Insurance Logistic Regression Project

Ying Cheng

Northwestern University

Predict 411: Generalized Linear Models

February 14, 2017

Insurance Logistic Regression Project

INTRODUCTION

The purpose of the project is to use logistic Regression to predict whether an auto insurance customer will have a car crash. We will also predict how much money the insurance company will pay to the claim for the customers who has car crash. Before we build the models, we will perform expletory data analysis to have some understanding of the insurance data. Then we will build different regression models and select the best ones.

DATA EXPLORATION AND PREPARATION

The data set consists of 8161 customers from an auto insurance company. The primary target variable is a binominal variable TARGET_FLAG, which indicating whether the customer has car crash. The secondary target variable is a numeric variable TARGET_AMT. If the customer do not a have car crash, then this number should be zero, otherwise, it will be greater than zero. There are 23 potential variables we can use from the insurance data set to predict both TARGET_FLAG and TARGET_AMT.

An Overall View of Response Variables

To have some general understanding of the dependent variable in the insurance data set. Below is the simple summary of the response variable TARGET_FLAG:

The FREQ Procedure

TARGET_FLAG	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	6008	73.62	6008	73.62
1	2153	26.38	8161	100.00

The exploratory data analysis below will focus on target variable TARGET_FLAG. Out of the 8161 sample, there are 2153 customers have car crash, taking 26.38% of the sample.

Numeric Variables

There are 13 independent numeric variables in the insurance data set. Below is an overall view:

The MEANS Procedure

Variable	Label	Mean	Median	Mode	N	N Miss	Minimum	Maximum	Std Dev	Lower 95% CL for Mean	Upper 95% CL for Mean
KIDSDRIV	#Driving Children	0	0	0	8161	0	0	4	1	0	0
AGE	Age	45	45	46	8155	6	16	81	9	45	45
HOMEKIDS	#Children @Home	1	0	0	8161	0	0	5	1	1	1
YOJ	Years on Job	10	11	12	7707	454	0	23	4	10	11
INCOME	Income	61898	54028	0	7716	445	0	367030	47573	60836	62960
HOME_VAL	Home Value	154867	161160	0	7697	464	0	885282	129124	151982	157752
TRAVTIME	Distance to Work	33	33	5	8161	0	5	142	16	33	34
BLUEBOOK	Value of Vehicle	15710	14440	1500	8161	0	1500	69740	8420	15527	15893
TIF	Time in Force	5	4	1	8161	0	1	25	4	5	5
OLDCLAIM	Total Claims(Past 5 Years)	4037	0	0	8161	0	0	57037	8777	3847	4228
CLM_FREQ	#Claims(Past 5 Years)	1	0	0	8161	0	0	5	1	1	1
MVR_PTS	Motor Vehicle Record Points	2	1	0	8161	0	0	13	2	2	2
CAR_AGE	Vehicle Age	8	8	1	7651	510	-3	28	6	8	8

From the table above, AGE, YOJ (Years on Job), INCOME, HOME_VAL (Home Value), and CAR_AGE (Vehicle Age) have missing values. Vehicle Age has error records. It should be never below zero. We see the minimum CAR_AGE is -3. We need to fix this.

Since AGE has only 6 records are missing, we will replace missing values with the median age 45.

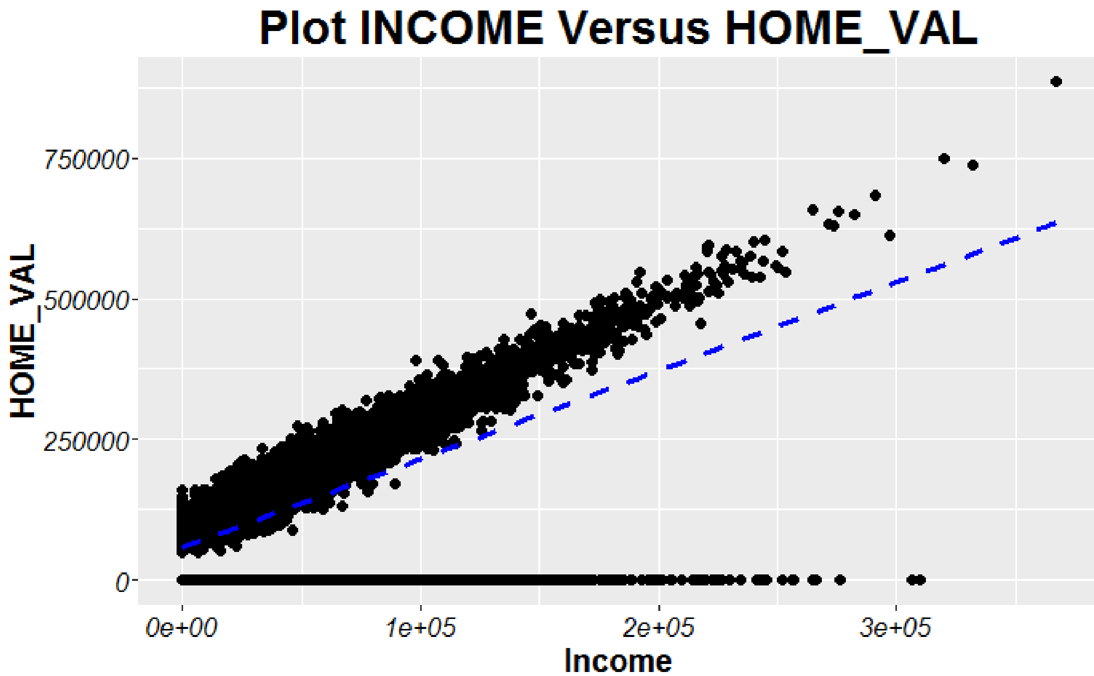
For the other four numeric variables has over 400 missing values, we need to be careful to handle the missing values. We have calculated the correlations for the numeric variables to check whether they have high correlations especially for the ones with missing values. We have found that INCOME and HOME_VAL has the highest correlation, which is 0.58. Followed by HOMEKIDS and AGE.

Correlation for Numeric Variables

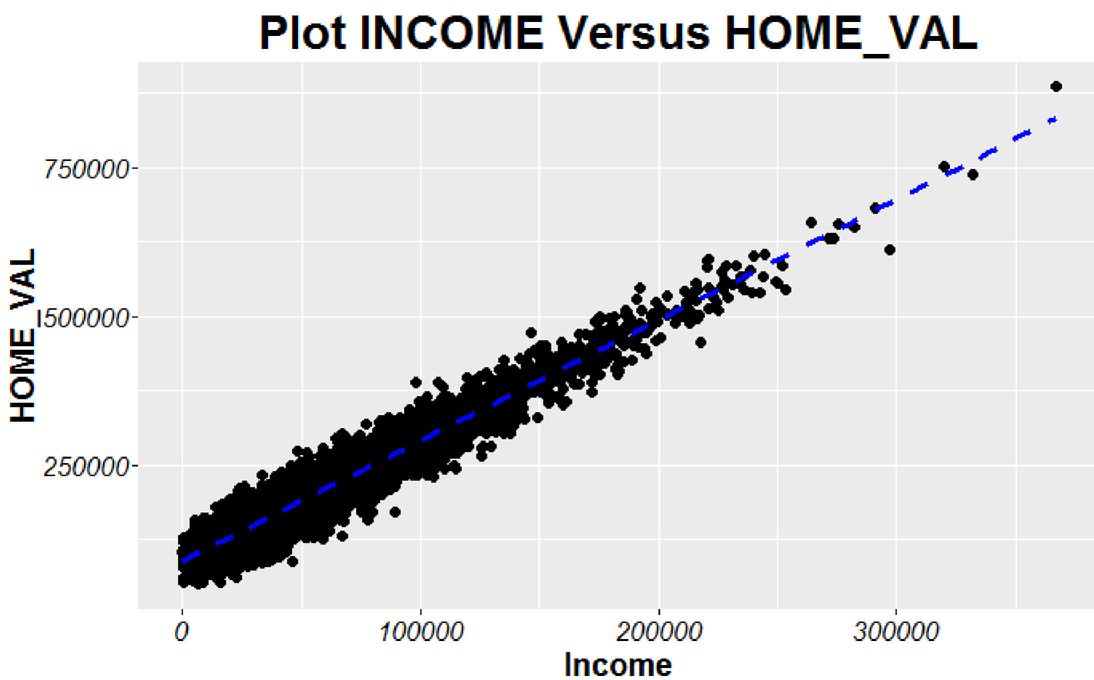
	KIDSDRIV	AGE	HOMEKIDS	YOJ	INCOME	HOME_VAL	TRAVTIME	BLUEBOOK	TIF	OLDCLAIM	CLM_FREQ	MVR_PTS	CAR_AGE
KIDSDRIV	1.00	-0.08	0.46	0.04	-0.05	-0.02	0.01	-0.02	0.00	0.02	0.04	0.05	-0.05
AGE	-0.08	1.00	-0.45	0.14	0.18	0.21	0.01	0.17	0.00	-0.03	-0.02	-0.07	0.18
HOMEKIDS	0.46	-0.45	1.00	0.09	-0.16	-0.11	-0.01	-0.11	0.01	0.03	0.03	0.06	-0.15
YOJ	0.04	0.14	0.09	1.00	0.29	0.27	-0.02	0.14	0.02	0.00	-0.03	-0.04	0.06
INCOME	-0.05	0.18	-0.16	0.29	1.00	0.58	-0.05	0.43	0.00	-0.05	-0.05	-0.06	0.41
HOME_VAL	-0.02	0.21	-0.11	0.27	0.58	1.00	-0.04	0.26	0.00	-0.07	-0.09	-0.09	0.22
TRAVTIME	0.01	0.01	-0.01	-0.02	-0.05	-0.04	1.00	-0.02	-0.01	-0.02	0.01	0.01	-0.04
BLUEBOOK	-0.02	0.17	-0.11	0.14	0.43	0.26	-0.02	1.00	-0.01	-0.03	-0.04	-0.04	0.19
TIF	0.00	0.00	0.01	0.02	0.00	0.00	-0.01	-0.01	1.00	-0.02	-0.02	-0.04	0.01
OLDCLAIM	0.02	-0.03	0.03	0.00	-0.05	-0.07	-0.02	-0.03	-0.02	1.00	0.50	0.26	-0.01
CLM_FREQ	0.04	-0.02	0.03	-0.03	-0.05	-0.09	0.01	-0.04	-0.02	0.50	1.00	0.40	-0.01
MVR_PTS	0.05	-0.07	0.06	-0.04	-0.06	-0.09	0.01	-0.04	-0.04	0.26	0.40	1.00	-0.02
CAR_AGE	-0.05	0.18	-0.15	0.06	0.41	0.22	-0.04	0.19	0.01	-0.01	-0.01	-0.02	1.00

We can not just base on the correlation values we got to see how strong the relationship is. Sometimes the outliers or other factors affect the values. Then we try to make plots to observe the relationship in a more careful way.

INCOME and HOME_VAL should have high correlations as we expected. People has high income have stronger purchasing power so they are more afford for expensive houses. Below is the plot for INCOME and HOME_VAL. As we can see there are not really outliers represented here affect the correlation. We have observed many zero values for both INCOME and HOME_VAL, which strongly affect the correlations we calculated.



After we extract the 0 values from both INCOME and HOME_VAL, we get the correlation 0.96 based on 4865 records. This is a much stronger relationship compare to what we got earlier. Below is the plot for INCOME and HOME_VAL with positive values only.



Only INCOME & HOME_VAL >0 included in above plot.

Insurance Logistic Regression Project

Base on the strong linear regression we got from INCOME and HOME_VAL. We can use this relationship to fill in the missing values for either of this. We run a simple linear regression for INCOME and HOME_VAL. Below is the fitted simple fitted linear regression summary.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	8.848563E12	8.848563E12	57006.7	<.0001
Error	4863	7.54833E11	155219618		
Corrected Total	4864	9.603396E12			

Root MSE	12459	R-Square	0.9214
Dependent Mean	68771	Adj R-Sq	0.9214
Coeff Var	18.11619		

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	-34863	469.36664	-74.28	<.0001	0
HOME_VAL	Home Value	1	0.45485	0.00191	238.76	<.0001	1.00000

The model has very high R-Square, which indicating a good fit.

INCOME = -34863 + 0.45485 * **HOME_VAL**

If HOME_VAL is not missing, we will use above SLR to filling the missing INCOME in the data set. We also have to consider that, since INCOME has to be at least 0, which indicating HOME_VAL has to be more than 76648 to apply this model.

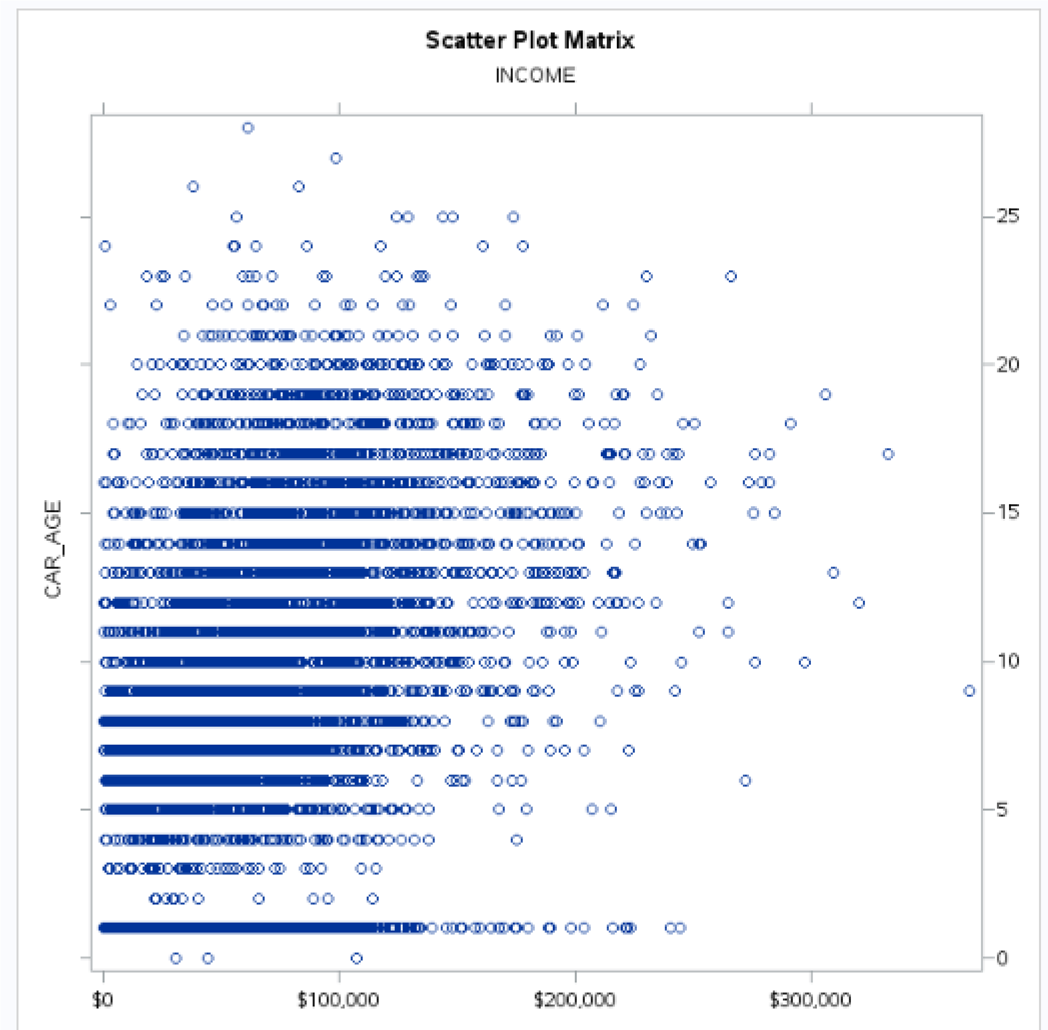
However, we still have some records have both missing values of INCOME and HOME_VAL. In reality, we know that income and job type should have some relationship. Below is the summary mean INCOME for different job categories.

Mean Income by Job Category

Job Category	Count	Mean Income
Missing	502	$ 118,853
Clerical	1198	$ 33,861
Doctor	238	$ 128,680
Home Maker	598	$ 12,073
Lawyer	799	$ 88,305
Manager	938	$ 87,462
Professional	1057	$ 76,593
Student	659	$ 6,310
z_Blue Collar	1727	$ 58,957

In theory, people who do not have a job should have YOJ 0. We have observed many 0 income in the dataset. For all the YOJ with 0 records, we can also find out that their income is also zero. For this reason, if the income is 0, we will replace the missing YOJ records with 0, other missing YOJ will be replaced by the median value 11.

CAR_AGE also has missing values. As we refer to the correlation table we got earlier, we find INCOME has the highest correlations with CAR_AGE. It might be high income customers more able to purchase new cars. However, the correlation is only 0.41. Even when we look the positive CAR_AGE and income observations, the correlation we get is 0.42547. We won't use a linear regression to replace CAR_AGE missing values. We has made a plot for this. Instead, we will just fill missing values with 8 (mean or median value).



The MEANS Procedure

Variable	Label	Mean	Median	Mode	N	N Miss	Minimum	Maximum	Std Dev	Lower 95% CL for Mean	Upper 95% CL for Mean
AGE	Age	45	45	46	8154	6	16	81	9	45	45
IMP_AGE		45	45	46	8160	0	16	81	9	45	45
YOJ	Years on Job	10	11	12	7706	454	0	23	4	10	11
IMP_YOJ		10	11	11	8160	0	0	23	4	10	11
INCOME	Income	61900	54028	0	7715	445	0	367030	47576	60838	62962
IMP_INCOME		61568	53916	0	8160	0	0	367030	47249	60542	62593
CAR_AGE	Vehicle Age	8	8	1	7650	510	0	28	6	8	8
IMP_CAR_AGE		8	8	1	8160	0	0	28	6	8	8

Categorical Variables

We have 10 categorical variables.

- **2 levels:** CAR_USE, MSTATUS (Marital Status), PARENT1(Single Parent), RED_CAR, REVOKED, SEX, URBANCITY
- **3 or more levels:** CAR_TYPE, EDUCATION, JOB,

Among those ten categorical variables, only JOB has missing values, which represents 6.45% (526 records) of the sample.

Job Category

JOB	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	526	6.45	526	6.45
z_Blue Collar	1825	22.37	2351	28.81
Clerical	1271	15.58	3622	44.39
Professional	1116	13.68	4738	58.06
Manager	988	12.11	5726	70.17
Lawyer	835	10.23	6561	80.40
Student	712	8.73	7273	89.13
Home Maker	641	7.86	7914	96.99
Doctor	246	3.01	8160	100.00

We have make frequency tables for each of the categorical variables. Since we will do similar analysis with response variable, the details will not be showing here. It makes more sense to analyze with TARGET_FLAG to have ideas how good the variable can separate out whether the customer will has a car crash.

Insurance Logistic Regression Project

Independent Variables With Response Variables

The graphs below are produced with WEKA. It is easy for us to have an general ideas how each variable can separate the response variable TARGET_FLAG.

KIDSDRIV (Driving Children): It is numeric, however, the majority customers has no children. It also seems in the graphic people has 2 or more children less likely to has car crash.

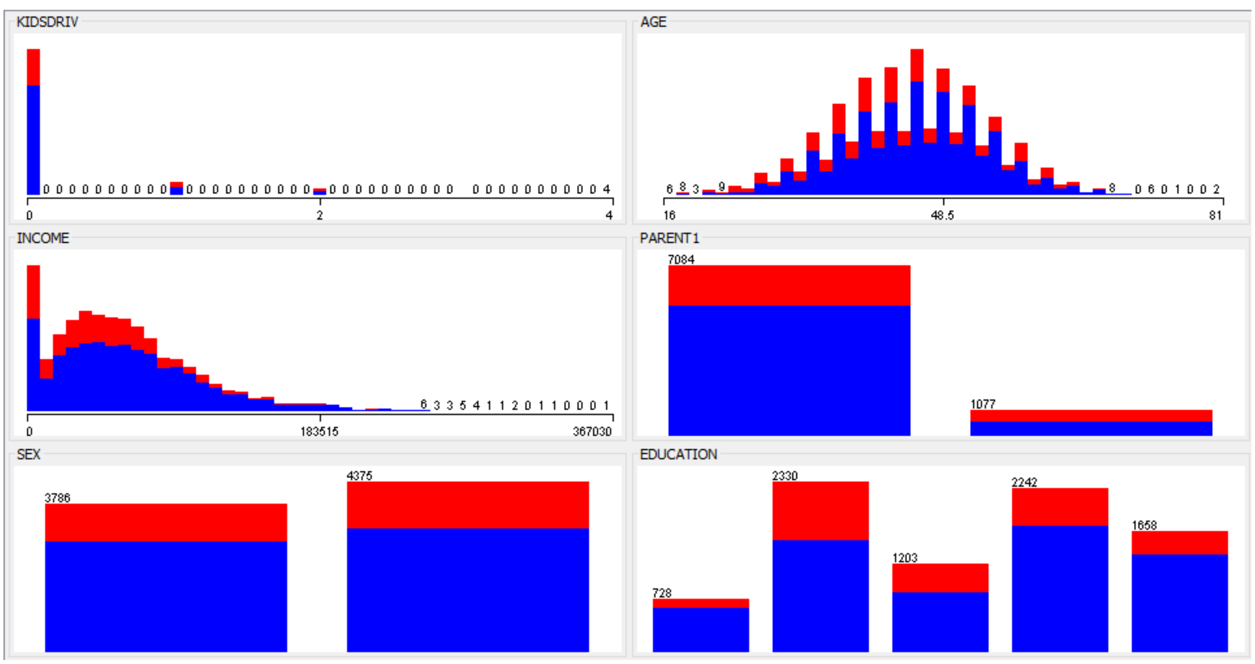
AGE: Looks normally distributed. Young age more likely has car crash.

