits
Mobile home
Single attached
Single de tached
Target


Own or Rent


Food Stamp



By looking at each visualization, it seemed following variables are significant to target variable.

Number of Bedrooms
Number of People
House Costs
Family Type
Own or Rent

Keeping above understanding in mind, we built logistic regression model. First, we incorporated all variables into logistic regression model and saw those statistical significance. Next we carefully eliminated variables which have less significance. We dropped variables which have p-value less than 0.05. If a categorical variable was split into multiple dummy variables and at least one of those variables had p-value less than 0.05, we dropped all related dummy variables from the model.

Build logistic regression model

```
##
## Call:
## glm(formula = Target ~ Acres + FamilyType + NumBedrooms + NumChildren
+
##     NumPeople + NumRooms + NumUnits + NumVehicles + NumWorkers +
##     OwnRent + YearBuilt + HouseCosts + ElectricBill + FoodStamp +
##     HeatingFuel + Insurance + Language, family = binomial(),
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3124  -0.6095  -0.3793  -0.1266   3.2645
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -1.125e+01  1.788e+00  -6.292 3.13e-10 ***
## Acres10+                 1.005e-01  1.162e-01   0.865 0.387088
## AcresSub 1               2.141e-01  5.601e-02   3.822 0.000133 ***
## FamilyTypeMale Head      3.427e-01  1.498e-01   2.288 0.022119 *
## FamilyTypeMarried        1.299e+00  8.951e-02  14.517  < 2e-16 ***
## NumBedrooms              8.013e-02  2.238e-02   3.581 0.000342 ***
## NumChildren              2.342e-02  2.619e-02   0.894 0.371174
## NumPeople               -1.296e-01  2.370e-02  -5.467 4.57e-08 ***
## NumRooms                 1.090e-01  9.802e-03  11.122  < 2e-16 ***
## NumUnitsSingle attached  2.367e+00  4.576e-01   5.173 2.30e-07 ***
## NumUnitsSingle detached  2.238e+00  4.532e-01   4.938 7.88e-07 ***
## NumVehicles              2.003e-01  2.370e-02   8.453  < 2e-16 ***
## NumWorkers               5.839e-01  3.121e-02  18.709  < 2e-16 ***
## OwnRentOutright          1.416e+00  2.229e-01   6.354 2.10e-10 ***
## OwnRentRented           -2.633e-01  1.057e-01  -2.491 0.012753 *
## YearBuilt1940-1949       1.761e+00  1.687e+00   1.044 0.296677
## YearBuilt1950-1959       1.908e+00  1.687e+00   1.131 0.257857
## YearBuilt1960-1969       1.878e+00  1.687e+00   1.113 0.265630
## YearBuilt1970-1979       1.788e+00  1.687e+00   1.060 0.289237
## YearBuilt1980-1989       2.136e+00  1.687e+00   1.266 0.205652
## YearBuilt1990-1999       2.087e+00  1.687e+00   1.237 0.216147
## YearBuilt2000-2004       2.036e+00  1.688e+00   1.206 0.227838
## YearBuilt2005            2.058e+00  1.696e+00   1.214 0.224938
## YearBuilt2006            1.968e+00  1.698e+00   1.159 0.246439
## YearBuilt2007            2.242e+00  1.700e+00   1.319 0.187243
## YearBuilt2008            1.545e+00  1.714e+00   0.901 0.367374
## YearBuilt2009            1.984e+00  1.717e+00   1.155 0.248019
## YearBuilt2010            2.246e+00  1.731e+00   1.297 0.194460
## YearBuiltBefore 1939     1.765e+00  1.686e+00   1.046 0.295355
## HouseCosts               5.915e-04  1.987e-05  29.763  < 2e-16 ***
## ElectricBill             1.165e-03  1.889e-04   6.169 6.89e-10 ***
## FoodStampYes            -5.517e-01  1.257e-01  -4.388 1.14e-05 ***
## HeatingFuelElectricity   7.160e-01  3.712e-01   1.929 0.053741 .
## HeatingFuelGas           8.921e-01  3.589e-01   2.486 0.012935 *
## HeatingFuelNone         -7.277e-01  1.134e+00  -0.642 0.521167
## HeatingFuelOil           9.072e-01  3.592e-01   2.526 0.011552 *
## HeatingFuelOther         8.087e-01  4.309e-01   1.877 0.060530 .
## HeatingFuelSolar         7.563e-01  1.258e+00   0.601 0.547730
## HeatingFuelWood          3.984e-02  3.767e-01   0.106 0.915755
## Insurance                2.941e-04  2.146e-05  13.702  < 2e-16 ***
## LanguageEnglish         -1.873e-01  9.655e-02  -1.940 0.052364 .
## LanguageOther           -1.874e-01  1.846e-01  -1.016 0.309858
## LanguageOther European  -2.020e-01  1.096e-01  -1.844 0.065168 .
## LanguageSpanish         -4.175e-01  1.162e-01  -3.593 0.000327 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 22808  on 22744  degrees of freedom
## Residual deviance: 17226  on 22701  degrees of freedom
## AIC: 17314
##
## Number of Fisher Scoring iterations: 7
```

As above coefficients report shows, some variables turned out to be less significant with p-value less than 0.05. We removed below variables.

