Assignment #2

Nitin Gaonkar PREDICT 411 section 56

INTRODUCTION

The purpose of this project is to analyze the data of the auto insurance and build logistic models. Each record in the given data represents a customer at an auto insurance company, each record has two target variables. The first target variable, TARGET_FLAG, is 1 or a 0. A '1' means that the person was in car crash. A zero means that the person was not in a car cash, we will build three models using logistics regression and then the models will be compared to get the best fit. After getting the best model it will be further analyzed to see if its the best fit.

1. Data Exploration:

The auto insurance dataset has around 8000 records and each record represents a Customer at an auto insurance company. each record has two target variables. The first target variable, TARGET_FLAG, is 1 or a 0. A '1' means that the person was in car crash. A zero means that the person was not in a car cash. The second target variable is TARGET_AMT. This value is zero if the person did not crash their car. But if they did crash their car, this number will be a value greater than zero. There are three data sets provided, one is the training data set containing around 8000 observations, the other is the test data set and the last one is the random data set.

we begin our data exploration by examining the data dictionary and the definitions given in the dictionary, after observing the data dictionary:

VARIABLE NAME	TYPE	DEFINITION
AGE	continuous	Age of Driver
BLUEBOOK	continuous	Value of Vehicle
CAR_AGE	continuous	Vehicle Age
CAR_TYPE	categorical	Type of Car
CAR_USE	categorical	Vehicle Use
CLM_FREQ	continuous	Claims(Past 5 Years)
EDUCATION	categorical	Max Education Level
HOMEKIDS	continuous	Children @Home
HOME_VAL	continuous	Home Value
INCOME	continuous	Income
JOB	categorical	Job Category
KIDSDRIV	categorical	Driving Children

MSTATUS	categorical	Marital Status
MVR_PTS	continuous	Motor Vehicle Record Points
OLDCLAIM	continuous	Total Claims(Past 5 Years)
PARENT1	categorical	Single Parent
RED_CAR	categorical	A Red Car
REVOKED	categorical	License Revoked (Past 7 Years)
SEX	categorical	Gender
TIF	continuous	Time in Force
TRAVTIME	continuous	Distance to Work
URBANICITY	categorical	Home/Work Area
YOJ	continuous	Years on Job

a. Just to give a bit insight on the data, I have calculated and listed the mean and the standard deviation along with the min and max value of each variable in the below table.

The MEANS Procedure

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
INDEX		8161	5151.87	2978.89	1.0000000	10302.00
TARGET_FLAG		8161	0.2638157	0.4407276	0	1.0000000
TARGET_AMT		8161	1504.32	4704.03	0	107586.14
KIDSDRIV	#Driving Children	8161	0.1710575	0.5115341	0	4.0000000
AGE	Age	8155	44.7903127	8.6275895	16.0000000	81.0000000
HOMEKIDS	#Children @Home	8161	0.7212351	1.1163233	0	5.0000000
YOJ	Years on Job	7707	10.4992864	4.0924742	0	23.0000000
INCOME	Income	7716	61898.10	47572.69	0	367030.26
HOME_VAL	Home Value	7697	154867.29	129123.78	0	885282.34
TRAVTIME	Distance to Work	8161	33.4887972	15.9047470	5.0000000	142.1206304
BLUEBOOK	Value of Vehicle	8161	15709.90	8419.73	1500.00	69740.00
TIF	Time in Force	8161	5.3513050	4.1466353	1.0000000	25.0000000
OLDCLAIM	Total Claims(Past 5 Years)	8161	4037.08	8777.14	0	57037.00
CLM_FREQ	#Claims(Past 5 Years)	8161	0.7985541	1.1584527	0	5.0000000
MVR_PTS	Motor Vehicle Record Points	8161	1.6955030	2.1471117	0	13.0000000
CAR_AGE	Vehicle Age	7651	8.3283231	5.7007424	-3.0000000	28.0000000

Table 1

So let's focus on the continuous variables and try to explore those variables and see get the means and the missing values from the data:

b. Missing values:

Based on the above table we can tell that in the target flag 0 we have around 5 variables which have missing values and for the target flag 1 we have around 5 variables which are missing the values. We will use imputation where we use the the mean for the missing values. We will further analyze the variables by creating the histogram as well tests for normality.
From the table 2 we can see that we have missing values for the variables like car_age, home_val, income,yoj.