INCOME: Not normally distributed, has a very long right tail. It seems low income more likely has car crash.

PARENT1 (Single Parent): A small percentage of customers is a single parent. They has a much higher probability has car crash comparing to non-single parent.

SEX : Not looks predictive.

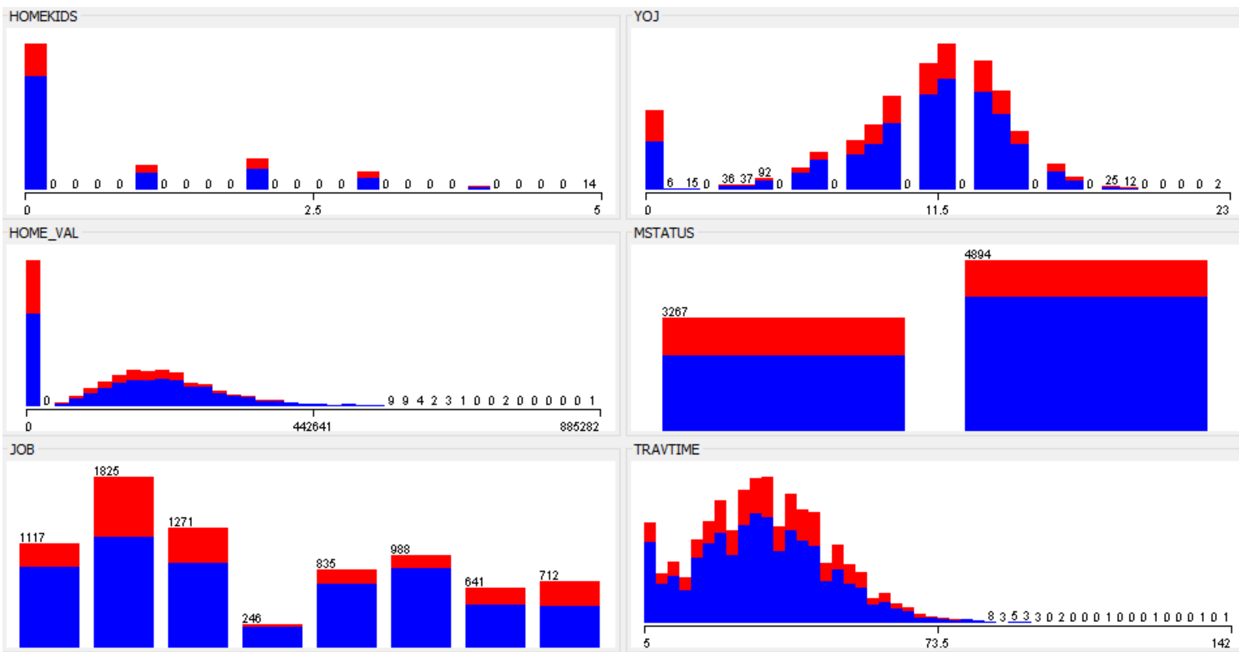
Education: We can observe different proportion of in car crash in different education level. This will help us to regroup EDUCATION.

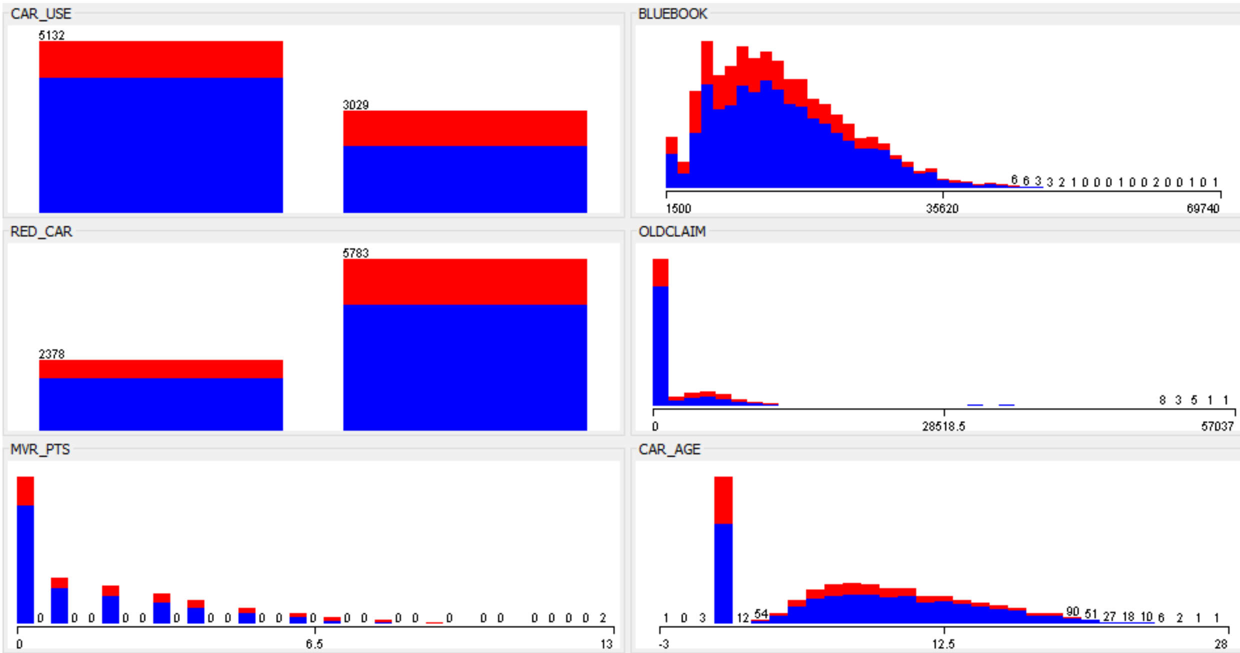


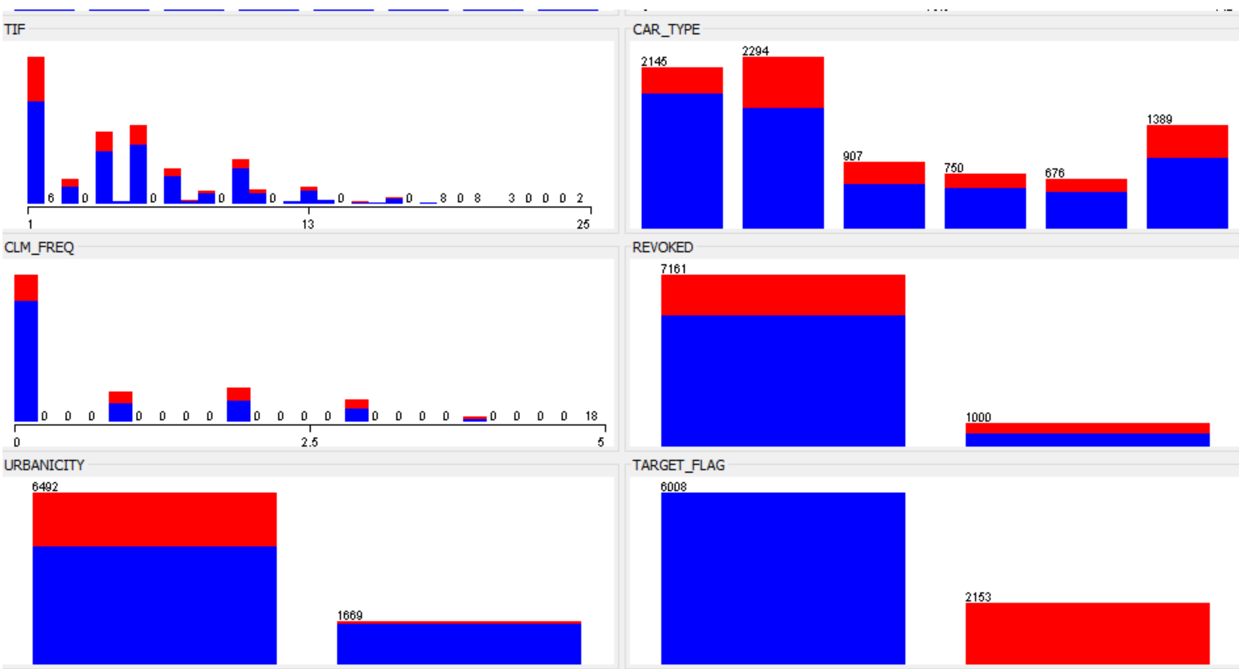
HOMEKIDS (Children @Home): Similar to KIDSDRIV discussed earlier. We might it to categorical variable.

YOJ (Years on Job), **HOME_VAL** (Home Value), **OLDCLAIM** (Total Claims-Past 5 Years), and **CAR_AGE** (Vehicle Age) : They look normal distributed except the large observations falling at zero. We might need to create a new variable in each indicating the ones has no job, no house, no old claims or just purchase a new car recently.

TRAVTIME (Distance to Work) and **BLUEBOOK** (Value of Vehicle): They not look normally distributed, we need to transform.




CAR_USE
5132
3029
BLUEBOOK
1500
35620
69740
RED_CAR
2378
5783
OLDCLAIM
0
28518.5
57037
MVR_PTS
0
6.5
13
CAR_AGE
-3
12.5
28


TIF
1
13
25
CAR_TYPE
2145
2294
907
750
676
1389
CLM_FREQ
0
2.5
5
REVOKED
7161
1000
URBANICITY
6492
1669
TARGET_FLAG
6008
2153

The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of KIDSDRIV by TARGET_FLAG

KIDSDRIV(#Driving Children)	TARGET_FLAG 0	TARGET_FLAG 1	Total
0	5407 66.26 75.32 90.00	1772 21.72 24.68 82.34	7179 87.98
1	400 4.90 62.89 6.66	236 2.89 37.11 10.97	636 7.79
2	168 2.06 60.22 2.80	111 1.36 39.78 5.16	279 3.42
3	31 0.38 50.00 0.52	31 0.38 50.00 1.44	62 0.76
4	2 0.02 50.00 0.03	2 0.02 50.00 0.09	4 0.05
Total	6008 73.63	2152 26.37	8160 100.00

Customers do not have driving children has car crash possibility lower than overall. They take almost 88% of all the customers. The other groups has a much higher car crash possibility but they take a very small portion. This variable will be grouped into whether customers have driving children.

In a similar way for HOMEKIDS (#Children @Home), we also find out that customer with no children at home has lower risk. A new variable created indicating whether the customer has children at home.

The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of CLM_FREQ by TARGET_FLAG

CLM_FREQ(#Claims(Past 5 Years))	TARGET_FLAG 0	TARGET_FLAG 1	Total
0	4111 50.37 82.07 68.43	898 11.00 17.93 41.71	5009 61.38
1	612 7.50 61.38 10.19	385 4.72 38.62 17.88	997 12.22
2	702 8.60 59.95 11.68	469 5.75 40.05 21.78	1171 14.35
3	462 5.66 59.54 7.69	314 3.85 40.46 14.58	776 9.51
4	110 1.35 57.89 1.83	80 0.98 42.11 3.72	190 2.33
5	11 0.13 61.11 0.18	7 0.09 38.89 0.33	18 0.22
Total	6008 73.62	2153 26.38	8161 100.00

Customers have no claims in the past 5 years have 18% possibility in car crash. Once they did had claims, no matter how many they are, customers have about 40% chances in car crash. We will create a new dummy variable indicating whether the customer has a claims in the past 5 years.

The FREQ Procedure

Frequency Percent Row Pct Col Pct

Table of CAR_TYPE by TARGET_FLAG

CAR_TYPE(Type of Car)	TARGET_FLAG 0	TARGET_FLAG 1	Total
Minivan	1796 22.01 83.73 29.89	349 4.28 16.27 16.22	2145 26.29
Panel Truck	498 6.10 73.67 8.29	178 2.18 26.33 8.27	676 8.28
Pickup	946 11.59 68.16 15.75	442 5.42 31.84 20.54	1388 17.01
Sports Car	603 7.39 66.48 10.04	304 3.73 33.52 14.13	907 11.12
Van	549 6.73 73.20 9.14	201 2.46 26.80 9.34	750 9.19
z_SUV	1616 19.80 70.44 26.90	678 8.31 29.56 31.51	2294 28.11
Total	6008 73.63	2152 26.37	8160 100.00

From above car type summary with target variable. It is appears that minivan has much lower risk. Pickup and sports car have much higher risk. We will create two new dummy variables based on CAR_TYPE.

1. CAR_TYPE_MV: Whether the customer use a minivan
2. CAR_TYPE_PS : Whether the customer use a Pickup or a Sport car

Insurance Logistic Regression Project

The FREQ Procedure

Frequency Percent Row Pct Col Pct

Table of EDUCATION by TARGET_FLAG

EDUCATION(Max Education Level)	TARGET_FLAG 0	TARGET_FLAG 1	Total
<High School	818 10.02 68.00 13.62	385 4.72 32.00 17.89	1203 14.74
Bachelors	1719 21.07 76.71 28.61	522 6.40 23.29 24.26	2241 27.46
Masters	1331 16.31 80.28 22.15	327 4.01 19.72 15.20	1658 20.32
PhD	603 7.39 82.83 10.04	125 1.53 17.17 5.81	728 8.92
z_High School	1537 18.84 65.97 25.58	793 9.72 34.03 36.85	2330 28.55
Total	6008 73.63	2152 26.37	8160 100.00

From above education type summary with target variable. It appears that phD has much lower risk. But only takes 8.92% of the sample. Customers with Masters also have much lower risk than overall. Customers with high school degree or below have over 30% chances having car crashes. Two dummy variables will be created based on this summary.

1. HighEducation: Whether the customer has a Masters or phD
2. LowEducation : Whether the customer's education is high school or below

Insurance Logistic Regression Project

The FREQ Procedure

Frequency Percent Row Pct Col Pct

Table of JOB by TARGET_FLAG			
	TARGET_FLAG		
JOB(Job Category)	0	1	Total
	390 4.78 74.14 6.49	136 1.67 25.86 6.32	526 6.45
Clerical	900 11.03 70.81 14.98	371 4.55 29.19 17.24	1271 15.58
Doctor	217 2.66 88.21 3.61	29 0.36 11.79 1.35	246 3.01
Home Maker	461 5.65 71.92 7.67	180 2.21 28.08 8.36	641 7.86
Lawyer	682 8.36 81.68 11.35	153 1.88 18.32 7.11	835 10.23
Manager	851 10.43 86.13 14.16	137 1.68 13.87 6.37	988 12.11
Professional	870 10.66 77.96 14.48	246 3.01 22.04 11.43	1116 13.68
Student	446 5.47 62.64 7.42	266 3.26 37.36 12.36	712 8.73
z_Blue Collar	1191 14.60 65.26 19.82	634 7.77 34.74 29.46	1825 22.37
Total	6008 73.63	2152 26.37	8160 100.00

From above job type summary with target variable. It appears that Doctor and Manager have very low risk. Lawyer also has much lower risk compare to the overall. Student and Z_Blue Collar have very high risk. Two dummy variables will be created based on this summary.

1. JOB_WHITE_COLLAR: Whether the customer's job is Doctor, Lawyer, or Manager
2. JOB_BLUE_STUDENT: Whether the customer is blue collar or a student