Acres
NumChildren
YearBuilt
HeatingFuel
Language

```
##
## Call:
## glm(formula = Target ~ FamilyType + NumBedrooms + NumPeople +
##     NumRooms + NumUnits + NumVehicles + NumWorkers + OwnRent +
##     HouseCosts + ElectricBill + FoodStamp + Insurance, family = binomi
al(),
##     data = data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3958  -0.6117  -0.3962  -0.1299   3.0992
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -8.529e+00  4.674e-01 -18.248  < 2e-16 ***
## FamilyTypeMale Head      3.170e-01  1.497e-01   2.117 0.034288 *
## FamilyTypeMarried        1.302e+00  8.908e-02  14.620  < 2e-16 ***
## NumBedrooms              8.118e-02  2.199e-02   3.691 0.000223 ***
## NumPeople               -1.200e-01  1.583e-02  -7.578 3.52e-14 ***
## NumRooms                  1.054e-01  9.551e-03  11.031  < 2e-16 ***
## NumUnitsSingle attached  2.389e+00  4.563e-01   5.236 1.64e-07 ***
## NumUnitsSingle detached  2.222e+00  4.521e-01   4.916 8.85e-07 ***
## NumVehicles               1.886e-01  2.213e-02   8.523  < 2e-16 ***
## NumWorkers                5.760e-01  2.936e-02  19.621  < 2e-16 ***
## OwnRentOutright           1.615e+00  2.194e-01   7.358 1.86e-13 ***
## OwnRentRented            -2.739e-01  1.051e-01  -2.606 0.009166 ** 
## HouseCosts                6.265e-04  1.928e-05  32.489  < 2e-16 ***
## ElectricBill              1.212e-03  1.858e-04   6.525 6.82e-11 ***
## FoodStampYes             -5.898e-01  1.238e-01  -4.763 1.90e-06 ***
## Insurance                 2.871e-04  2.124e-05  13.515  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 22808  on 22744  degrees of freedom
## Residual deviance: 17376  on 22729  degrees of freedom
## AIC: 17408
## 
## Number of Fisher Scoring iterations: 7
```

Compared with previous model, AIC and deviance residual increased just a little. We could say there is no negative impact on revised model.

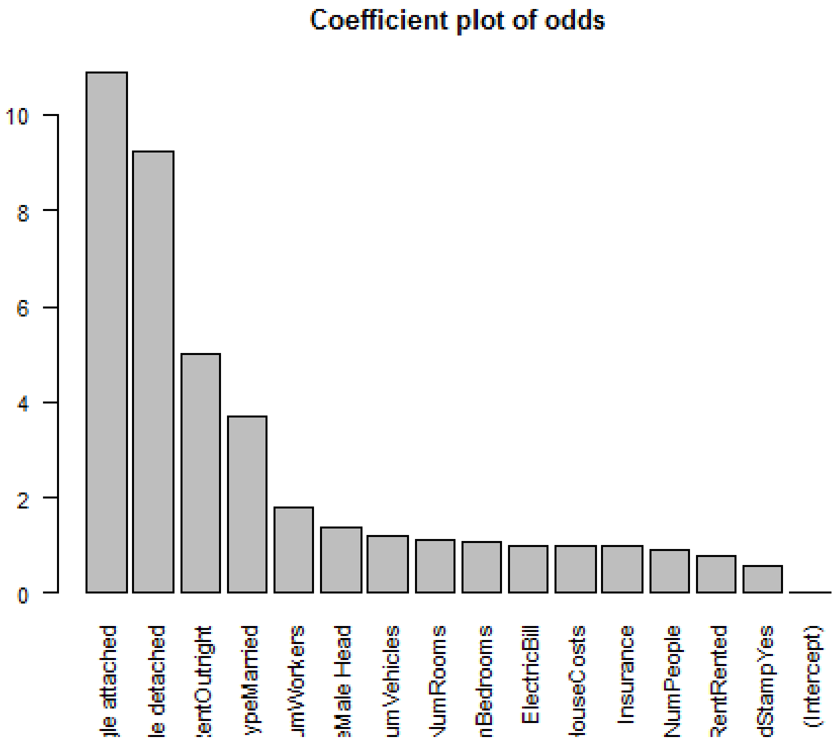
```
##          AIC full var        AIC reduced var Percentage of increase 
##         17313.7592267          17407.9505501              0.5440258

##    Deviance Residual full var Deviance Resudial reduced var 
##                 17225.7592267                 17375.9505501 
##        Percentage of increase 
##                     0.8718996
```

Interpreting coefficient value

We interpreted coefficient value of given model. To understand quantitative impact on response variable easily, we exponentiated coefficient of each variable. An Exponentiated coefficient has a linear relationship with odds ratio of probability (p); p is a probability that each household has higher income larger than 150000.

```
## NumUnitsSingle attached NumUnitsSingle detached         OwnRentOutrigh
t 
##                   10.90                    9.23                    5.0
3 
##       FamilyTypeMarried              NumWorkers     FamilyTypeMale Hea
d 
##                    3.68                    1.78                    1.3
```

```
7
##               NumVehicles                     NumRooms                 NumBedroom
s
##                      1.21                         1.11                        1.0
8
##              ElectricBill                   HouseCosts                   Insuranc
e
##                      1.00                         1.00                        1.0
0
##                 NumPeople                OwnRentRented                 FoodStampYe
s
##                      0.89                         0.76                        0.5
5
##               (Intercept)
##                      0.00
```



Above coefficient values were interpretted as how much a probability of odds ratio increase per unit increased in each variable. For example, having House type "Single Unit attached" will increase odds ratio by 10.90 point which means very high contribution to response variable. As the coefficient plot tells, type of house spec and ownership will be better predictor of household income.

Prediction

At the end of this report, we calculated probability of family income larger than 150000 and save it with training data. We created ROC curve using this prediction (See "acs_ny_predict.xlsx")

```
## Warning in write.csv(data, file = "acs_ny_predict.csv", col.names = TR
UE):
## attempt to set 'col.names' ignored
```