TARGET_FLAG	N Obs	Variable	Label	N	N Miss	Mean	Std Dev	Range
0	6008	AGE	Age	6007	1	45.3227901	8.2022705	65.0000000
		BLUEBOOK	Value of Vehicle	6008	0	16230.95	8401.95	68240.00
		CAR_AGE	Vehicle Age	5640	368	8.6709220	5.7201267	28.0000000
		CLM_FREQ	#Claims(Past 5 Years)	6008	0	0.6486352	1.0860488	5.0000000
		HOMEKIDS	#Children @Home	6008	0	0.6439747	1.0762090	5.0000000
		HOME_VAL	Home Value	5665	343	169075.41	129938.83	885282.34
		INCOME	Income	5673	335	65951.97	48552.20	367030.26
		MVR_PTS	Motor Vehicle Record Points	6008	0	1.4137816	1.8916611	11.0000000
		OLDCLAIM	Total Claims(Past 5 Years)	6008	0	3311.59	8143.61	53986.00
		TIF	Time in Force	6008	0	5.5557590	4.2020970	24.0000000
		TRAVTIME	Distance to Work	6008	0	33.0303446	16.1312900	137.1206304
		YOJ	Years on Job	5677	331	10.6718337	3.9175259	23.0000000
1	2153	AGE	Age	2148	5	43.3012104	9.5646287	60.0000000
		BLUEBOOK	Value of Vehicle	2153	0	14255.90	8299.81	60740.00
		CAR_AGE	Vehicle Age	2011	142	7.3674789	5.5353874	28.0000000
		CLM_FREQ	#Claims(Past 5 Years)	2153	0	1.2169066	1.2483641	5.0000000
		HOMEKIDS	#Children @Home	2153	0	0.9368323	1.1954470	5.0000000
		HOME_VAL	Home Value	2032	121	115256.55	118150.14	750455.22
		INCOME	Income	2043	110	50641.30	42782.04	320126.98
		MVR_PTS	Motor Vehicle Record Points	2153	0	2.4816535	2.5791851	13.0000000
		OLDCLAIM	Total Claims(Past 5 Years)	2153	0	6061.55	10071.09	57037.00
		TIF	Time in Force	2153	0	4.7807710	3.9329017	20.0000000
		TRAVTIME	Distance to Work	2153	0	34.7681203	15.1853855	91.6143255
		YOJ	Years on Job	2030	123	10.0167488	4.5122598	19.0000000

Table 2

c. Below table gives us the correlation of the variables with the target flag: From the below table we can see that none of the variable are highly correlated with the targetflag

1 With Variables:	TARGET_FLAG
15 Variables:	INDEX TARGET_AMT KIDSDRIV AGE HOMEKIDS YOJ INCOME HOME_VAL TRAVTIME BLUEBOOK TIF OLDCLAIM CLM_FREQ MVR_PTS CAR_AGE

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
TARGET_FLAG	8161	0.26382	0.44073	2153	0	1.00000	
INDEX	8161	5152	2979	42044392	1.00000	10302	
TARGET_AMT	8161	1504	4704	12276793	0	107586	
KIDSDRIV	8161	0.17106	0.51153	1396	0	4.00000	#Driving Children
AGE	8155	44.79031	8.62759	365265	16.00000	81.00000	Age
HOMEKIDS	8161	0.72124	1.11632	5886	0	5.00000	#Children @Home
YOJ	7707	10.49929	4.09247	80918	0	23.00000	Years on Job
INCOME	7716	61898	47573	477605720	0	367030	Income
HOME_VAL	7697	154867	129124	1192013528	0	885282	Home Value
TRAVTIME	8161	33.48880	15.90475	273302	5.00000	142.12063	Distance to Work
BLUEBOOK	8161	15710	8420	128208490	1500	69740	Value of Vehicle
TIF	8161	5.35130	4.14664	43672	1.00000	25.00000	Time in Force
OLDCLAIM	8161	4037	8777	32946579	0	57037	Total Claims(Past 5 Years)
CLM_FREQ	8161	0.79855	1.15845	6517	0	5.00000	#Claims(Past 5 Years)
MVR_PTS	8161	1.69550	2.14711	13837	0	13.00000	Motor Vehicle Record Points
CAR_AGE	7651	8.32832	5.70074	63720	-3.00000	28.00000	Vehicle Age