BUILD MODELS AND SELECT MODELS

MODEL 1- Forward Selection

In the clean the data set, we have fixed the missing values and transformed the categorical variables. Below is the best logistic regression model returned by R with forward selection based on AIC. The first model has AIC value 7380. However, we notice that many parameters estimated are not significant. We might need to try to delete some based on the returned p values and AIC of the model.

```
Call:
glm(formula = TARGET_FLAG ~ MVR_PTS + No_CLM_FREQ + No_HOME +
    log_INCOME + No_Income + log_OLDCLAIM + No_HOMEKIDS + No_KIDSDRIV +
    log_BLUEBOOK + IMP_AGE + log_CAR_AGE + NewCar + TIF + IMP_YOJ +
    log_TRAVTIME + CAR_TYPE_MV + CAR_TYPE_PS + CAR_USE_C + HighEducation +
    LowEducation + JOB_WHITE_COLLAR + JOB_BLUE_STUDENT + MSTATUS_Y +
    PARENT_Y + RED_CAR_Y + REVOKED_Y + SEX_M + URBANICITY_HU,
    family = binomial(), data = CleanData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4204  -0.7153  -0.4036   0.6466   3.1376

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       1.742837   0.806387   2.161 0.030673 *
MVR_PTS           0.104420   0.014018   7.449 9.41e-14 ***
No_CLM_FREQ      -1.845063   0.402155  -4.588 4.48e-06 ***
No_HOME           0.229703   0.078549   2.924 0.003452 **
log_INCOME       -0.109407   0.038164  -2.867 0.004147 **
No_Income        -0.496924   0.423531  -1.173 0.240681
log_OLDCLAIM     -0.162199   0.045343  -3.577 0.000347 ***
No_HOMEKIDS      -0.221959   0.099966  -2.220 0.026395 *
No_KIDSDRIV      -0.569340   0.098065  -5.806 6.41e-09 ***
log_BLUEBOOK     -0.364761   0.051482  -7.085 1.39e-12 ***
IMP_AGE           0.001254   0.004130   0.304 0.761355
log_CAR_AGE       0.143987   0.122373   1.177 0.239347
NewCar            0.280964   0.193848   1.449 0.147225
TIF              -0.053360   0.007311  -7.299 2.90e-13 ***
IMP_YOJ           0.011604   0.011501   1.009 0.312992
log_TRAVTIME      0.434202   0.054151   8.018 1.07e-15 ***
CAR_TYPE_MV      -0.617948   0.083193  -7.428 1.10e-13 ***
CAR_TYPE_PS       0.054625   0.068710   0.795 0.426611
CAR_USE_C         0.602150   0.078190   7.701 1.35e-14 ***
HighEducation     0.011673   0.099883   0.117 0.906964
LowEducation      0.512870   0.084360   6.080 1.21e-09 ***
JOB_WHITE_COLLAR -0.476942   0.091075  -5.237 1.63e-07 ***
JOB_BLUE_STUDENT  0.073161   0.079180   0.924 0.355500
MSTATUS_Y        -0.586447   0.087578  -6.696 2.14e-11 ***
PARENT_Y          0.248031   0.120338   2.061 0.039291 *
RED_CAR_Y        -0.025109   0.086115  -0.292 0.770610
REVOKED_Y         0.876554   0.088206   9.938  < 2e-16 ***
SEX_M            -0.074836   0.084168  -0.889 0.373937
URBANICITY_HU     2.308090   0.112604  20.497  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Insurance Logistic Regression Project

The logistic regression model allows to establish a relationship between a binary outcome YHAT (Profitability=0, Profitability=1) and group of predictor variables. Then the logistic regression of YHAT on the predictive variables via maximum likelihood method of the following equation:

YHAT =	1.742837	+	(intercept)
	0.104420*	**MVR_PTS** +	(Motor Vehicle Record Points)
	-1.845063*	**No_CLM_FREQ** +	(No claim in the past 5 years)
	0.229703	**NO_HOME** +	(Not own a home)
	-0.109407*	**log_INCOME** +	(income +1 in log)
	-0.496924	**No_INCOME** +	(no income)
	-0.162199*	**log_OLDCLAIM** +	(past 5 year claim +1 in log)
	-0.221959*	**No_HOMEKIDS** +	(no kids at home)
	-0.569340*	**No_KIDSDRIV** +	(no driving children)
	-0.364761*	**log_BLUEBOOK** +	(Value of Vehicle + 1 in log)
	0.001254*	**IMP_AGE** +	(Age of Driver)
	0.1439873*	**log_CAR_AGE** +	(Vehicle Age +1 in log)
	0.280964	**NewCar** +	(whether the car within 3 years)
	-0.053360*	**TIF** +	(Time in Force)
	0.011604*	**IMP_YOJ** +	(Years on Job)
	0.434202*	**log_TRAVTIME** +	(Distance to Work +1 in log)
	-0.617948*	**CAR_TYPE_MV** +	(minivan)
	0.054625*	**CAR_TYPE_PS** +	(Pickup or a Sport car)
	0.602150*	**CAR_USE_C** +	(Commercial vehicles)
	0.011673*	**HighEducation** +	(Masters or phD)
	0.512870*	**LowEducation** +	(high school or below)
	-0.476942*	**JOB_WHITE_COLLAR** +	(Whether Doctor, Lawyer, or Manager)
	0.073161*	**JOB_BLUE_STUDENT** +	(whether blue collar or a student)
	-0.586447*	**MSTATUS_Y** +	(whether married)
	0.248031*	**PARENT_Y** +	(whether a single parent)
	-0.025109*	**RED_CAR_Y** +	(whether a red car)

Insurance Logistic Regression Project

0.876554*	**REVOKED_Y** +	(License Revoked (Past 7 Years))
-0.074836*	**SEX_M** +	(whether Gender is male)
2.308090*	**URBANICITY_HU**	(Highly Urban/ Urban)

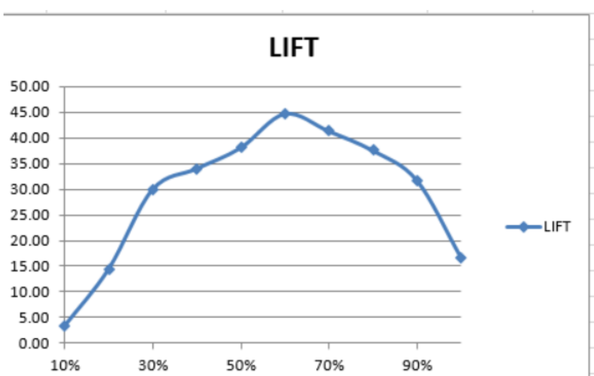
Let P be the probability of YHAT to be 1, P=prob(Profitability=1). It is the probability the customer will have a car crash. In terms of probabilities, the equation above is translated into:
P=exp(YHAT)/(1+exp(YHAT))

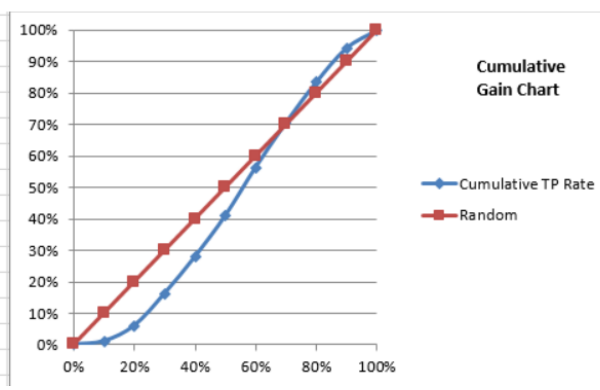
We have tried 10 cut off points from 1.0 to 0.0 with 0.1 reduction at each time. The maximum lift is at cut off 0.4 with 44.69% lift. The maximum gain is 4% at cut off 0.1.

GROUP	CUTOFF	OBS	True Positives (TP)	Total OBS	Total TP	Random TP	TP Rate	Random Rate	LIFT	Cumulative TP Rate	Random TP Rate	Difference
0	1.00000	0	0	0	0	0	0	0		0%	0%	0%
1	0.9	28	24	28	24	7	1%	0%	3.25	1%	10%	-9%
2	0.8	135	106	163	130	7	5%	0%	14.35	6%	20%	-14%
3	0.7	290	221	453	351	7	10%	0%	29.93	16%	30%	-14%
4	0.6	371	251	824	602	7	12%	0%	33.99	28%	40%	-12%
5	0.5	519	281	1343	883	7	13%	0%	38.05	41%	50%	-9%
6	0.4	699	330	2042	1213	7	15%	0%	44.69	56%	60%	-4%
7	0.3	918	305	2960	1518	7	14%	0%	41.30	71%	70%	1%
8	0.2	1143	277	4103	1795	7	13%	0%	37.51	83%	80%	3%
9	0.1	1583	234	5686	2029	7	11%	0%	31.69	94%	90%	4%
10	0	2474	123	8160	2152	7	6%	0%	16.66	100%	100%	0%

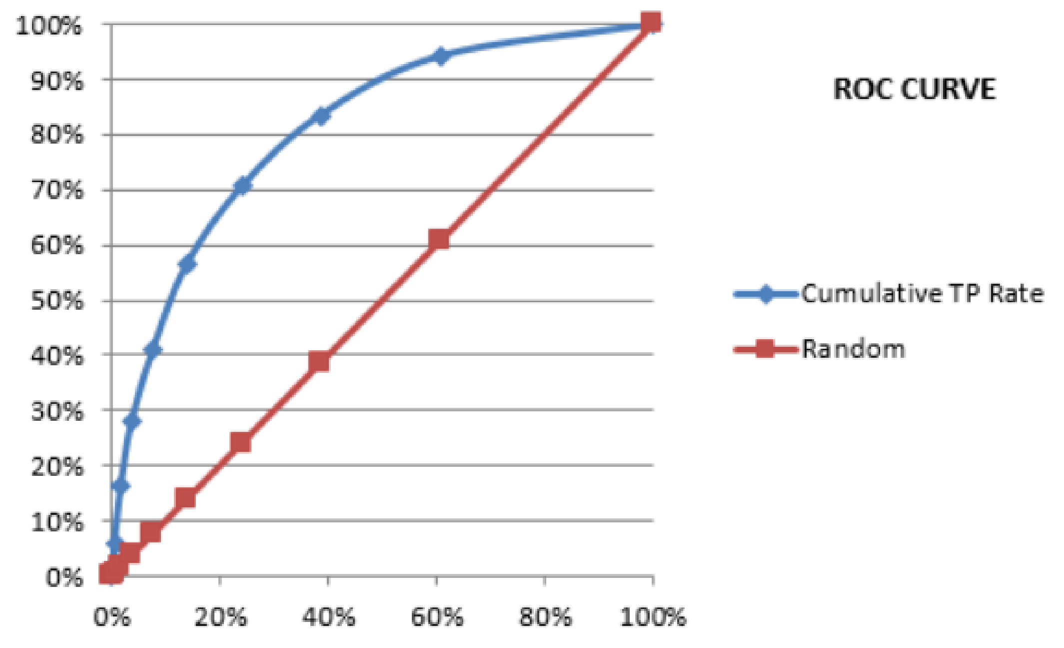
LIFT = 44.69

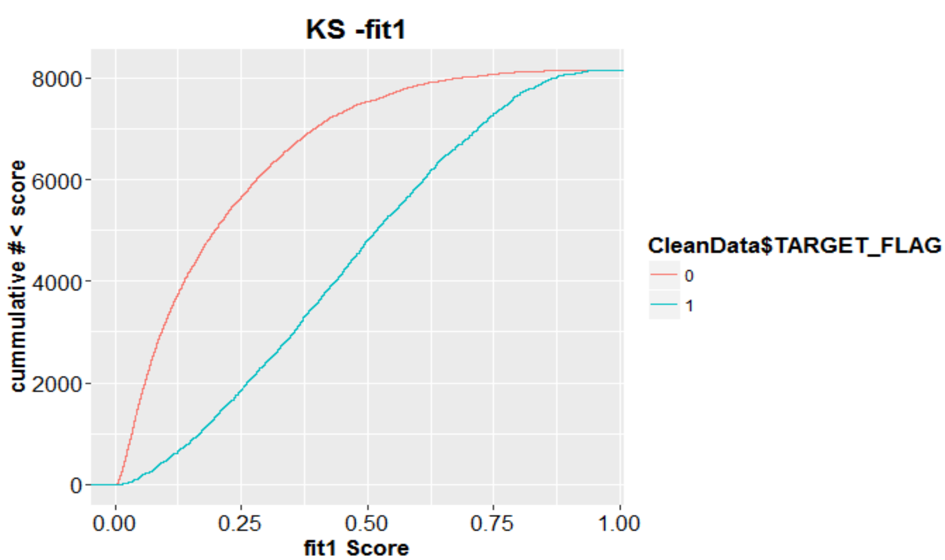
GAIN = 4%





The ROC curve showing model 1 is better than the random model, as it has more area over the curve than the triangle line. The KS graph also showing model 1 differentiate car crash well.


ROC CURVE
Cumulative TP Rate
Random
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
0%
20%
40%
60%
80%
100%


KS -fit1
cummulative # < score
8000
6000
4000
2000
0
0.00
0.25
0.50
0.75
1.00
fit1 Score
CleanData$TARGET_FLAG
0
1

GROUP	CUTOFF	True Positives (TP)	False Postive (FP)	Total TP	Total FP	TP Rate	FP Rate	Cumulative TP Rate	Cumulative FP Rate	Difference
0	1.00000	0	0	0	0	0	0	0%	0%	0%
1	0.78811	24	4	24	4	1%	0%	1%	0%	1%
2	0.38716	106	29	130	33	5%	0%	6%	1%	5%
3	0.11277	221	69	351	102	10%	1%	16%	2%	15%
4	0.06775	251	120	602	222	12%	2%	28%	4%	24%
5	0.05238	281	238	883	460	13%	4%	41%	8%	33%
6	0.03777	330	369	1213	829	15%	6%	56%	14%	43%
7	0.02718	305	613	1518	1442	14%	10%	71%	24%	47%
8	0.01581	277	866	1795	2308	13%	14%	83%	38%	45%
9	0.00375	234	1349	2029	3657	11%	22%	94%	61%	33%
10	0.00013	123	2351	2152	6008	6%	39%	100%	100%	0%
									KS =	47%

MODEL 2- Adjusted Based on Model1

As we have observed some parameters estimated in mode 1 are not very significant. We try to delete those variables one by one based on the p-value until all the parameters are significant. Below is what final returned:

Insurance Logistic Regression Project

```
Call:
glm(formula = TARGET_FLAG ~ MVR_PTS + No_CLM_FREQ + No_HOME +
    log_INCOME + log_OLDCLAIM + No_HOMEKIDS + No_KIDSDRIV + log_BLUEBOOK +
    TIF + log_TRAVTIME + CAR_TYPE_MV + CAR_USE_C + LowEducation +
    JOB_WHITE_COLLAR + MSTATUS_Y + PARENT_Y + REVOKED_Y + URBANICITY_HU,
    family = binomial(), data = CleanData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4716  -0.7175  -0.4024   0.6439   3.1250

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       1.821353   0.644024   2.828 0.004683 **
MVR_PTS           0.104026   0.013997   7.432 1.07e-13 ***
No_CLM_FREQ      -1.820780   0.401462  -4.535 5.75e-06 ***
No_HOME           0.248516   0.077117   3.223 0.001270 **
log_INCOME       -0.054345   0.010222  -5.316 1.06e-07 ***
log_OLDCLAIM     -0.159289   0.045262  -3.519 0.000433 ***
No_HOMEKIDS      -0.234877   0.087916  -2.672 0.007549 **
No_KIDSDRIV      -0.572909   0.095616  -5.992 2.08e-09 ***
log_BLUEBOOK     -0.381609   0.048903  -7.803 6.03e-15 ***
TIF              -0.053123   0.007295  -7.282 3.28e-13 ***
log_TRAVTIME      0.435091   0.054107   8.041 8.89e-16 ***
CAR_TYPE_MV      -0.662794   0.074034  -8.953  < 2e-16 ***
CAR_USE_C         0.589742   0.064763   9.106  < 2e-16 ***
LowEducation      0.547460   0.065711   8.331  < 2e-16 ***
JOB_WHITE_COLLAR -0.480619   0.084475  -5.690 1.27e-08 ***
MSTATUS_Y        -0.560750   0.086283  -6.499 8.09e-11 ***
PARENT_Y          0.240783   0.120001   2.007 0.044802 *
REVOKED_Y         0.870447   0.088054   9.885  < 2e-16 ***
URBANICITY_HU     2.299562   0.112398  20.459  < 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: 9415.3  on 8159  degrees of freedom
Residual deviance: 7331.9  on 8141  degrees of freedom
AIC: 7369.9

Number of Fisher Scoring iterations: 5
```

YHAT =	1.821353		+	(intercept)
	0.104026*	**MVR_PTS**	+	(Motor Vehicle Record Points)
	-1.820780*	**No_CLM_FREQ**	+	(No claim in the past 5 years)
	0.248516	**NO_HOME**	+	(Not own a home)
	-0.054345*	**log_INCOME**	+	(income +1 in log)
	-0.159289*	**log_OLDCLAIM**	+	(past 5 year claim +1 in log)
	-0.234877*	**No_HOMEKIDS**	+	(no kids at home)

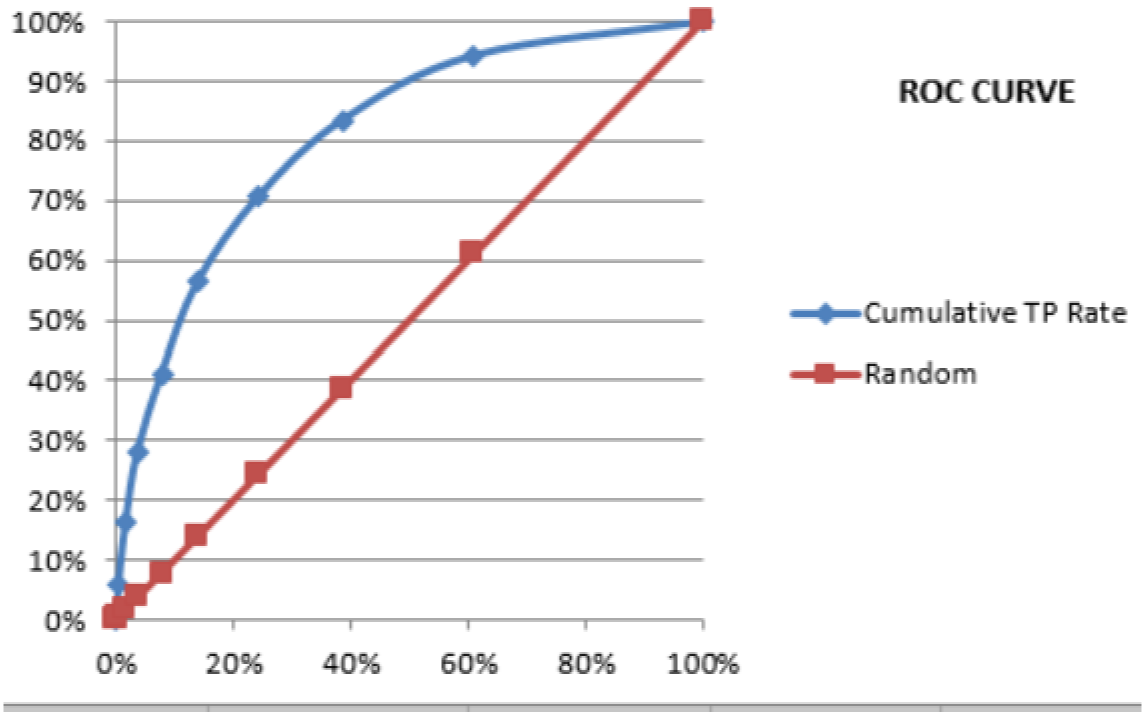
-0.572909*	**No_KIDSDRIV** +	(no driving children)
-0.381609*	**log_BLUEBOOK** +	(Value of Vehicle + 1 in log)
-0.053123*	**TIF** +	(Time in Force)
0.435091	**log_TRAVTIME** +	(Distance to Work +1 in log)
-0.662794*	**CAR_TYPE_MV** +	(minivan)
0.589742*	**CAR_USE_C** +	(Commercial vehicles)
0.547460*	**LowEducation** +	(high school or below)
-0.480619*	**JOB_WHITE_COLLAR** +	(Whether Doctor, Lawyer, or Manager)
-0.560750*	**MSTATUS_Y** +	(whether married)
0.240783*	**PARENT_Y** +	(whether a single parent)
0.870447*	**REVOKED_Y** +	(License Revoked (Past 7 Years))
2.299562*	**URBANICITY_HU**	(Highly Urban/ Urban)

The maximum lift for model 2 is at cut off 0.4 with 44.69% lift. The maximum gain is 4% at cut off 0.1. This is the same observation we got from model1.

GROUP	CUTOFF	OBS	True Positives (TP)	Total OBS	Total TP	Random TP	TP Rate	Random Rate	LIFT	Cumulative TP Rate	Random TP Rate	Difference
0	1.00000	0	0	0	0	0	0	0		0%	0%	0%
1	0.9	28	24	28	24	7	1%	0%	3.25	1%	10%	-9%
2	0.8	129	106	157	130	7	5%	0%	14.35	6%	20%	-14%
3	0.7	293	221	450	351	7	10%	0%	29.93	16%	30%	-14%
4	0.6	369	251	819	602	7	12%	0%	33.99	28%	40%	-12%
5	0.5	529	281	1348	883	7	13%	0%	38.05	41%	50%	-9%
6	0.4	697	330	2045	1213	7	15%	0%	44.69	56%	60%	-4%
7	0.3	923	305	2968	1518	7	14%	0%	41.30	71%	70%	1%
8	0.2	1136	277	4104	1795	7	13%	0%	37.51	83%	80%	3%
9	0.1	1586	234	5690	2029	7	11%	0%	31.69	94%	90%	4%
10	0	2470	123	8160	2152	7	6%	0%	16.66	100%	100%	0%
								LIFT =	44.69		GAIN =	4%

Insurance Logistic Regression Project

Below are the ROC curve which indicating model2 is a much better model than random. It seems the best cut off point near to 0.4.



The KS value we get for model 2 is 46% at cut off point 0.30.

GROUP	CUTOFF	True Positives (TP)	False Postive (FP)	Total TP	Total FP	TP Rate	FP Rate	Cumulative TP Rate	Cumulative FP Rate	Difference
0	1.00000	0	0	0	0	0	0	0%	0%	0%
1	0.78811	24	4	24	4	1%	0%	1%	0%	1%
2	0.38716	106	23	130	27	5%	0%	6%	0%	6%
3	0.11277	221	72	351	99	10%	1%	16%	2%	15%
4	0.06775	251	118	602	217	12%	2%	28%	4%	24%
5	0.05238	281	248	883	465	13%	4%	41%	8%	33%
6	0.03777	330	367	1213	832	15%	6%	56%	14%	43%
7	0.02718	305	618	1518	1450	14%	10%	71%	24%	46%
8	0.01581	277	859	1795	2309	13%	14%	83%	38%	45%
9	0.00375	234	1352	2029	3661	11%	23%	94%	61%	33%
10	0.00013	123	2347	2152	6008	6%	39%	100%	100%	0%
									KS =	46%

MODEL 3- Taken Variable Effect in Practice

Before we do any analysis, we have some idea how we expect the variable will impact on car crash. Comparing to the coefficients in model 2 with theory. We only find out one variables have the opposite effect than we thought. It is the only claims is the past 5 years. In theory, we expect that If customers' total payout over the past five years was high, this suggests future payouts will be high. In the model 2, it is telling us a different story. If we want make sure everything work out the same way as we expect. We delete the log_OLDCLAIM variable out of the model. We get the below result:

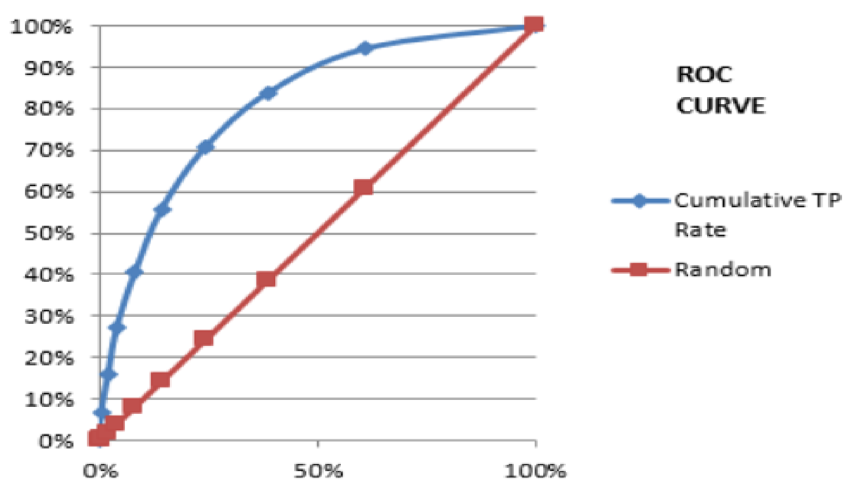
```
Call:
glm(formula = TARGET_FLAG ~ MVR_PTS + No_CLM_FREQ + No_HOME +
    log_INCOME + No_HOMEKIDS + No_KIDSDRIV + log_BLUEBOOK + TIF +
    log_TRAVTIME + CAR_TYPE_MV + CAR_USE_C + LowEducation + JOB_WHITE_COLLAR +
    MSTATUS_Y + PARENT_Y + REVOKED_Y + URBANICITY_HU, family = binomial(),
    data = CleanData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4777  -0.7180  -0.4044   0.6373   3.1224

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       0.455917   0.513500   0.888  0.37461
MVR_PTS           0.104446   0.013992   7.464 8.37e-14 ***
No_CLM_FREQ      -0.424701   0.064847  -6.549 5.78e-11 ***
No_HOME           0.249780   0.077039   3.242  0.00119 **
log_INCOME       -0.054118   0.010211  -5.300 1.16e-07 ***
No_HOMEKIDS      -0.231597   0.087828  -2.637  0.00837 **
No_KIDSDRIV      -0.583316   0.095396  -6.115 9.68e-10 ***
log_BLUEBOOK     -0.383646   0.048853  -7.853 4.06e-15 ***
TIF              -0.052988   0.007283  -7.276 3.45e-13 ***
log_TRAVTIME      0.436512   0.054053   8.076 6.71e-16 ***
CAR_TYPE_MV      -0.665632   0.073937  -9.003  < 2e-16 ***
CAR_USE_C         0.593599   0.064700   9.175  < 2e-16 ***
LowEducation      0.545124   0.065641   8.305  < 2e-16 ***
JOB_WHITE_COLLAR -0.485390   0.084399  -5.751 8.86e-09 ***
MSTATUS_Y        -0.561022   0.086245  -6.505 7.77e-11 ***
PARENT_Y          0.243603   0.119873   2.032  0.04214 *
REVOKED_Y         0.738659   0.080013   9.232  < 2e-16 ***
URBANICITY_HU     2.309246   0.112249  20.573  < 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: 9415.3  on 8159  degrees of freedom
Residual deviance: 7344.4  on 8142  degrees of freedom
AIC: 7380.4

Number of Fisher Scoring iterations: 5
```

The AIC has increased a little bit comparing to model2. All the parameters for the variables in the model are predictive. However, the estimate of the intercept tends out not good. The p value is very high.

YHAT =	0.455917	+	(intercept)
	0.104446*	**MVR_PTS** +	(Motor Vehicle Record Points)
	-0.424701*	**No_CLM_FREQ** +	(No claim in the past 5 years)
	0.249780*	**NO_HOME** +	(Not own a home)
	-0.054118*	**log_INCOME** +	(income +1 in log)
	-0.231597*	**No_HOMEKIDS** +	(no kids at home)
	-0.583316*	**No_KIDSDRIV** +	(no driving children)
	-0.383646*	**log_BLUEBOOK** +	(Value of Vehicle + 1 in log)
	-0.052988*	**TIF** +	(Time in Force)
	0.4356512*	**log_TRAVTIME** +	(Distance to Work +1 in log)
	-0.665632*	**CAR_TYPE_MV** +	(minivan)
	0.593599*	**CAR_USE_C** +	(Commercial vehicles)
	0.545124*	**LowEducation** +	(high school or below)
	-0.485390*	**JOB_WHITE_COLLAR** +	(Whether Doctor, Lawyer, or Manager)
	-0.561022*	**MSTATUS_Y** +	(whether married)
	0.243603*	**PARENT_Y** +	(whether a single parent)
	0.738659*	**REVOKED_Y** +	(License Revoked (Past 7 Years))
	2.309246*	**URBANICITY_HU**	(Highly Urban/ Urban)

The maximum lift for model 3 is at cut off 0.4 with 44.55% lift, slightly lower than the value we got from model2. The maximum gain is 4% at cut off 0.1. This is the same we go from both model 1 and 2.

GROUP	CUTOFF	OBS	True Positives (TP)	Total OBS	Total TP	Random TP	TP Rate	Random Rate	LIFT	Cumulative TP Rate	Random TP Rate	Difference
0	1.00000	0	0	0	0	0	0	0		0%	0%	0%
1	0.9	28	20	28	20	7	1%	0%	2.71	1%	10%	-9%
2	0.8	129	121	157	141	7	6%	0%	16.39	7%	20%	-13%
3	0.7	293	199	450	340	7	9%	0%	26.95	16%	30%	-14%
4	0.6	369	248	819	588	7	12%	0%	33.58	27%	40%	-13%
5	0.5	529	281	1348	869	7	13%	0%	38.05	40%	50%	-10%
6	0.4	697	329	2045	1198	7	15%	0%	44.55	56%	60%	-4%
7	0.3	923	322	2968	1520	7	15%	0%	43.61	71%	70%	1%
8	0.2	1136	281	4104	1801	7	13%	0%	38.05	84%	80%	4%
9	0.1	1586	232	5690	2033	7	11%	0%	31.42	94%	90%	4%
10	0	2470	119	8160	2152	7	6%	0%	16.12	100%	100%	0%
								LIFT =	44.55		GAIN =	4%



The KS value we get for model 3 is 47% at cut off point 0.30.

GROUP	CUTOFF	True Positives (TP)	False Postive (FP)	Total TP	Total FP	TP Rate	FP Rate	Cumulative TP Rate	Cumulative FP Rate	Difference
0	1.00000	0	0	0	0	0	0	0%	0%	0%
1	0.9	20	8	20	8	1%	0%	1%	0%	1%
2	0.8	121	8	141	16	6%	0%	7%	0%	6%
3	0.7	199	94	340	110	9%	2%	16%	2%	14%
4	0.6	248	121	588	231	12%	2%	27%	4%	23%
5	0.5	281	248	869	479	13%	4%	40%	8%	32%
6	0.4	329	368	1198	847	15%	6%	56%	14%	42%
7	0.3	322	601	1520	1448	15%	10%	71%	24%	47%
8	0.2	281	855	1801	2303	13%	14%	84%	38%	45%
9	0.1	232	1354	2033	3657	11%	23%	94%	61%	34%
10	0	119	2351	2152	6008	6%	39%	100%	100%	0%
									KS =	47%

All the three models we have build so far have similar performance. They have very clos KS values, ROC curve and AIC. Actually, we can just choose one of this. However, I still prefer model2 for the consideration of the significance of parameters.

CONCLUSION

The logistic regression models build this time do not have significant difference in performance. I should have try to build a bench mark model with all data as original as possible. However, from the ROC curves and the KS values, it is apparent that all those three models are much better than the random guess. We do not have much guidelines how business decision markers want to achieve. We try the best to predict a customer will have a car crash or not. At the insurance company side, they can get the percentage of car crash in control. Maybe they can try to lower it from 26% to 20%. We have a secondary model predicting the claim loss. We have to think about how to achieve those two goals. We also have to take the opportunity cost for false positive and false negative to select the best model.

BINGO BONUS:

WEKA

** Variable visualization with target variable already in the report.

=== Run information ===

Evaluator: weka.attributeSelection.ChiSquaredAttributeEval

Search:weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1

Relation: logit_insurance_weka

Instances: 8161

Attributes: 24

KIDSDRIV

AGE

HOMEKIDS

YOJ

INCOME

PARENT1

HOME_VAL

MSTATUS

SEX

EDUCATION

JOB

TRAVTIME

CAR_USE

BLUEBOOK

TIF

CAR_TYPE

RED_CAR

OLDCLAIM

CLM_FREQ

REVOKED

MVR_PTS

CAR_AGE

URBANICITY

TARGET_FLAG

Evaluation mode:10-fold cross-validation

=== Attribute selection 10 fold cross-validation (stratified), seed: 1 ===

average merit	average rank	attribute
432.682 +-12.043	1 +- 0	18 OLDCLAIM
429.604 +-11.769	2 +- 0	19 CLM_FREQ
369.42 +- 9.504	3.2 +- 0.4	23 URBANICITY
359.088 +-14.296	3.8 +- 0.4	21 MVR_PTS
240.171 +-12.417	5.1 +- 0.3	7 HOME_VAL
223.746 +- 7.148	6.3 +- 0.46	2 AGE
220.341 +-11.599	6.6 +- 0.66	11 JOB
182.632 +-10.519	8.3 +- 0.46	6 PARENT1
171.055 +- 8.595	9.4 +- 0.92	5 INCOME
169.67 +- 8.835	9.5 +- 0.81	20 REVOKED
153.932 +- 7.004	11.4 +- 0.8	16 CAR_TYPE
153.897 +- 8.558	12 +- 1	10 EDUCATION
149.562 +- 5.465	12.5 +- 0.81	13 CAR_USE
134.239 +- 8.274	14.4 +- 0.8	8 MSTATUS
124.372 +-13.676	14.9 +- 0.7	14 BLUEBOOK

123.145 +- 6.526 15.6 +- 0.66 3 HOMEKIDS

79.057 +- 7.568 17 +- 0 1 KIDSDRIV

69.211 +- 3.772 18 +- 0 22 CAR_AGE

51.808 +- 6.731 19.2 +- 0.4 4 YOJ

45.854 +- 3.467 19.8 +- 0.4 15 TIF

22.06 +- 2.085 21 +- 0 12 TRAVTIME

3.4 +- 1.376 22 +- 0 9 SEX

0.465 +- 0.426 23 +- 0 17 RED_CAR

R

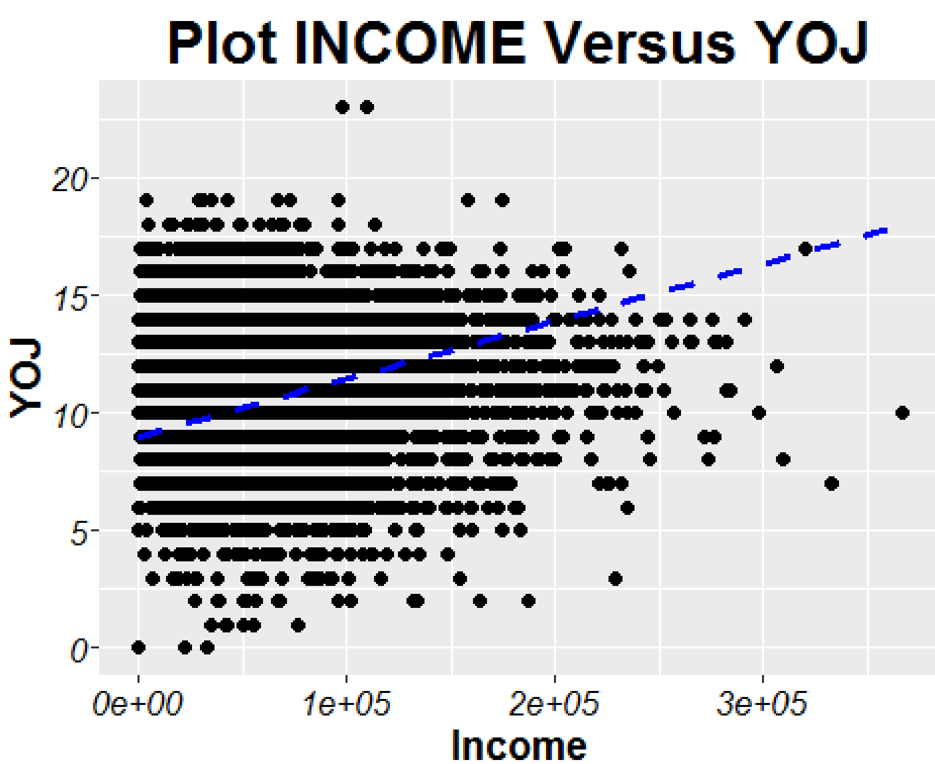
Unit02_Insurance

Ying Cheng

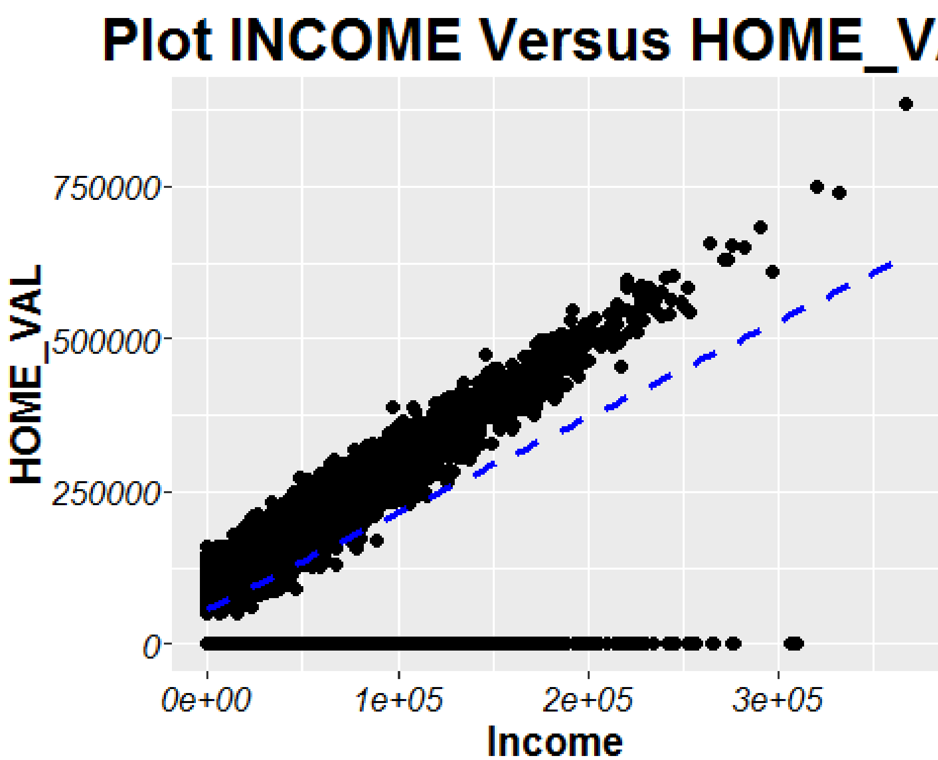
February 12, 2017

```
library(readr)
logit_insurance <- read_csv("C:/Users/Admin/Dropbox/Northwestern University/P
redict411/Auto Insurance Problem/logit_insurance.csv")
logit_insurance$TARGET_FLAG<- factor(logit_insurance$TARGET_FLAG)

### aov test for numeric data
### Get all the numeric columns

    nums <- sapply(logit_insurance, is.numeric)
    Numeric_data <- logit_insurance[ , nums]


    AOV_test <- function(x)
    {
      summary(aov(x~TARGET_FLAG,  data = logit_insurance))[[1]][["Pr(>F)"]]
    }

    sapply(Numeric_data, AOV_test)

##          INDEX TARGET_AMT     KIDSDRIV          AGE     HOMEKIDS
## [1,] 0.8801258          0 6.052406e-21 9.230158e-21 1.083837e-25
## [2,]        NA         NA           NA           NA           NA
##            YOJ       INCOME     HOME_VAL     TRAVTIME     BLUEBOOK
```

```
## [1,] 5.75827e-10 4.764099e-36 2.036088e-59 1.234536e-05 7.741376e-21
## [2,]          NA           NA           NA           NA           NA
##               TIF     OLDCLAIM     CLM_FREQ      MVR_PTS      CAR_AGE
## [1,] 9.145383e-14 4.962696e-36 6.332803e-87 2.320264e-89 1.095702e-18
## [2,]           NA           NA           NA           NA           NA

library(ggplot2)
library(gridExtra)

  ggplot(data=logit_insurance, aes(x=INCOME, y=YOJ)) +
        geom_point(pch=16, color="black", size=2) +
        geom_smooth(method="lm",  se = FALSE, color="blue",  size=1.2, linety
pe=2) +
        labs(title="Plot INCOME Versus YOJ", x="Income", y="YOJ") +
        theme(
                    title = element_text(size = 18, color = "black", face = "
bold"),
                    axis.text = element_text(colour = "black", size = 12, fac
e = "italic"),
                    axis.text.y = element_text(colour = "black",size = 12),
                    axis.title = element_text(size = 15, color = "black", fac
e = "bold"),
                    axis.title.y = element_text(size = 15, color = "black", f
ace = "bold")

                  )
```

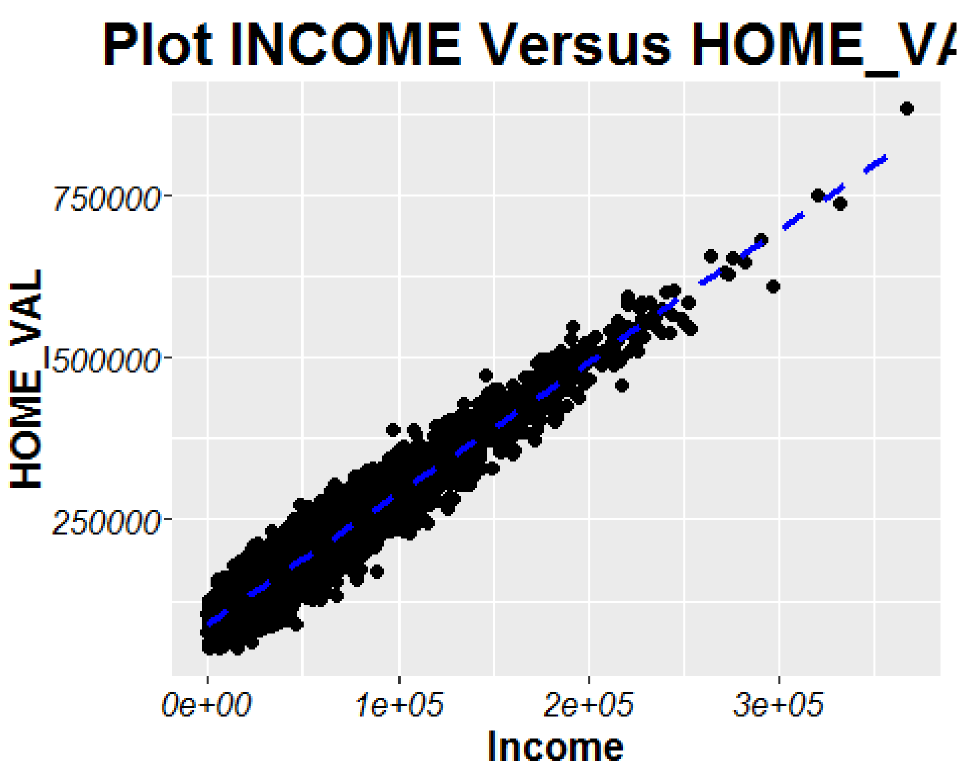


```
library(ggplot2)
library(gridExtra)

  ggplot(data=logit_insurance, aes(x=INCOME, y=HOME_VAL)) +
        geom_point(pch=16, color="black", size=2) +
        geom_smooth(method="lm",  se = FALSE, color="blue",  size=1.2, linety
pe=2) +
        labs(title="Plot INCOME Versus HOME_VAL", x="Income", y="HOME_VAL") +
        theme(
                    title = element_text(size = 18, color = "black", face = "
bold"),
                    axis.text = element_text(colour = "black", size = 12, fac
e = "italic"),
                    axis.text.y = element_text(colour = "black",size = 12),
                    axis.title = element_text(size = 15, color = "black", fac
e = "bold"),
                    axis.title.y = element_text(size = 15, color = "black", f
ace = "bold")

                  )
```




```
data1 <- subset(logit_insurance, logit_insurance$INCOME>0 & logit_insurance$H
OME_VAL>0)
cor(data1$INCOME, data1$HOME_VAL)

## [1] 0.9598955
```

```
library(ggplot2)
library(gridExtra)

  ggplot(data=data1, aes(x=INCOME, y=HOME_VAL)) +
        geom_point(pch=16, color="black", size=2) +
        geom_smooth(method="lm",  se = FALSE, color="blue",  size=1.2, linety
pe=2) +
        labs(title="Plot INCOME Versus HOME_VAL", x="Income", y="HOME_VAL") +
        theme(
                    title = element_text(size = 18, color = "black", face = "
bold"),
                    axis.text = element_text(colour = "black", size = 12, fac
e = "italic"),
                    axis.text.y = element_text(colour = "black",size = 12),
                    axis.title = element_text(size = 15, color = "black", fac
e = "bold"),
                    axis.title.y = element_text(size = 15, color = "black", f
ace = "bold")

                  )
```




```
options(scipen=999)

lm.model_income_homeV <- lm(INCOME~ HOME_VAL, data1)
summary(lm.model_income_homeV)
```

```
##
## Call:
## lm(formula = INCOME ~ HOME_VAL, data = data1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -44515  -8454    -54   8352  54235
##
## Coefficients:
##                  Estimate    Std. Error t value            Pr(>|t|)
## (Intercept) -34863.046802    469.366901  -74.28 <0.0000000000000002 ***
## HOME_VAL         0.454851      0.001905  238.76 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12460 on 4863 degrees of freedom
## Multiple R-squared:  0.9214, Adjusted R-squared:  0.9214
## F-statistic: 5.701e+04 on 1 and 4863 DF,  p-value: < 0.00000000000000022
```

```
    Factor_Proportion_test <- function(y) {

      Proportion_test <- prop.test(table(y, logit_insurance$TARGET_FLAG), cor
rect=FALSE)
      Proportion_test_p <- Proportion_test$p.value

      return(Proportion_test_p)
    }


     Factor_Proportion_test(logit_insurance$CAR_USE)

## [1] 0.00000000000000000000000000000000000005198618

     Factor_Proportion_test(logit_insurance$PARENT1)

## [1] 0.000000000000000000000000000000000000000000005223583

     Factor_Proportion_test(logit_insurance$RED_CAR)

## [1] 0.5302639

     Factor_Proportion_test(logit_insurance$REVOKED)

## [1] 0.0000000000000000000000000000000000000000007103044

     Factor_Proportion_test(logit_insurance$SEX)