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0
Number of Observations

	INDEX	TARGET_AMT	KIDSDRIV	AGE	HOMEKIDS	YOJ	INCOME	HOME_VAL	TRAVTIME	BLUEBOOK	TIF	OLDCLAIM	CLM_FREQ	MVR_PTS
TARGET_FLAG	-0.00167 0.8801 8161	0.53425 <.0001 8161	0.10367 <.0001 8161	-0.10322 <.0001 8155	0.11562 <.0001 8161	-0.07051 <.0001 7707	-0.14201 <.0001 7716	-0.18374 <.0001 7697	0.04815 <.0001 8161	-0.10338 <.0001 8161	-0.08237 <.0001 8161	0.13808 <.0001 8161	0.21620 <.0001 8161	0.21920 <.0001 8161

d. In this section we will explore the distribution of variables, after going through all the distribution of the variables we could see that there are few variables which have extreme values and the high skewness. Below we will discuss few of the variables with the extreme values and skewness.

INCOME:

As we can observe from the below graph that the mean is around 5151.86766 the data is normally distributed as the skewness is about 0.00200461 which is good.

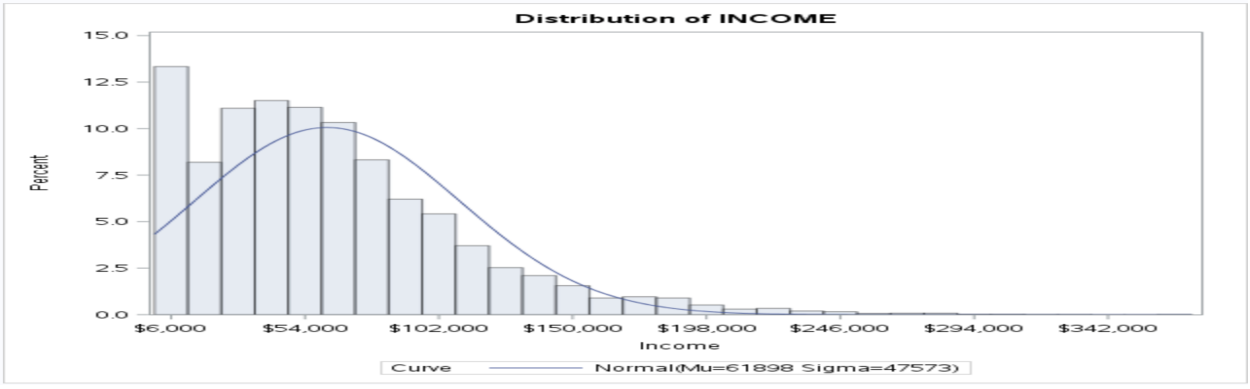


Fig:1 distributions of INCOME

MVR_PTS:

As we can observe from the below graph that the mean is around 1.695503 but there is long tail for this graph and the skewness is about 2.14711174 which is high, thus we would definitely would require to have a look at this variable.