## [1] 0.0568841

     Factor_Proportion_test(logit_insurance$URBANICITY)
```

```
## [1] 0.00000000000000000000000000000000000000000000000000000000000000000000000000000000000000002993051

    Factor_Proportion_test(logit_insurance$MSTATUS)

## [1] 0.000000000000000000000000000000002854425

library(gmodels)

CrossTable(logit_insurance$CAR_TYPE,logit_insurance$TARGET_FLAG, digits=2, prop.c=FALSE,
           prop.t=FALSE, prop.chisq=FALSE, chisq = FALSE, fisher=FALSE, mcnemar=FALSE,
           resid=FALSE, sresid=FALSE, asresid=FALSE,  format="SPSS")

## 
##    Cell Contents
## |-------------------------|
## |                   Count |
## |             Row Percent |
## |-------------------------|
## 
## Total Observations in Table:  8161 
## 
##                          | logit_insurance$TARGET_FLAG 
## logit_insurance$CAR_TYPE |        0  |        1  | Row Total | 
## -------------------------|-----------|-----------|-----------|
##                  Minivan |     1796  |      349  |     2145  | 
##                          |   83.73%  |   16.27%  |   26.28%  | 
## -------------------------|-----------|-----------|-----------|
##              Panel Truck |      498  |      178  |      676  | 
##                          |   73.67%  |   26.33%  |    8.28%  | 
## -------------------------|-----------|-----------|-----------|
##                   Pickup |      946  |      443  |     1389  | 
##                          |   68.11%  |   31.89%  |   17.02%  | 
## -------------------------|-----------|-----------|-----------|
##               Sports Car |      603  |      304  |      907  | 
##                          |   66.48%  |   33.52%  |   11.11%  | 
## -------------------------|-----------|-----------|-----------|
##                      Van |      549  |      201  |      750  | 
##                          |   73.20%  |   26.80%  |    9.19%  | 
## -------------------------|-----------|-----------|-----------|
##                    z_SUV |     1616  |      678  |     2294  | 
##                          |   70.44%  |   29.56%  |   28.11%  | 
## -------------------------|-----------|-----------|-----------|
##             Column Total |     6008  |     2153  |     8161  | 
## -------------------------|-----------|-----------|-----------|
## 
## 
```

```
CrossTable(logit_insurance$CAR_USE,logit_insurance$TARGET_FLAG, digits=2, pro
p.c=FALSE,
           prop.t=FALSE, prop.chisq=FALSE, chisq = FALSE, fisher=FALSE, mcnem
ar=FALSE,
           resid=FALSE, sresid=FALSE, asresid=FALSE,  format="SPSS")

## 
##    Cell Contents
## |-------------------------|
## |                   Count |
## |             Row Percent |
## |-------------------------|
## 
## Total Observations in Table:  8161 
## 
##                         | logit_insurance$TARGET_FLAG 
## logit_insurance$CAR_USE |        0  |        1  | Row Total | 
## ------------------------|-----------|-----------|-----------|
##              Commercial |     1982  |     1047  |     3029  | 
##                         |   65.43%  |   34.57%  |   37.12%  | 
## ------------------------|-----------|-----------|-----------|
##                 Private |     4026  |     1106  |     5132  | 
##                         |   78.45%  |   21.55%  |   62.88%  | 
## ------------------------|-----------|-----------|-----------|
##            Column Total |     6008  |     2153  |     8161  | 
## ------------------------|-----------|-----------|-----------|
## 
## 

CrossTable(logit_insurance$EDUCATION,logit_insurance$TARGET_FLAG, digits=2, p
rop.c=FALSE,
           prop.t=FALSE, prop.chisq=FALSE, chisq = FALSE, fisher=FALSE, mcnem
ar=FALSE,
           resid=FALSE, sresid=FALSE, asresid=FALSE,  format="SPSS")

## 
##    Cell Contents
## |-------------------------|
## |                   Count |
## |             Row Percent |
## |-------------------------|
## 
## Total Observations in Table:  8161 
## 
##                           | logit_insurance$TARGET_FLAG 
## logit_insurance$EDUCATION |        0  |        1  | Row Total | 
## --------------------------|-----------|-----------|-----------|
##              <High School |      818  |      385  |     1203  | 
##                           |   68.00%  |   32.00%  |   14.74%  | 
## --------------------------|-----------|-----------|-----------|
```

```
##                 Bachelors |      1719 |       523 |      2242 |
##                           |    76.67% |    23.33% |    27.47% |
## --------------------------|-----------|-----------|-----------|
##                   Masters |      1331 |       327 |      1658 |
##                           |    80.28% |    19.72% |    20.32% |
## --------------------------|-----------|-----------|-----------|
##                       PhD |       603 |       125 |       728 |
##                           |    82.83% |    17.17% |     8.92% |
## --------------------------|-----------|-----------|-----------|
##             z_High School |      1537 |       793 |      2330 |
##                           |    65.97% |    34.03% |    28.55% |
## --------------------------|-----------|-----------|-----------|
##              Column Total |      6008 |      2153 |      8161 |
## --------------------------|-----------|-----------|-----------|
##
##
```

```
CrossTable(logit_insurance$JOB,logit_insurance$TARGET_FLAG, digits=2, prop.c=
FALSE,
           prop.t=FALSE, prop.chisq=FALSE, chisq = FALSE, fisher=FALSE, mcnem
ar=FALSE,
           resid=FALSE, sresid=FALSE, asresid=FALSE,  format="SPSS")
```

```
##
##    Cell Contents
## |-------------------------|
## |                   Count |
## |             Row Percent |
## |-------------------------|
##
## Total Observations in Table:  7635
##
##                      | logit_insurance$TARGET_FLAG
## logit_insurance$JOB |         0 |         1 | Row Total |
## --------------------|-----------|-----------|-----------|
##            Clerical |       900 |       371 |      1271 |
##                     |    70.81% |    29.19% |    16.65% |
## --------------------|-----------|-----------|-----------|
##              Doctor |       217 |        29 |       246 |
##                     |    88.21% |    11.79% |     3.22% |
## --------------------|-----------|-----------|-----------|
##           Home Maker |       461 |       180 |       641 |
##                     |    71.92% |    28.08% |     8.40% |
## --------------------|-----------|-----------|-----------|
##              Lawyer |       682 |       153 |       835 |
##                     |    81.68% |    18.32% |    10.94% |
## --------------------|-----------|-----------|-----------|
##             Manager |       851 |       137 |       988 |
##                     |    86.13% |    13.87% |    12.94% |
## --------------------|-----------|-----------|-----------|
```

```
##         Professional |       870 |       247 |      1117 |
##                      |    77.89% |    22.11% |    14.63% |
## ---------------------|-----------|-----------|-----------|
##              Student |       446 |       266 |       712 |
##                      |    62.64% |    37.36% |     9.33% |
## ---------------------|-----------|-----------|-----------|
##        z_Blue Collar |      1191 |       634 |      1825 |
##                      |    65.26% |    34.74% |    23.90% |
## ---------------------|-----------|-----------|-----------|
##         Column Total |      5618 |      2017 |      7635 |
## ---------------------|-----------|-----------|-----------|
## 
## 
```

```
CrossTable(logit_insurance$MSTATUS,logit_insurance$TARGET_FLAG, digits=2, pro
p.c=FALSE,
           prop.t=FALSE, prop.chisq=FALSE, chisq = FALSE, fisher=FALSE, mcnem
ar=FALSE,
           resid=FALSE, sresid=FALSE, asresid=FALSE,  format="SPSS")
```

```
## 
##    Cell Contents
## |-------------------------|
## |                   Count |
## |             Row Percent |
## |-------------------------|
## 
## Total Observations in Table:  8161 
## 
##                         | logit_insurance$TARGET_FLAG 
## logit_insurance$MSTATUS |         0 |         1 | Row Total | 
## ------------------------|-----------|-----------|-----------|
##                     Yes |      3841 |      1053 |      4894 | 
##                         |    78.48% |    21.52% |    59.97% | 
## ------------------------|-----------|-----------|-----------|
##                    z_No |      2167 |      1100 |      3267 | 
##                         |    66.33% |    33.67% |    40.03% | 
## ------------------------|-----------|-----------|-----------|
##            Column Total |      6008 |      2153 |      8161 | 
## ------------------------|-----------|-----------|-----------|
## 
## 
```

```
CrossTable(logit_insurance$PARENT1,logit_insurance$TARGET_FLAG, digits=2, pro
p.c=FALSE,
           prop.t=FALSE, prop.chisq=FALSE, chisq = FALSE, fisher=FALSE, mcnem
ar=FALSE,
           resid=FALSE, sresid=FALSE, asresid=FALSE,  format="SPSS")
```

```
## 
##    Cell Contents
```

```
## |-------------------------|
## |                   Count |
## |             Row Percent |
## |-------------------------|
##
## Total Observations in Table:  8161
##
##                          | logit_insurance$TARGET_FLAG
## logit_insurance$PARENT1 |         0 |         1 | Row Total |
## ------------------------|-----------|-----------|-----------|
##                      No |      5407 |      1677 |      7084 |
##                         |    76.33% |    23.67% |    86.80% |
## ------------------------|-----------|-----------|-----------|
##                     Yes |       601 |       476 |      1077 |
##                         |    55.80% |    44.20% |    13.20% |
## ------------------------|-----------|-----------|-----------|
##            Column Total |      6008 |      2153 |      8161 |
## ------------------------|-----------|-----------|-----------|
##
##
```

```
CrossTable(logit_insurance$RED_CAR,logit_insurance$TARGET_FLAG, digits=2, pro
p.c=FALSE,
           prop.t=FALSE, prop.chisq=FALSE, chisq = FALSE, fisher=FALSE, mcnem
ar=FALSE,
           resid=FALSE, sresid=FALSE, asresid=FALSE,  format="SPSS")
```

```
##
##    Cell Contents
## |-------------------------|
## |                   Count |
## |             Row Percent |
## |-------------------------|
##
## Total Observations in Table:  8161
##
##                          | logit_insurance$TARGET_FLAG
## logit_insurance$RED_CAR |         0 |         1 | Row Total |
## ------------------------|-----------|-----------|-----------|
##                      no |      4246 |      1537 |      5783 |
##                         |    73.42% |    26.58% |    70.86% |
## ------------------------|-----------|-----------|-----------|
##                     yes |      1762 |       616 |      2378 |
##                         |    74.10% |    25.90% |    29.14% |
## ------------------------|-----------|-----------|-----------|
##            Column Total |      6008 |      2153 |      8161 |
## ------------------------|-----------|-----------|-----------|
##
##
```

```
aov.INCOME_insurance<- aov(INCOME~JOB, logit_insurance)
summary(aov.INCOME_insurance)
```

```
##                Df        Sum Sq      Mean Sq F value              Pr(>F)
## JOB             7 6824037927768 974862561110   981.3 <0.0000000000000002
## Residuals    7206 7158691798456    993434887                            
##                
## JOB         ***
## Residuals      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 947 observations deleted due to missingness
```

```
TukeyHSD(aov.INCOME_insurance)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
## 
## Fit: aov(formula = INCOME ~ JOB, data = logit_insurance)
## 
## $JOB
##                                  diff         lwr          upr     p adj
## Doctor-Clerical            94818.5113   88036.994  101600.0283 0.0000000
## Home Maker-Clerical       -21787.8500  -26572.395  -17003.3046 0.0000000
## Lawyer-Clerical            54443.6136   50078.920   58808.3071 0.0000000
## Manager-Clerical           53600.3693   49434.196   57766.5425 0.0000000
## Professional-Clerical      42731.9094   38699.414   46764.4049 0.0000000
## Student-Clerical          -27551.5291  -32186.013  -22917.0454 0.0000000
## z_Blue Collar-Clerical     25095.8254   21502.845   28688.8061 0.0000000
## Home Maker-Doctor        -116606.3614 -123930.062 -109282.6606 0.0000000
## Lawyer-Doctor             -40374.8977  -47431.474  -33318.3216 0.0000000
## Manager-Doctor            -41218.1420  -48153.682  -34282.6020 0.0000000
## Professional-Doctor       -52086.6019  -58942.675  -45230.5289 0.0000000
## Student-Doctor           -122370.0404 -129596.599 -115143.4814 0.0000000
## z_Blue Collar-Doctor      -69722.6859  -76329.820  -63115.5514 0.0000000
## Lawyer-Home Maker          76231.4636   71064.436   81398.4908 0.0000000
## Manager-Home Maker         75388.2193   70387.757   80388.6812 0.0000000
## Professional-Home Maker    64519.7594   59630.113   69409.4056 0.0000000
## Student-Home Maker         -5763.6791  -11160.535    -366.8227 0.0266072
## z_Blue Collar-Home Maker   46883.6755   42349.678   51417.6726 0.0000000
## Manager-Lawyer              -843.2443   -5443.602    3757.1136 0.9993252
## Professional-Lawyer       -11711.7042  -16191.360   -7232.0486 0.0000000
## Student-Lawyer            -81995.1427  -87023.535  -76966.7500 0.0000000
## z_Blue Collar-Lawyer      -29347.7882  -33436.285  -25259.2916 0.0000000
## Professional-Manager      -10868.4599  -15154.923   -6581.9971 0.0000000
## Student-Manager           -81151.8984  -86008.974  -76294.8229 0.0000000
## z_Blue Collar-Manager     -28504.5439  -32380.399  -24628.6887 0.0000000
## Student-Professional      -70283.4385  -75026.349  -65540.5276 0.0000000
## z_Blue Collar-Professional -17636.0840 -21367.876  -13904.2916 0.0000000
## z_Blue Collar-Student      52647.3545   48272.004   57022.7052 0.0000000
```

```
## upload the final cleaned data from SAS

CleanData <- read_csv("C:/Users/Admin/Dropbox/Northwestern University/Predict
411/Auto Insurance Problem/CleanData.csv")
CleanData$TARGET_FLAG<- factor(CleanData$TARGET_FLAG)

library(MASS)

fit1 <- glm(TARGET_FLAG ~
          MVR_PTS +
                    No_CLM_FREQ  +
                    No_HOME +
                    log_INCOME+
                    No_Income +
                    log_OLDCLAIM+
            #       No_OLDCLAIM +
                    No_HOMEKIDS +
                    No_KIDSDRIV+
                    log_BLUEBOOK +
                    IMP_AGE+
                    log_CAR_AGE+
                    NewCar+
                    TIF +
                    IMP_YOJ +
                    log_TRAVTIME+
                    CAR_TYPE_MV  +
                    CAR_TYPE_PS+
                    CAR_USE_C +
                    HighEducation +
                    LowEducation +
                    JOB_WHITE_COLLAR +
                    JOB_BLUE_STUDENT+
                    MSTATUS_Y +
                    PARENT_Y +
                    RED_CAR_Y +
                    REVOKED_Y +
                    SEX_M +
                    URBANICITY_HU ,
                    data = CleanData, family=binomial())

stepAIC(fit1, direction="forward")

## Start:  AIC=7380.29
## TARGET_FLAG ~ MVR_PTS + No_CLM_FREQ + No_HOME + log_INCOME +
##     No_Income + log_OLDCLAIM + No_HOMEKIDS + No_KIDSDRIV + log_BLUEBOOK +
##     IMP_AGE + log_CAR_AGE + NewCar + TIF + IMP_YOJ + log_TRAVTIME +
##     CAR_TYPE_MV + CAR_TYPE_PS + CAR_USE_C + HighEducation + LowEducation +

##     JOB_WHITE_COLLAR + JOB_BLUE_STUDENT + MSTATUS_Y + PARENT_Y +
##     RED_CAR_Y + REVOKED_Y + SEX_M + URBANICITY_HU
```

```
##
## Call:  glm(formula = TARGET_FLAG ~ MVR_PTS + No_CLM_FREQ + No_HOME +
##     log_INCOME + No_Income + log_OLDCLAIM + No_HOMEKIDS + No_KIDSDRIV +
##     log_BLUEBOOK + IMP_AGE + log_CAR_AGE + NewCar + TIF + IMP_YOJ +
##     log_TRAVTIME + CAR_TYPE_MV + CAR_TYPE_PS + CAR_USE_C + HighEducation +

##     LowEducation + JOB_WHITE_COLLAR + JOB_BLUE_STUDENT + MSTATUS_Y +
##     PARENT_Y + RED_CAR_Y + REVOKED_Y + SEX_M + URBANICITY_HU,
##     family = binomial(), data = CleanData)
##
## Coefficients:
##      (Intercept)           MVR_PTS       No_CLM_FREQ           No_HOME
##         1.742837          0.104420         -1.845063          0.229703
##       log_INCOME         No_Income      log_OLDCLAIM       No_HOMEKIDS
##        -0.109407         -0.496924         -0.162199         -0.221959
##      No_KIDSDRIV      log_BLUEBOOK           IMP_AGE       log_CAR_AGE
##        -0.569340         -0.364761          0.001254          0.143987
##           NewCar               TIF           IMP_YOJ      log_TRAVTIME
##         0.280964         -0.053360          0.011604          0.434202
##      CAR_TYPE_MV       CAR_TYPE_PS         CAR_USE_C     HighEducation
##        -0.617948          0.054625          0.602150          0.011673
##     LowEducation  JOB_WHITE_COLLAR  JOB_BLUE_STUDENT         MSTATUS_Y
##         0.512870         -0.476942          0.073161         -0.586447
##         PARENT_Y         RED_CAR_Y         REVOKED_Y             SEX_M
##         0.248031         -0.025109          0.876554         -0.074836
##    URBANICITY_HU
##         2.308090
##
## Degrees of Freedom: 8159 Total (i.e. Null);  8131 Residual
## Null Deviance:       9415
## Residual Deviance: 7322  AIC: 7380
```

```
summary(fit1)
```

```
##
## Call:
## glm(formula = TARGET_FLAG ~ MVR_PTS + No_CLM_FREQ + No_HOME +
##     log_INCOME + No_Income + log_OLDCLAIM + No_HOMEKIDS + No_KIDSDRIV +
##     log_BLUEBOOK + IMP_AGE + log_CAR_AGE + NewCar + TIF + IMP_YOJ +
##     log_TRAVTIME + CAR_TYPE_MV + CAR_TYPE_PS + CAR_USE_C + HighEducation +

##     LowEducation + JOB_WHITE_COLLAR + JOB_BLUE_STUDENT + MSTATUS_Y +
##     PARENT_Y + RED_CAR_Y + REVOKED_Y + SEX_M + URBANICITY_HU,
##     family = binomial(), data = CleanData)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4204  -0.7153  -0.4036   0.6466   3.1376
##
## Coefficients:
```

```
##                      Estimate Std. Error z value             Pr(>|z|)
## (Intercept)          1.742837   0.806387   2.161             0.030673 *
## MVR_PTS              0.104420   0.014018   7.449  0.00000000000009406 ***
## No_CLM_FREQ         -1.845063   0.402155  -4.588  0.00000447631834662 ***
## No_HOME              0.229703   0.078549   2.924             0.003452 **
## log_INCOME          -0.109407   0.038164  -2.867             0.004147 **
## No_Income           -0.496924   0.423531  -1.173             0.240681
## log_OLDCLAIM        -0.162199   0.045343  -3.577             0.000347 ***
## No_HOMEKIDS         -0.221959   0.099966  -2.220             0.026395 *
## No_KIDSDRIV         -0.569340   0.098065  -5.806  0.00000000640832140 ***
## log_BLUEBOOK        -0.364761   0.051482  -7.085  0.00000000000138804 ***
## IMP_AGE              0.001254   0.004130   0.304             0.761355
## log_CAR_AGE          0.143987   0.122373   1.177             0.239347
## NewCar               0.280964   0.193848   1.449             0.147225
## TIF                 -0.053360   0.007311  -7.299  0.00000000000028974 ***
## IMP_YOJ              0.011604   0.011501   1.009             0.312992
## log_TRAVTIME         0.434202   0.054151   8.018  0.00000000000000107 ***
## CAR_TYPE_MV         -0.617948   0.083193  -7.428  0.00000000000011032 ***
## CAR_TYPE_PS          0.054625   0.068710   0.795             0.426611
## CAR_USE_C            0.602150   0.078190   7.701  0.00000000000001349 ***
## HighEducation        0.011673   0.099883   0.117             0.906964
## LowEducation         0.512870   0.084360   6.080  0.00000000120504938 ***
## JOB_WHITE_COLLAR -0.476942   0.091075  -5.237  0.00000016337625302 ***
## JOB_BLUE_STUDENT  0.073161   0.079180   0.924             0.355500
## MSTATUS_Y           -0.586447   0.087578  -6.696  0.00000000002137554 ***
## PARENT_Y             0.248031   0.120338   2.061             0.039291 *
## RED_CAR_Y           -0.025109   0.086115  -0.292             0.770610
## REVOKED_Y            0.876554   0.088206   9.938 < 0.0000000000000002 ***
## SEX_M               -0.074836   0.084168  -0.889             0.373937
## URBANICITY_HU        2.308090   0.112604  20.497 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9415.3  on 8159  degrees of freedom
## Residual deviance: 7322.3  on 8131  degrees of freedom
## AIC: 7380.3
##
## Number of Fisher Scoring iterations: 5
```

```
fit2 <- glm(TARGET_FLAG ~ MVR_PTS +
      No_CLM_FREQ +
      No_HOME +
      log_INCOME +
      log_OLDCLAIM +
      No_HOMEKIDS +
      No_KIDSDRIV +
      log_BLUEBOOK +
      TIF +
```

```
      log_TRAVTIME +
      CAR_TYPE_MV +
      CAR_USE_C +
      LowEducation +
      JOB_WHITE_COLLAR +
      MSTATUS_Y +
      PARENT_Y +
      REVOKED_Y +
      URBANICITY_HU,
    family = binomial(), data = CleanData)

summary(fit2)

##
## Call:
## glm(formula = TARGET_FLAG ~ MVR_PTS + No_CLM_FREQ + No_HOME +
##     log_INCOME + log_OLDCLAIM + No_HOMEKIDS + No_KIDSDRIV + log_BLUEBOOK +

##     TIF + log_TRAVTIME + CAR_TYPE_MV + CAR_USE_C + LowEducation +
##     JOB_WHITE_COLLAR + MSTATUS_Y + PARENT_Y + REVOKED_Y + URBANICITY_HU,
##     family = binomial(), data = CleanData)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4716  -0.7175  -0.4024   0.6439   3.1250
##
## Coefficients:
##                   Estimate Std. Error z value             Pr(>|z|)
## (Intercept)       1.821353   0.644024   2.828             0.004683 **
## MVR_PTS           0.104026   0.013997   7.432 0.000000000000107039 ***
## No_CLM_FREQ      -1.820780   0.401462  -4.535 0.000005750209327984 ***
## No_HOME           0.248516   0.077117   3.223             0.001270 **
## log_INCOME       -0.054345   0.010222  -5.316 0.000000105810314329 ***
## log_OLDCLAIM     -0.159289   0.045262  -3.519             0.000433 ***
## No_HOMEKIDS      -0.234877   0.087916  -2.672             0.007549 **
## No_KIDSDRIV      -0.572909   0.095616  -5.992 0.000000002075458104 ***
## log_BLUEBOOK     -0.381609   0.048903  -7.803 0.000000000000006026 ***
## TIF              -0.053123   0.007295  -7.282 0.000000000000327774 ***
## log_TRAVTIME      0.435091   0.054107   8.041 0.000000000000000889 ***
## CAR_TYPE_MV      -0.662794   0.074034  -8.953 < 0.0000000000000002 ***
## CAR_USE_C         0.589742   0.064763   9.106 < 0.0000000000000002 ***
## LowEducation      0.547460   0.065711   8.331 < 0.0000000000000002 ***
## JOB_WHITE_COLLAR -0.480619   0.084475  -5.690 0.000000012740421200 ***
## MSTATUS_Y        -0.560750   0.086283  -6.499 0.000000000080858581 ***
## PARENT_Y          0.240783   0.120001   2.007             0.044802 *
## REVOKED_Y         0.870447   0.088054   9.885 < 0.0000000000000002 ***
## URBANICITY_HU     2.299562   0.112398  20.459 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9415.3  on 8159  degrees of freedom
## Residual deviance: 7331.9  on 8141  degrees of freedom
## AIC: 7369.9
##
## Number of Fisher Scoring iterations: 5
```

```
fit3 <- glm(TARGET_FLAG ~ MVR_PTS +
      No_CLM_FREQ +
      No_HOME +
      log_INCOME +
      No_HOMEKIDS +
      No_KIDSDRIV +
      log_BLUEBOOK +
      TIF +
      log_TRAVTIME +
      CAR_TYPE_MV +
      CAR_USE_C +
      LowEducation +
      JOB_WHITE_COLLAR +
      MSTATUS_Y +
      PARENT_Y +
      REVOKED_Y +
      URBANICITY_HU,
    family = binomial(), data = CleanData)

summary(fit3)
```

```
##
## Call:
## glm(formula = TARGET_FLAG ~ MVR_PTS + No_CLM_FREQ + No_HOME +
##     log_INCOME + No_HOMEKIDS + No_KIDSDRIV + log_BLUEBOOK + TIF +
##     log_TRAVTIME + CAR_TYPE_MV + CAR_USE_C + LowEducation + JOB_WHITE_COLL
AR +
##     MSTATUS_Y + PARENT_Y + REVOKED_Y + URBANICITY_HU, family = binomial(),

##     data = CleanData)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4777  -0.7180  -0.4044   0.6373   3.1224
##
## Coefficients:
##                   Estimate Std. Error z value             Pr(>|z|)
## (Intercept)       0.455917   0.513500   0.888              0.37461
## MVR_PTS           0.104446   0.013992   7.464 0.000000000000083659 ***
## No_CLM_FREQ      -0.424701   0.064847  -6.549 0.000000000057818654 ***
## No_HOME           0.249780   0.077039   3.242              0.00119 **
## log_INCOME       -0.054118   0.010211  -5.300 0.000000115828502400 ***
```

```
## No_HOMEKIDS      -0.231597   0.087828  -2.637              0.00837 **
## No_KIDSDRIV      -0.583316   0.095396  -6.115 0.000000000967666161 ***
## log_BLUEBOOK     -0.383646   0.048853  -7.853 0.000000000000004062 ***
## TIF              -0.052988   0.007283  -7.276 0.000000000000344872 ***
## log_TRAVTIME      0.436512   0.054053   8.076 0.000000000000000671 ***
## CAR_TYPE_MV      -0.665632   0.073937  -9.003 < 0.0000000000000002 ***
## CAR_USE_C         0.593599   0.064700   9.175 < 0.0000000000000002 ***
## LowEducation      0.545124   0.065641   8.305 < 0.0000000000000002 ***
## JOB_WHITE_COLLAR -0.485390   0.084399  -5.751 0.000000008864216767 ***
## MSTATUS_Y        -0.561022   0.086245  -6.505 0.000000000077692510 ***
## PARENT_Y          0.243603   0.119873   2.032              0.04214 *
## REVOKED_Y         0.738659   0.080013   9.232 < 0.0000000000000002 ***
## URBANICITY_HU     2.309246   0.112249  20.573 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9415.3  on 8159  degrees of freedom
## Residual deviance: 7344.4  on 8142  degrees of freedom
## AIC: 7380.4
##
## Number of Fisher Scoring iterations: 5
```

```
library(caret)


predict_1 <- predict(fit1, type = 'response')
predict_2 <- predict(fit2, type = 'response')
predict_3 <- predict(fit3, type = 'response')

addmargins(table(CleanData$TARGET_FLAG, predict_1 <=0.2 ))
```

```
##
##       FALSE TRUE  Sum
##   0    2308 3700 6008
##   1    1795  357 2152
##   Sum  4103 4057 8160
```

```
addmargins(table(CleanData$TARGET_FLAG, predict_2 <=0.2637255 ))
```

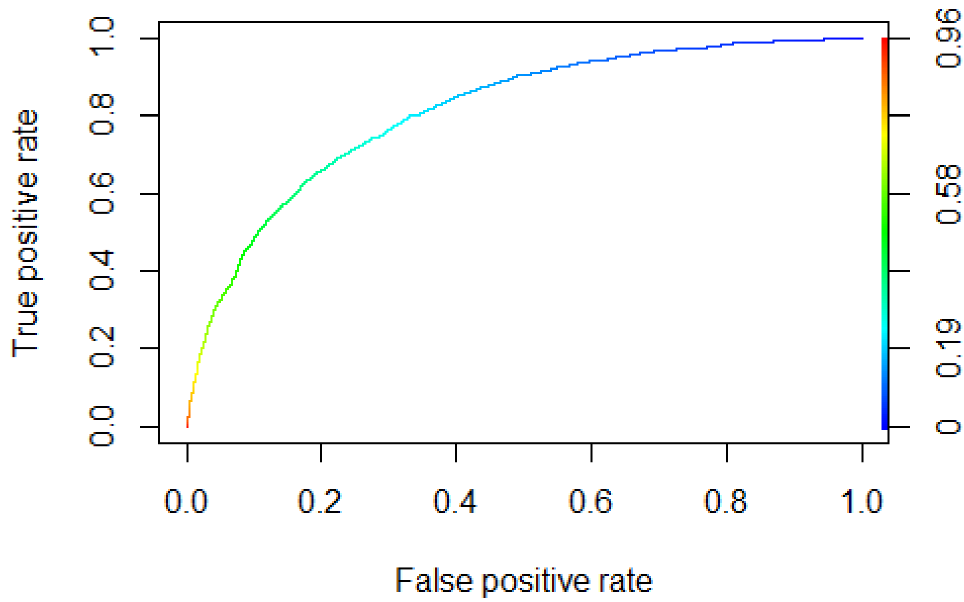
```
##
##       FALSE TRUE  Sum
##   0    1738 4270 6008
##   1    1624  528 2152
##   Sum  3362 4798 8160
```

```
addmargins(table(CleanData$TARGET_FLAG, predict_3 <=0.2637255 ))
```

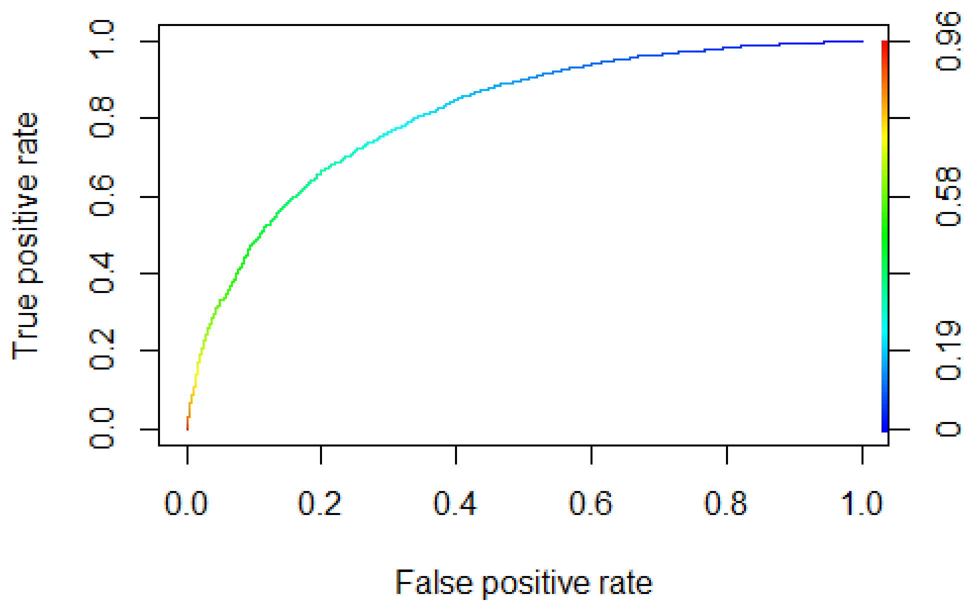
```
##
##       FALSE TRUE  Sum
```

```
## 0      1753 4255 6008
## 1      1623  529 2152
## Sum    3376 4784 8160
```

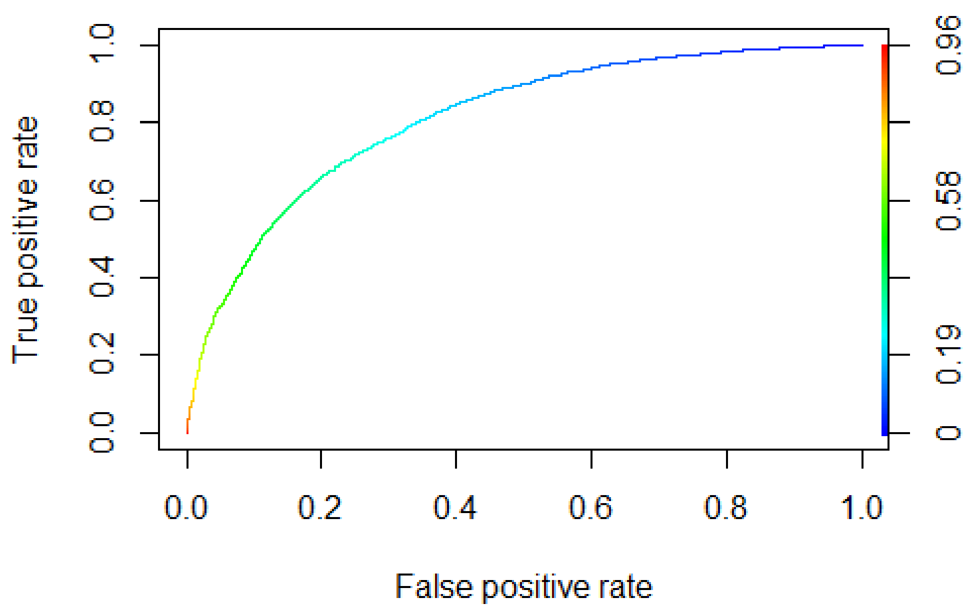
```
#ROCR Curve
library(ROCR)
ROCRpred_1 <- prediction(predict_1, CleanData$TARGET_FLAG)
ROCRperf_1 <- performance(ROCRpred_1,  'tpr','fpr')
plot(ROCRperf_1, colorize = TRUE, text.adj = c(-0.2,1.7))
```



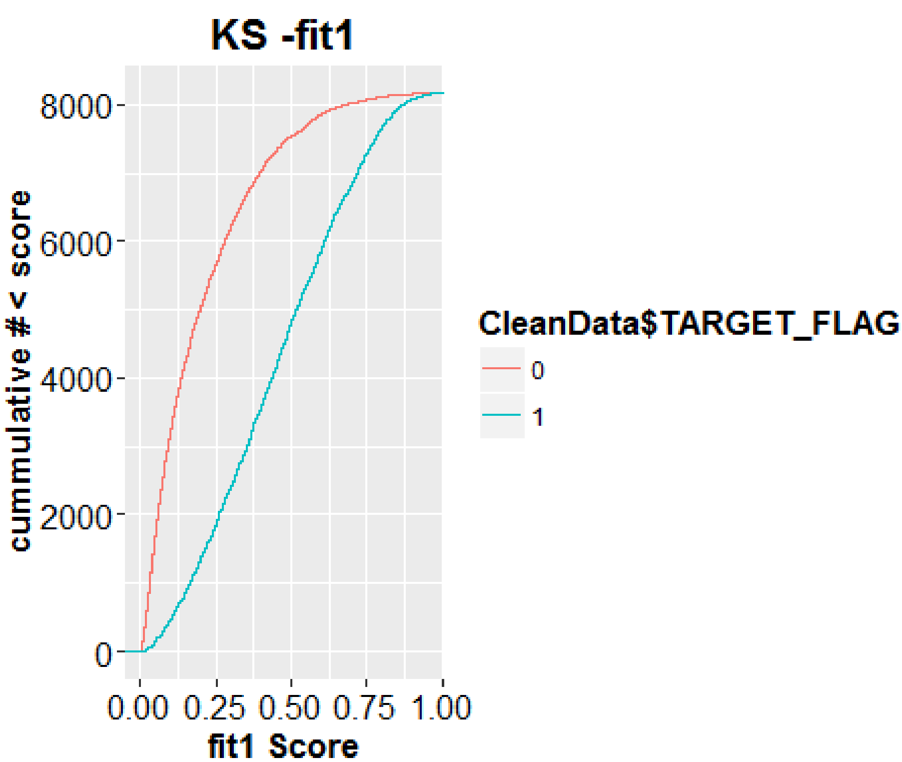

```
ROCRpred_2 <- prediction(predict_2, CleanData$TARGET_FLAG)
ROCRperf_2 <- performance(ROCRpred_2,  'tpr','fpr')
plot(ROCRperf_2, colorize = TRUE, text.adj = c(-0.2,1.7))
```




```
ROCRpred_3 <- prediction(predict_3, CleanData$TARGET_FLAG)
ROCRperf_3 <- performance(ROCRpred_3,  'tpr','fpr')
plot(ROCRperf_3, colorize = TRUE, text.adj = c(-0.2,1.7))
```

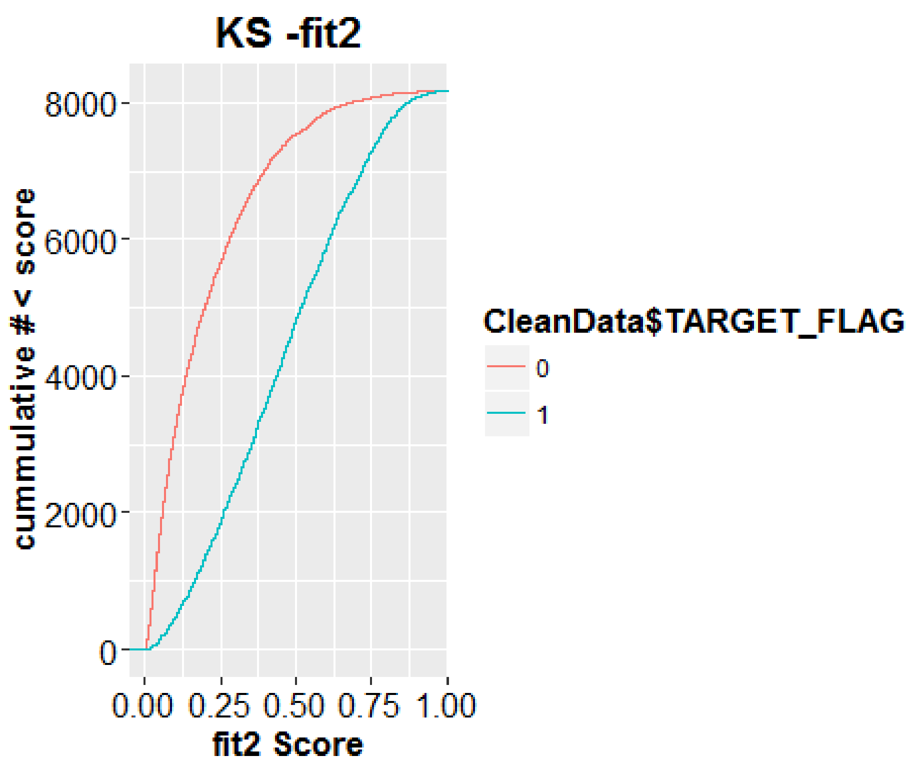


```
library(ggplot2)
library(gridExtra)


library(plyr)

# KS1
df1 <- ddply(CleanData,.(CleanData$TARGET_FLAG),transform,len=length(predict_
1))

ggplot(df1,aes(x=predict_1,color=CleanData$TARGET_FLAG)) + geom_step(aes(len=
len,y=..y.. * len),stat="ecdf")+

  labs(title= "KS -fit1", x="fit1 Score", y="cummulative # < score") +

  theme(
    title = element_text(size = 13, color = "black", face = "bold"),
    axis.text = element_text(colour = "black", size = 12),
    axis.text.y = element_text(colour = "black",size = 12),
    axis.title = element_text(size = 13, color = "black", face = "bold"),
    axis.title.y = element_text(size = 13, color = "black", face = "bold")
  )
```



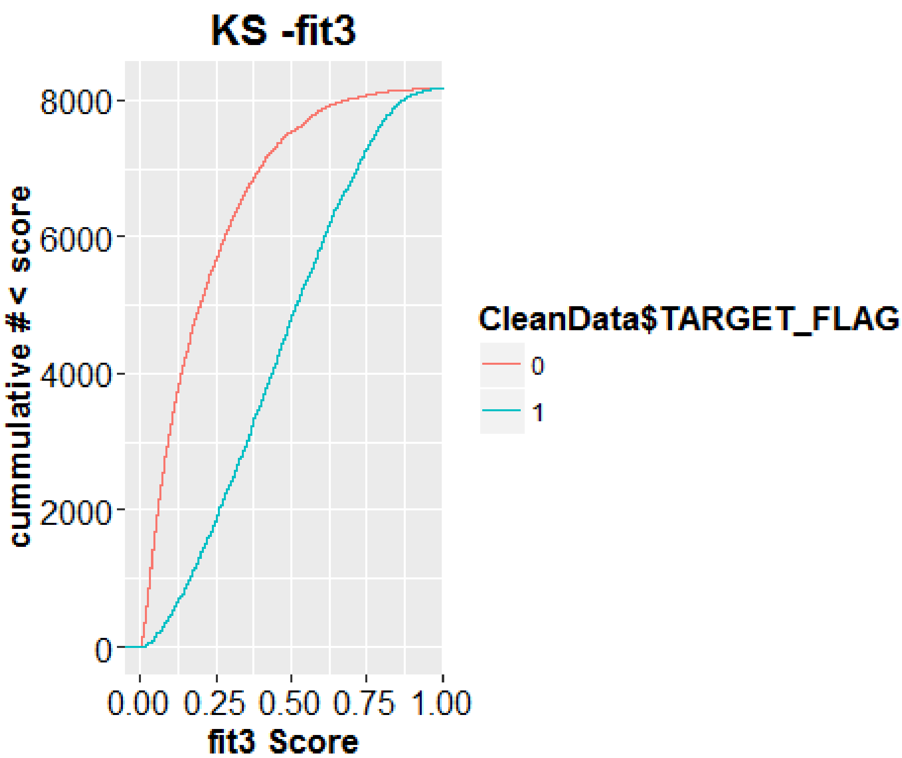

```
# KS2

df2 <- ddply(CleanData,.(CleanData$TARGET_FLAG),transform,len=length(predict_
2))
```

```
ggplot(df2,aes(x=predict_1,color=CleanData$TARGET_FLAG)) + geom_step(aes(len=
len,y=..y.. * len),stat="ecdf")+

  labs(title= "KS -fit2", x="fit2 Score", y="cummulative # < score") +

  theme(
    title = element_text(size = 13, color = "black", face = "bold"),
    axis.text = element_text(colour = "black", size = 12),
    axis.text.y = element_text(colour = "black",size = 12),
    axis.title = element_text(size = 13, color = "black", face = "bold"),
    axis.title.y = element_text(size = 13, color = "black", face = "bold")
  )
```




```
# KS3

df3 <- ddply(CleanData,.(CleanData$TARGET_FLAG),transform,len=length(predict_
3))

ggplot(df3,aes(x=predict_1,color=CleanData$TARGET_FLAG)) + geom_step(aes(len=
len,y=..y.. * len),stat="ecdf")+

  labs(title= "KS -fit3", x="fit3 Score", y="cummulative # < score") +

  theme(
    title = element_text(size = 13, color = "black", face = "bold"),
    axis.text = element_text(colour = "black", size = 12),
```

```
    axis.text.y = element_text(colour = "black",size = 12),
    axis.title = element_text(size = 13, color = "black", face = "bold"),
    axis.title.y = element_text(size = 13, color = "black", face = "bold")
  )
```




```
data_with_response <- data.frame(CleanData, predict_1,predict_2,predict_3)

# write.csv(data_with_response, file = "data_with_response.csv")

# Model 2:

# cumulative TRUE POSITIVE:5053/5919
# Cumulative FALSE POSITIVE  Rate :955/6008
# KS = 69.47%
```

Evaluate the models

```
# there are a number of pseudo R2 metrics that could be of value. Most notabl
e is McFadden's R2, which is defined as 1???[ln(LM)/ln(L0)] where ln(LM) is t
he log likelihood value for the fitted model and ln(L0) is the log likelihood
 for the null model with only an intercept as a predictor. The measure ranges
 from 0 to just under 1, with values closer to zero indicating that the model
 has no predictive power.

library(pscl)

pR2(fit1)  # look for 'McFadden'
```

```
##           llh       llhNull            G2      McFadden          r2ML
## -3661.1436348 -4707.6484705  2093.0096715     0.2222988     0.2262421
##          r2CU
##     0.3304854
```

Build a molde to predict claim amount

```
CleanData_claim <- subset(CleanData, CleanData$TARGET_AMT>0)
```

Best model selected by backward selection

```
# Backward stepwise selection

library(MASS)

fit <-  lm(TARGET_AMT ~
          MVR_PTS +
                    No_CLM_FREQ  +
                    No_HOME +
                    log_INCOME+
                    No_Income +
                    log_OLDCLAIM+
                    No_HOMEKIDS +
                    No_KIDSDRIV+
                    log_BLUEBOOK +
                    IMP_AGE+
                    log_CAR_AGE+
                    NewCar+
                    TIF +
                    IMP_YOJ +
                    log_TRAVTIME+
                    CAR_TYPE_MV  +
                    CAR_TYPE_PS+
                    CAR_USE_C +
                    HighEducation +
                    LowEducation +
                    JOB_WHITE_COLLAR +
                    JOB_BLUE_STUDENT+
                    MSTATUS_Y +
                    PARENT_Y +
                    RED_CAR_Y +
                    REVOKED_Y +
                    SEX_M +
                    URBANICITY_HU , data =CleanData_claim)

stepAIC(fit, direction="backward")

## Start:  AIC=38539.43
## TARGET_AMT ~ MVR_PTS + No_CLM_FREQ + No_HOME + log_INCOME + No_Income +
##     log_OLDCLAIM + No_HOMEKIDS + No_KIDSDRIV + log_BLUEBOOK +
##     IMP_AGE + log_CAR_AGE + NewCar + TIF + IMP_YOJ + log_TRAVTIME +
##     CAR_TYPE_MV + CAR_TYPE_PS + CAR_USE_C + HighEducation + LowEducation +
```

```
##     JOB_WHITE_COLLAR + JOB_BLUE_STUDENT + MSTATUS_Y + PARENT_Y +
##     RED_CAR_Y + REVOKED_Y + SEX_M + URBANICITY_HU
##
##                    Df  Sum of Sq          RSS   AIC
## - CAR_USE_C         1      40664 125535970329 38537
## - URBANICITY_HU     1     127727 125536057392 38537
## - IMP_YOJ           1     315707 125536245371 38537
## - log_TRAVTIME      1    1194476 125537124141 38537
## - No_KIDSDRIV       1    1202357 125537132022 38537
## - CAR_TYPE_PS       1    1798537 125537728202 38537
## - PARENT_Y          1    4749420 125540679085 38538
## - log_OLDCLAIM      1    6482957 125542412622 38538
## - No_CLM_FREQ       1    7465414 125543395078 38538
## - TIF               1    8753493 125544683158 38538
## - log_INCOME        1    9972735 125545902400 38538
## - RED_CAR_Y         1   13386115 125549315780 38538
## - JOB_BLUE_STUDENT  1   15467886 125551397551 38538
## - No_Income         1   16129076 125552058741 38538
## - CAR_TYPE_MV       1   21028611 125556958276 38538
## - No_HOMEKIDS       1   45385266 125581314931 38538
## - NewCar            1   75234825 125611164490 38539
## - IMP_AGE           1   76365405 125612295070 38539
## - LowEducation      1   96080000 125632009665 38539
## <none>                           125535929665 38539
## - HighEducation     1  123420958 125659350623 38540
## - JOB_WHITE_COLLAR  1  144422396 125680352060 38540
## - No_HOME           1  148828534 125684758199 38540
## - log_CAR_AGE       1  150002890 125685932555 38540
## - SEX_M             1  151523240 125687452904 38540
## - REVOKED_Y         1  166436991 125702366655 38540
## - MSTATUS_Y         1  179883078 125715812743 38541
## - MVR_PTS           1  192759115 125728688780 38541
## - log_BLUEBOOK      1 1249617291 126785546956 38559
##
## Step:  AIC=38537.43
## TARGET_AMT ~ MVR_PTS + No_CLM_FREQ + No_HOME + log_INCOME + No_Income +
##     log_OLDCLAIM + No_HOMEKIDS + No_KIDSDRIV + log_BLUEBOOK +
##     IMP_AGE + log_CAR_AGE + NewCar + TIF + IMP_YOJ + log_TRAVTIME +
##     CAR_TYPE_MV + CAR_TYPE_PS + HighEducation + LowEducation +
##     JOB_WHITE_COLLAR + JOB_BLUE_STUDENT + MSTATUS_Y + PARENT_Y +
##     RED_CAR_Y + REVOKED_Y + SEX_M + URBANICITY_HU
##
##                    Df  Sum of Sq          RSS   AIC
## - URBANICITY_HU     1     130742 125536101071 38535
## - IMP_YOJ           1     312685 125536283013 38535
## - log_TRAVTIME      1    1174662 125537144990 38535
## - No_KIDSDRIV       1    1199058 125537169386 38535
## - CAR_TYPE_PS       1    1843340 125537813668 38535
## - PARENT_Y          1    4757728 125540728057 38536
```

```
## - log_OLDCLAIM      1    6476386 125542446714 38536
## - No_CLM_FREQ       1    7461336 125543431664 38536
## - TIF               1    8860867 125544831195 38536
## - log_INCOME        1   10460659 125546430988 38536
## - RED_CAR_Y         1   13400030 125549370359 38536
## - No_Income         1   16702678 125552673007 38536
## - JOB_BLUE_STUDENT  1   18986123 125554956452 38536
## - CAR_TYPE_MV       1   21822189 125557792517 38536
## - No_HOMEKIDS       1   45502997 125581473326 38536
## - NewCar            1   75595041 125611565369 38537
## - IMP_AGE           1   77204560 125613174889 38537
## - LowEducation      1   96090364 125632060693 38537
## <none>                          125535970329 38537
## - HighEducation     1  123446902 125659417231 38538
## - No_HOME           1  148853269 125684823597 38538
## - log_CAR_AGE       1  150633652 125686603981 38538
## - JOB_WHITE_COLLAR  1  151112071 125687082400 38538
## - SEX_M             1  162737668 125698707997 38538
## - REVOKED_Y         1  166752235 125702722564 38538
## - MSTATUS_Y         1  179842414 125715812743 38539
## - MVR_PTS           1  193021897 125728992226 38539
## - log_BLUEBOOK      1 1334191529 126870161858 38558
##
## Step:  AIC=38535.43
## TARGET_AMT ~ MVR_PTS + No_CLM_FREQ + No_HOME + log_INCOME + No_Income +
##     log_OLDCLAIM + No_HOMEKIDS + No_KIDSDRIV + log_BLUEBOOK +
##     IMP_AGE + log_CAR_AGE + NewCar + TIF + IMP_YOJ + log_TRAVTIME +
##     CAR_TYPE_MV + CAR_TYPE_PS + HighEducation + LowEducation +
##     JOB_WHITE_COLLAR + JOB_BLUE_STUDENT + MSTATUS_Y + PARENT_Y +
##     RED_CAR_Y + REVOKED_Y + SEX_M
##
##                    Df  Sum of Sq          RSS   AIC
## - IMP_YOJ           1     308912 125536409982 38533
## - log_TRAVTIME      1    1131879 125537232950 38533
## - No_KIDSDRIV       1    1191042 125537292113 38533
## - CAR_TYPE_PS       1    1831031 125537932101 38533
## - PARENT_Y          1    4752287 125540853357 38534
## - log_OLDCLAIM      1    6608132 125542709202 38534
## - No_CLM_FREQ       1    7628202 125543729273 38534
## - TIF               1    8905018 125545006089 38534
## - log_INCOME        1   10598926 125546699997 38534
## - RED_CAR_Y         1   13441207 125549542278 38534
## - No_Income         1   16812462 125552913532 38534
## - JOB_BLUE_STUDENT  1   18969308 125555070379 38534
## - CAR_TYPE_MV       1   21749096 125557850166 38534
## - No_HOMEKIDS       1   45675570 125581776641 38534
## - NewCar            1   75482940 125611584010 38535
## - IMP_AGE           1   77532586 125613633657 38535
## - LowEducation      1   96045065 125632146136 38535
## <none>                          125536101071 38535
```

```
## - HighEducation     1  123318671 125659419741 38536
## - No_HOME           1  148831670 125684932741 38536
## - log_CAR_AGE       1  150511153 125686612223 38536
## - JOB_WHITE_COLLAR  1  151416005 125687517076 38536
## - SEX_M             1  163091328 125699192398 38536
## - REVOKED_Y         1  167622573 125703723643 38536
## - MSTATUS_Y         1  180746081 125716847152 38537
## - MVR_PTS           1  192900634 125729001705 38537
## - log_BLUEBOOK      1 1334971679 126871072750 38556
##
## Step:  AIC=38533.44
## TARGET_AMT ~ MVR_PTS + No_CLM_FREQ + No_HOME + log_INCOME + No_Income +
##     log_OLDCLAIM + No_HOMEKIDS + No_KIDSDRIV + log_BLUEBOOK +
##     IMP_AGE + log_CAR_AGE + NewCar + TIF + log_TRAVTIME + CAR_TYPE_MV +
##     CAR_TYPE_PS + HighEducation + LowEducation + JOB_WHITE_COLLAR +
##     JOB_BLUE_STUDENT + MSTATUS_Y + PARENT_Y + RED_CAR_Y + REVOKED_Y +
##     SEX_M
##
##                    Df  Sum of Sq          RSS   AIC
## - No_KIDSDRIV       1    1110961 125537520943 38531
## - log_TRAVTIME      1    1113592 125537523574 38531
## - CAR_TYPE_PS       1    1831303 125538241286 38531
## - PARENT_Y          1    4791450 125541201432 38532
## - log_OLDCLAIM      1    6550895 125542960878 38532
## - No_CLM_FREQ       1    7575795 125543985777 38532
## - TIF               1    8817487 125545227469 38532
## - log_INCOME        1   10844083 125547254065 38532
## - RED_CAR_Y         1   13331015 125549740998 38532
## - No_Income         1   16771661 125553181643 38532
## - JOB_BLUE_STUDENT  1   18928948 125555338930 38532
## - CAR_TYPE_MV       1   22009761 125558419743 38532
## - No_HOMEKIDS       1   46103888 125582513870 38532
## - NewCar            1   75391252 125611801234 38533
## - IMP_AGE           1   79144154 125615554136 38533
## - LowEducation      1   96612713 125633022695 38533
## <none>                           125536409982 38533
## - HighEducation     1  123227486 125659637468 38534
## - No_HOME           1  148523411 125684933393 38534
## - log_CAR_AGE       1  150310612 125686720595 38534
## - JOB_WHITE_COLLAR  1  152055163 125688465146 38534
## - SEX_M             1  162857907 125699267889 38534
## - REVOKED_Y         1  168051698 125704461680 38534
## - MSTATUS_Y         1  184283818 125720693800 38535
## - MVR_PTS           1  193016072 125729426054 38535
## - log_BLUEBOOK      1 1334921554 126871331537 38554
##
## Step:  AIC=38531.46
## TARGET_AMT ~ MVR_PTS + No_CLM_FREQ + No_HOME + log_INCOME + No_Income +
##     log_OLDCLAIM + No_HOMEKIDS + log_BLUEBOOK + IMP_AGE + log_CAR_AGE +
##     NewCar + TIF + log_TRAVTIME + CAR_TYPE_MV + CAR_TYPE_PS +
```

```
##     HighEducation + LowEducation + JOB_WHITE_COLLAR + JOB_BLUE_STUDENT +
##     MSTATUS_Y + PARENT_Y + RED_CAR_Y + REVOKED_Y + SEX_M
##
##                    Df  Sum of Sq          RSS   AIC
## - log_TRAVTIME      1    1098811 125538619754 38529
## - CAR_TYPE_PS       1    1888938 125539409881 38529
## - PARENT_Y          1    4902127 125542423070 38530
## - log_OLDCLAIM      1    6800010 125544320953 38530
## - No_CLM_FREQ       1    7858197 125545379140 38530
## - TIF               1    8652996 125546173939 38530
## - log_INCOME        1   11026964 125548547907 38530
## - RED_CAR_Y         1   13088162 125550609105 38530
## - No_Income         1   16910596 125554431539 38530
## - JOB_BLUE_STUDENT  1   19314181 125556835124 38530
## - CAR_TYPE_MV       1   22308875 125559829818 38530
## - No_HOMEKIDS       1   49861024 125587381967 38530
## - NewCar            1   75195211 125612716154 38531
## - IMP_AGE           1   79753905 125617274848 38531
## - LowEducation      1   95917682 125633438625 38531
## <none>                           125537520943 38531
## - HighEducation     1  123192640 125660713583 38532
## - No_HOME           1  149041257 125686562200 38532
## - log_CAR_AGE       1  149933980 125687454923 38532
## - JOB_WHITE_COLLAR  1  152838628 125690359571 38532
## - SEX_M             1  162610499 125700131442 38532
## - REVOKED_Y         1  170164675 125707685619 38532
## - MSTATUS_Y         1  183980470 125721501413 38533
## - MVR_PTS           1  193386852 125730907795 38533
## - log_BLUEBOOK      1 1333910023 126871430966 38552
##
## Step:  AIC=38529.48
## TARGET_AMT ~ MVR_PTS + No_CLM_FREQ + No_HOME + log_INCOME + No_Income +
##     log_OLDCLAIM + No_HOMEKIDS + log_BLUEBOOK + IMP_AGE + log_CAR_AGE +
##     NewCar + TIF + CAR_TYPE_MV + CAR_TYPE_PS + HighEducation +
##     LowEducation + JOB_WHITE_COLLAR + JOB_BLUE_STUDENT + MSTATUS_Y +
##     PARENT_Y + RED_CAR_Y + REVOKED_Y + SEX_M
##
##                    Df  Sum of Sq          RSS   AIC
## - CAR_TYPE_PS       1    1768749 125540388503 38528
## - PARENT_Y          1    4985059 125543604813 38528
## - log_OLDCLAIM      1    6844107 125545463861 38528
## - No_CLM_FREQ       1    7911263 125546531017 38528
## - TIF               1    8470996 125547090750 38528
## - log_INCOME        1   11073809 125549693563 38528
## - RED_CAR_Y         1   13204403 125551824157 38528
## - No_Income         1   16896303 125555516057 38528
## - JOB_BLUE_STUDENT  1   19604289 125558224043 38528
## - CAR_TYPE_MV       1   22337461 125560957215 38528
## - No_HOMEKIDS       1   50077171 125588696925 38528
## - NewCar            1   75126911 125613746665 38529
```

```
## - IMP_AGE            1   79477989 125618097742 38529
## - LowEducation       1   95542684 125634162438 38529
## <none>                            125538619754 38529
## - HighEducation      1  123186025 125661805778 38530
## - No_HOME            1  149522217 125688141970 38530
## - log_CAR_AGE        1  149984106 125688603859 38530
## - JOB_WHITE_COLLAR   1  152546367 125691166121 38530
## - SEX_M              1  162834902 125701454656 38530
## - REVOKED_Y          1  169923550 125708543303 38530
## - MSTATUS_Y          1  184507990 125723127744 38531
## - MVR_PTS            1  193017818 125731637572 38531
## - log_BLUEBOOK       1 1333021458 126871641212 38550
##
## Step:  AIC=38527.51
## TARGET_AMT ~ MVR_PTS + No_CLM_FREQ + No_HOME + log_INCOME + No_Income +
##     log_OLDCLAIM + No_HOMEKIDS + log_BLUEBOOK + IMP_AGE + log_CAR_AGE +
##     NewCar + TIF + CAR_TYPE_MV + HighEducation + LowEducation +
##     JOB_WHITE_COLLAR + JOB_BLUE_STUDENT + MSTATUS_Y + PARENT_Y +
##     RED_CAR_Y + REVOKED_Y + SEX_M
##
##                     Df  Sum of Sq          RSS   AIC
## - PARENT_Y           1    4947846 125545336348 38526
## - log_OLDCLAIM       1    6654103 125547042606 38526
## - No_CLM_FREQ        1    7750823 125548139326 38526
## - TIF                1    8518451 125548906954 38526
## - log_INCOME         1   10723146 125551111649 38526
## - RED_CAR_Y          1   13409992 125553798494 38526
## - No_Income          1   16506706 125556895209 38526
## - JOB_BLUE_STUDENT   1   19272173 125559660675 38526
## - CAR_TYPE_MV        1   20687835 125561076337 38526
## - No_HOMEKIDS        1   50341291 125590729793 38526
## - NewCar             1   75103965 125615492468 38527
## - IMP_AGE            1   80090235 125620478737 38527
## - LowEducation       1   96385743 125636774245 38527
## <none>                            125540388503 38528
## - HighEducation      1  123409287 125663797790 38528
## - No_HOME            1  149452899 125689841402 38528
## - log_CAR_AGE        1  150069827 125690458329 38528
## - JOB_WHITE_COLLAR   1  155467155 125695855658 38528
## - SEX_M              1  161125124 125701513627 38528
## - REVOKED_Y          1  170216730 125710605232 38528
## - MSTATUS_Y          1  184632888 125725021391 38529
## - MVR_PTS            1  192349487 125732737990 38529
## - log_BLUEBOOK       1 1407814025 126948202527 38550
##
## Step:  AIC=38525.59
## TARGET_AMT ~ MVR_PTS + No_CLM_FREQ + No_HOME + log_INCOME + No_Income +
##     log_OLDCLAIM + No_HOMEKIDS + log_BLUEBOOK + IMP_AGE + log_CAR_AGE +
##     NewCar + TIF + CAR_TYPE_MV + HighEducation + LowEducation +
##     JOB_WHITE_COLLAR + JOB_BLUE_STUDENT + MSTATUS_Y + RED_CAR_Y +
```

```
##     REVOKED_Y + SEX_M
##
##                    Df  Sum of Sq          RSS   AIC
## - log_OLDCLAIM      1    6802773 125552139122 38524
## - No_CLM_FREQ       1    7971564 125553307913 38524
## - TIF               1    9037232 125554373581 38524
## - log_INCOME        1   10823882 125556160230 38524
## - RED_CAR_Y         1   13536541 125558872889 38524
## - No_Income         1   16580065 125561916414 38524
## - JOB_BLUE_STUDENT  1   18728874 125564065222 38524
## - CAR_TYPE_MV       1   20220491 125565556839 38524
## - NewCar            1   77177647 125622513995 38525
## - IMP_AGE           1   81684975 125627021324 38525
## - LowEducation      1   96384790 125641721138 38525
## <none>                           125545336348 38526
## - HighEducation     1  122869293 125668205641 38526
## - No_HOMEKIDS       1  132365588 125677701937 38526
## - No_HOME           1  149059282 125694395630 38526
## - log_CAR_AGE       1  151981385 125697317733 38526
## - JOB_WHITE_COLLAR  1  155774995 125701111343 38526
## - SEX_M             1  160373021 125705709370 38526
## - REVOKED_Y         1  171313185 125716649534 38527
## - MVR_PTS           1  194420808 125739757157 38527
## - MSTATUS_Y         1  337215812 125882552160 38529
## - log_BLUEBOOK      1 1407061364 126952397712 38548
##
## Step:  AIC=38523.71
## TARGET_AMT ~ MVR_PTS + No_CLM_FREQ + No_HOME + log_INCOME + No_Income +
##     No_HOMEKIDS + log_BLUEBOOK + IMP_AGE + log_CAR_AGE + NewCar +
##     TIF + CAR_TYPE_MV + HighEducation + LowEducation + JOB_WHITE_COLLAR +
##     JOB_BLUE_STUDENT + MSTATUS_Y + RED_CAR_Y + REVOKED_Y + SEX_M
##
##                    Df  Sum of Sq          RSS   AIC
## - No_CLM_FREQ       1    2563034 125554702156 38522
## - TIF               1    9742847 125561881969 38522
## - log_INCOME        1   10699140 125562838262 38522
## - RED_CAR_Y         1   12698729 125564837851 38522
## - No_Income         1   16343412 125568482534 38522
## - JOB_BLUE_STUDENT  1   18123584 125570262705 38522
## - CAR_TYPE_MV       1   19875451 125572014573 38522
## - NewCar            1   77127762 125629266884 38523
## - IMP_AGE           1   83352302 125635491424 38523
## - LowEducation      1   94482247 125646621369 38523
## <none>                           125552139122 38524
## - HighEducation     1  123610966 125675750088 38524
## - No_HOMEKIDS       1  132346099 125684485221 38524
## - No_HOME           1  148997274 125701136396 38524
## - log_CAR_AGE       1  152002596 125704141717 38524
## - JOB_WHITE_COLLAR  1  154093635 125706232756 38524
## - SEX_M             1  157054586 125709193708 38524
```

```
## - REVOKED_Y         1  191472844 125743611966 38525
## - MVR_PTS           1  193006856 125745145977 38525
## - MSTATUS_Y         1  339023961 125891163083 38528
## - log_BLUEBOOK      1 1400962320 126953101442 38546
##
## Step:  AIC=38521.75
## TARGET_AMT ~ MVR_PTS + No_HOME + log_INCOME + No_Income + No_HOMEKIDS +
##     log_BLUEBOOK + IMP_AGE + log_CAR_AGE + NewCar + TIF + CAR_TYPE_MV +
##     HighEducation + LowEducation + JOB_WHITE_COLLAR + JOB_BLUE_STUDENT +
##     MSTATUS_Y + RED_CAR_Y + REVOKED_Y + SEX_M
##
##                    Df  Sum of Sq          RSS   AIC
## - TIF               1   10032712 125564734868 38520
## - log_INCOME        1   10682308 125565384464 38520
## - RED_CAR_Y         1   12490416 125567192572 38520
## - No_Income         1   16260985 125570963141 38520
## - JOB_BLUE_STUDENT  1   18207917 125572910073 38520
## - CAR_TYPE_MV       1   19874499 125574576655 38520
## - NewCar            1   78005307 125632707463 38521
## - IMP_AGE           1   83254926 125637957082 38521
## - LowEducation      1   93940671 125648642827 38521
## <none>                           125554702156 38522
## - HighEducation     1  124374583 125679076739 38522
## - No_HOMEKIDS       1  133032131 125687734287 38522
## - No_HOME           1  149590683 125704292839 38522
## - JOB_WHITE_COLLAR  1  152754310 125707456466 38522
## - log_CAR_AGE       1  153945070 125708647226 38522
## - SEX_M             1  156948383 125711650539 38522
## - REVOKED_Y         1  192879165 125747581321 38523
## - MVR_PTS           1  204714964 125759417120 38523
## - MSTATUS_Y         1  341024010 125895726166 38526
## - log_BLUEBOOK      1 1401964084 126956666240 38544
##
## Step:  AIC=38519.92
## TARGET_AMT ~ MVR_PTS + No_HOME + log_INCOME + No_Income + No_HOMEKIDS +
##     log_BLUEBOOK + IMP_AGE + log_CAR_AGE + NewCar + CAR_TYPE_MV +
##     HighEducation + LowEducation + JOB_WHITE_COLLAR + JOB_BLUE_STUDENT +
##     MSTATUS_Y + RED_CAR_Y + REVOKED_Y + SEX_M
##
##                    Df  Sum of Sq          RSS   AIC
## - log_INCOME        1   10773410 125575508278 38518
## - RED_CAR_Y         1   12269176 125577004045 38518
## - No_Income         1   16339771 125581074639 38518
## - JOB_BLUE_STUDENT  1   18122748 125582857616 38518
## - CAR_TYPE_MV       1   19509740 125584244608 38518
## - NewCar            1   77407715 125642142583 38519
## - IMP_AGE           1   82991337 125647726205 38519
## - LowEducation      1   94195991 125658930859 38520
## <none>                           125564734868 38520
## - HighEducation     1  123588701 125688323569 38520
```

```
## - No_HOMEKIDS       1  131561269 125696296137 38520
## - No_HOME           1  148662386 125713397254 38520
## - JOB_WHITE_COLLAR  1  149465813 125714200682 38520
## - log_CAR_AGE       1  153099847 125717834715 38521
## - SEX_M             1  157401444 125722136312 38521
## - REVOKED_Y         1  191582078 125756316947 38521
## - MVR_PTS           1  207597326 125772332194 38521
## - MSTATUS_Y         1  336205222 125900940091 38524
## - log_BLUEBOOK      1 1402173284 126966908152 38542
##
## Step:  AIC=38518.11
## TARGET_AMT ~ MVR_PTS + No_HOME + No_Income + No_HOMEKIDS + log_BLUEBOOK +
##     IMP_AGE + log_CAR_AGE + NewCar + CAR_TYPE_MV + HighEducation +
##     LowEducation + JOB_WHITE_COLLAR + JOB_BLUE_STUDENT + MSTATUS_Y +
##     RED_CAR_Y + REVOKED_Y + SEX_M
##
##                    Df  Sum of Sq          RSS   AIC
## - RED_CAR_Y         1   11728102 125587236380 38516
## - No_Income         1   13171432 125588679710 38516
## - JOB_BLUE_STUDENT  1   16404451 125591912730 38516
## - CAR_TYPE_MV       1   19048487 125594556765 38516
## - NewCar            1   76990799 125652499078 38517
## - IMP_AGE           1   80065402 125655573680 38517
## - LowEducation      1   84788189 125660296467 38518
## <none>                           125575508278 38518
## - HighEducation     1  117307092 125692815371 38518
## - No_HOMEKIDS       1  130154645 125705662923 38518
## - No_HOME           1  139960255 125715468534 38519
## - log_CAR_AGE       1  152428245 125727936524 38519
## - SEX_M             1  153043239 125728551518 38519
## - JOB_WHITE_COLLAR  1  157886722 125733395001 38519
## - REVOKED_Y         1  190764532 125766272810 38519
## - MVR_PTS           1  206153506 125781661785 38520
## - MSTATUS_Y         1  326260386 125901768665 38522
## - log_BLUEBOOK      1 1419922864 126995431142 38540
##
## Step:  AIC=38516.31
## TARGET_AMT ~ MVR_PTS + No_HOME + No_Income + No_HOMEKIDS + log_BLUEBOOK +
##     IMP_AGE + log_CAR_AGE + NewCar + CAR_TYPE_MV + HighEducation +
##     LowEducation + JOB_WHITE_COLLAR + JOB_BLUE_STUDENT + MSTATUS_Y +
##     REVOKED_Y + SEX_M
##
##                    Df  Sum of Sq          RSS   AIC
## - No_Income         1   13725420 125600961800 38515
## - JOB_BLUE_STUDENT  1   15822479 125603058859 38515
## - CAR_TYPE_MV       1   18846453 125606082833 38515
## - NewCar            1   79348030 125666584410 38516
## - LowEducation      1   82612289 125669848669 38516
## - IMP_AGE           1   82762229 125669998609 38516
## <none>                           125587236380 38516
```

```
## - HighEducation     1  117751965 125704988345 38516
## - No_HOMEKIDS       1  130988579 125718224959 38517
## - No_HOME           1  141479042 125728715422 38517
## - log_CAR_AGE       1  155019750 125742256130 38517
## - JOB_WHITE_COLLAR  1  157966330 125745202710 38517
## - SEX_M             1  177538361 125764774741 38517
## - REVOKED_Y         1  189354075 125776590455 38518
## - MVR_PTS           1  203658528 125790894908 38518
## - MSTATUS_Y         1  327707552 125914943932 38520
## - log_BLUEBOOK      1 1420447720 127007684100 38539
##
## Step:  AIC=38514.54
## TARGET_AMT ~ MVR_PTS + No_HOME + No_HOMEKIDS + log_BLUEBOOK +
##     IMP_AGE + log_CAR_AGE + NewCar + CAR_TYPE_MV + HighEducation +
##     LowEducation + JOB_WHITE_COLLAR + JOB_BLUE_STUDENT + MSTATUS_Y +
##     REVOKED_Y + SEX_M
##
##                    Df  Sum of Sq          RSS   AIC
## - JOB_BLUE_STUDENT  1   14914129 125615875929 38513
## - CAR_TYPE_MV       1   18973890 125619935689 38513
## - NewCar            1   77619175 125678580975 38514
## - IMP_AGE           1   80712996 125681674796 38514
## - LowEducation      1   82534327 125683496126 38514
## - HighEducation     1  114656364 125715618164 38515
## <none>                           125600961800 38515
## - No_HOMEKIDS       1  128408707 125729370507 38515
## - JOB_WHITE_COLLAR  1  149312684 125750274483 38515
## - log_CAR_AGE       1  153054148 125754015947 38515
## - No_HOME           1  167002399 125767964199 38515
## - REVOKED_Y         1  186176848 125787138648 38516
## - SEX_M             1  192001086 125792962885 38516
## - MVR_PTS           1  201671884 125802633683 38516
## - MSTATUS_Y         1  348750276 125949712076 38519
## - log_BLUEBOOK      1 1524018959 127124980759 38538
##
## Step:  AIC=38512.8
## TARGET_AMT ~ MVR_PTS + No_HOME + No_HOMEKIDS + log_BLUEBOOK +
##     IMP_AGE + log_CAR_AGE + NewCar + CAR_TYPE_MV + HighEducation +
##     LowEducation + JOB_WHITE_COLLAR + MSTATUS_Y + REVOKED_Y +
##     SEX_M
##
##                    Df  Sum of Sq          RSS   AIC
## - CAR_TYPE_MV       1   18793324 125634669252 38511
## - LowEducation      1   73890282 125689766211 38512
## - NewCar            1   77249594 125693125523 38512
## - IMP_AGE           1   80987178 125696863107 38512
## - HighEducation     1  102388085 125718264014 38513
## <none>                           125615875929 38513
## - No_HOMEKIDS       1  126130762 125742006691 38513
## - log_CAR_AGE       1  150873871 125766749800 38513
```

```
## - No_HOME           1  155842578 125771718506 38513
## - JOB_WHITE_COLLAR  1  160833358 125776709287 38514
## - REVOKED_Y         1  183646896 125799522825 38514
## - MVR_PTS           1  203829051 125819704980 38514
## - SEX_M             1  203944995 125819820924 38514
## - MSTATUS_Y         1  337348864 125953224793 38517
## - log_BLUEBOOK      1 1522455512 127138331441 38537
##
## Step:  AIC=38511.12
## TARGET_AMT ~ MVR_PTS + No_HOME + No_HOMEKIDS + log_BLUEBOOK +
##     IMP_AGE + log_CAR_AGE + NewCar + HighEducation + LowEducation +
##     JOB_WHITE_COLLAR + MSTATUS_Y + REVOKED_Y + SEX_M
##
##                    Df  Sum of Sq          RSS   AIC
## - LowEducation      1   74933385 125709602638 38510
## - NewCar            1   76159601 125710828853 38510
## - IMP_AGE           1   83502690 125718171942 38511
## - HighEducation     1  108412018 125743081270 38511
## <none>                           125634669252 38511
## - No_HOMEKIDS       1  125410522 125760079775 38511
## - log_CAR_AGE       1  149592732 125784261985 38512
## - No_HOME           1  158127388 125792796640 38512
## - JOB_WHITE_COLLAR  1  172313402 125806982654 38512
## - REVOKED_Y         1  179860545 125814529797 38512
## - SEX_M             1  186280577 125820949829 38512
## - MVR_PTS           1  209714748 125844384001 38513
## - MSTATUS_Y         1  333682108 125968351360 38515
## - log_BLUEBOOK      1 1523266970 127157936223 38535
##
## Step:  AIC=38510.4
## TARGET_AMT ~ MVR_PTS + No_HOME + No_HOMEKIDS + log_BLUEBOOK +
##     IMP_AGE + log_CAR_AGE + NewCar + HighEducation + JOB_WHITE_COLLAR +
##     MSTATUS_Y + REVOKED_Y + SEX_M
##
##                    Df  Sum of Sq          RSS   AIC
## - NewCar            1   57476527 125767079165 38509
## - IMP_AGE           1   85395714 125794998352 38510
## - log_CAR_AGE       1  104493135 125814095773 38510
## <none>                           125709602638 38510
## - No_HOMEKIDS       1  122687074 125832289712 38511
## - JOB_WHITE_COLLAR  1  151580770 125861183408 38511
## - HighEducation     1  156708024 125866310662 38511
## - No_HOME           1  165867497 125875470135 38511
## - REVOKED_Y         1  180295699 125889898337 38511
## - SEX_M             1  197874279 125907476917 38512
## - MVR_PTS           1  217587998 125927190636 38512
## - MSTATUS_Y         1  365439881 126075042519 38515
## - log_BLUEBOOK      1 1607573837 127317176474 38536
##
## Step:  AIC=38509.39
```

```
## TARGET_AMT ~ MVR_PTS + No_HOME + No_HOMEKIDS + log_BLUEBOOK +
##     IMP_AGE + log_CAR_AGE + HighEducation + JOB_WHITE_COLLAR +
##     MSTATUS_Y + REVOKED_Y + SEX_M
##
##                    Df  Sum of Sq          RSS   AIC
## - IMP_AGE           1   80493634 125847572799 38509
## - log_CAR_AGE       1   85865334 125852944499 38509
## - HighEducation     1  105633839 125872713004 38509
## <none>                           125767079165 38509
## - No_HOMEKIDS       1  119585825 125886664990 38509
## - JOB_WHITE_COLLAR  1  154372269 125921451434 38510
## - No_HOME           1  157611258 125924690423 38510
## - REVOKED_Y         1  176216364 125943295529 38510
## - SEX_M             1  202068419 125969147584 38511
## - MVR_PTS           1  214717794 125981796959 38511
## - MSTATUS_Y         1  354300824 126121379989 38513
## - log_BLUEBOOK      1 1592327303 127359406468 38534
##
## Step:  AIC=38508.76
## TARGET_AMT ~ MVR_PTS + No_HOME + No_HOMEKIDS + log_BLUEBOOK +
##     log_CAR_AGE + HighEducation + JOB_WHITE_COLLAR + MSTATUS_Y +
##     REVOKED_Y + SEX_M
##
##                    Df  Sum of Sq          RSS   AIC
## - No_HOMEKIDS       1   58051324 125905624123 38508
## - log_CAR_AGE       1   85504537 125933077336 38508
## <none>                           125847572799 38509
## - HighEducation     1  124827586 125972400385 38509
## - JOB_WHITE_COLLAR  1  137631203 125985204002 38509
## - No_HOME           1  165469898 126013042697 38510
## - REVOKED_Y         1  169927919 126017500718 38510
## - SEX_M             1  192075279 126039648077 38510
## - MVR_PTS           1  216021835 126063594634 38510
## - MSTATUS_Y         1  319904777 126167477576 38512
## - log_BLUEBOOK      1 1695073875 127542646674 38536
##
## Step:  AIC=38507.76
## TARGET_AMT ~ MVR_PTS + No_HOME + log_BLUEBOOK + log_CAR_AGE +
##     HighEducation + JOB_WHITE_COLLAR + MSTATUS_Y + REVOKED_Y +
##     SEX_M
##
##                    Df  Sum of Sq          RSS   AIC
## - log_CAR_AGE       1   89860792 125995484914 38507
## - HighEducation     1  114040812 126019664934 38508
## <none>                           125905624123 38508
## - JOB_WHITE_COLLAR  1  130555112 126036179235 38508
## - No_HOME           1  153557471 126059181594 38508
## - REVOKED_Y         1  164603531 126070227653 38509
## - SEX_M             1  171132127 126076756250 38509
## - MVR_PTS           1  226251634 126131875757 38510
```

```
## - MSTATUS_Y         1  301721878 126207346001 38511
## - log_BLUEBOOK      1 1668783165 127574407288 38534
##
## Step:  AIC=38507.29
## TARGET_AMT ~ MVR_PTS + No_HOME + log_BLUEBOOK + HighEducation +
##     JOB_WHITE_COLLAR + MSTATUS_Y + REVOKED_Y + SEX_M
##
##                    Df  Sum of Sq          RSS   AIC
## - HighEducation     1   59494031 126054978945 38506
## <none>                           125995484914 38507
## - JOB_WHITE_COLLAR  1  141661520 126137146434 38508
## - No_HOME           1  153868728 126149353642 38508
## - REVOKED_Y         1  168123428 126163608343 38508
## - SEX_M             1  181286562 126176771477 38508
## - MVR_PTS           1  237667216 126233152131 38509
## - MSTATUS_Y         1  293073769 126288558684 38510
## - log_BLUEBOOK      1 1636411182 127631896097 38533
##
## Step:  AIC=38506.31
## TARGET_AMT ~ MVR_PTS + No_HOME + log_BLUEBOOK + JOB_WHITE_COLLAR +
##     MSTATUS_Y + REVOKED_Y + SEX_M
##
##                    Df  Sum of Sq          RSS   AIC
## - JOB_WHITE_COLLAR  1   83221627 126138200572 38506
## <none>                           126054978945 38506
## - No_HOME           1  161780146 126216759092 38507
## - REVOKED_Y         1  164451254 126219430199 38507
## - SEX_M             1  194656530 126249635476 38508
## - MVR_PTS           1  238675074 126293654019 38508
## - MSTATUS_Y         1  301721725 126356700671 38509
## - log_BLUEBOOK      1 1791362236 127846341182 38535
##
## Step:  AIC=38505.73
## TARGET_AMT ~ MVR_PTS + No_HOME + log_BLUEBOOK + MSTATUS_Y + REVOKED_Y +
##     SEX_M
##
##                Df  Sum of Sq          RSS   AIC
## <none>                       126138200572 38506
## - No_HOME       1  151278812 126289479385 38506
## - REVOKED_Y     1  158076374 126296276947 38506
## - SEX_M         1  225105643 126363306216 38508
## - MVR_PTS       1  253537468 126391738041 38508
## - MSTATUS_Y     1  282904015 126421104588 38509
## - log_BLUEBOOK  1 1728255452 127866456025 38533

##
## Call:
## lm(formula = TARGET_AMT ~ MVR_PTS + No_HOME + log_BLUEBOOK +
##     MSTATUS_Y + REVOKED_Y + SEX_M, data = CleanData_claim)
##
```

```
## Coefficients:
##  (Intercept)       MVR_PTS       No_HOME  log_BLUEBOOK     MSTATUS_Y
##      -6893.0         133.3        -667.5        1366.2        -889.0
##    REVOKED_Y         SEX_M
##       -672.0         653.0
```

```
bestfit_claim <- lm(TARGET_AMT ~
          MVR_PTS +
                    log_BLUEBOOK +
                    REVOKED_Y +
                    SEX_M ,
                 data =CleanData_claim)

summary(bestfit_claim)
```

```
##
## Call:
## lm(formula = TARGET_AMT ~ MVR_PTS + log_BLUEBOOK + REVOKED_Y +
##     SEX_M, data = CleanData_claim)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -7318  -3184  -1619    423 100198
##
## Coefficients:
##              Estimate Std. Error t value     Pr(>|t|)
## (Intercept)  -8013.70    2363.73  -3.390     0.000711 ***
## MVR_PTS        129.08      64.18   2.011     0.044408 *
## log_BLUEBOOK  1413.24     251.09   5.628 0.0000000206 ***
## REVOKED_Y     -682.88     409.51  -1.668     0.095554 .
## SEX_M          642.61     333.98   1.924     0.054474 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7674 on 2147 degrees of freedom
## Multiple R-squared:  0.01998,    Adjusted R-squared:  0.01816
## F-statistic: 10.95 on 4 and 2147 DF,  p-value: 0.000000008712
```