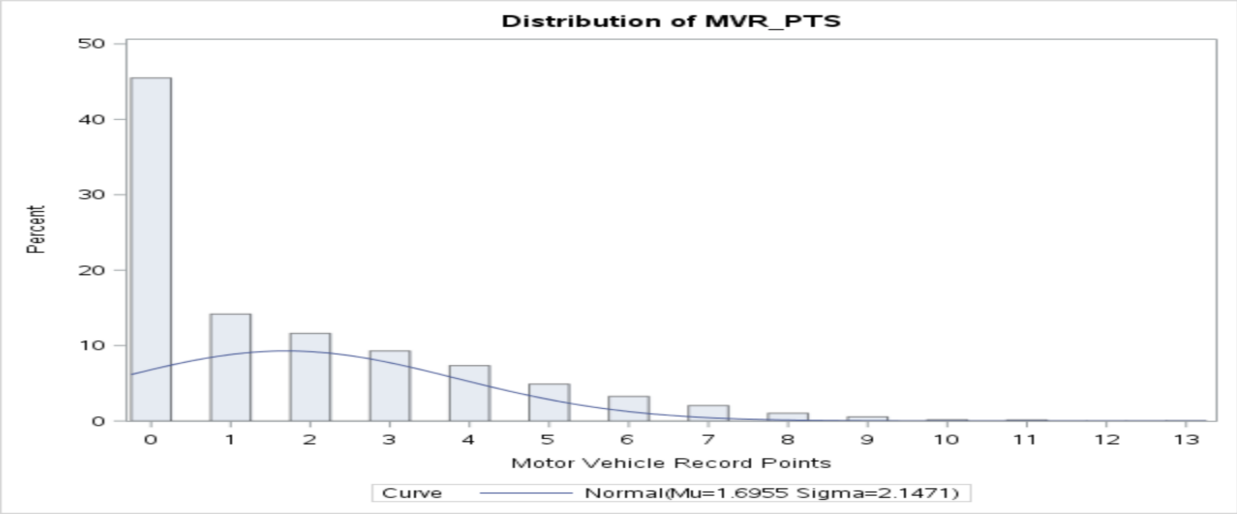


Fig :2 distributions of mvr_pts

HOMEKIDS

As we can observe from the below graph that the mean is around 0.72123514 but there is long tail for this graph and the skewness is about 1.34162023 which is high, thus we would definitely would require to have a look at this variable as well.

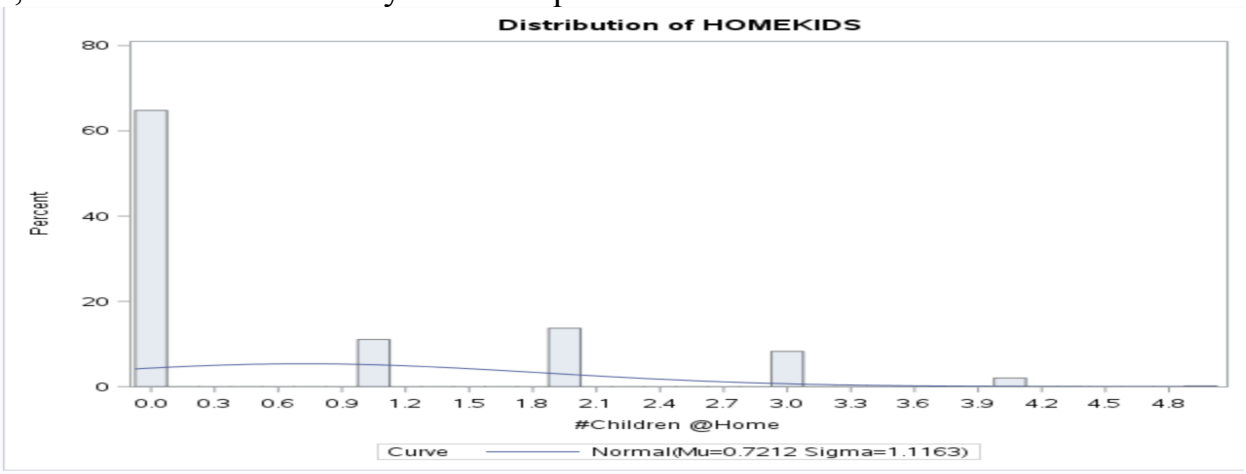


Fig :3 distribution of Homekids

OLDCLAIM:

As we can observe from the below graph that the mean is around 4037.07622 but there is long tail for this graph and the skewness is about 3.12018688 which is high, thus we would definitely would require to have a look at this variable as well.

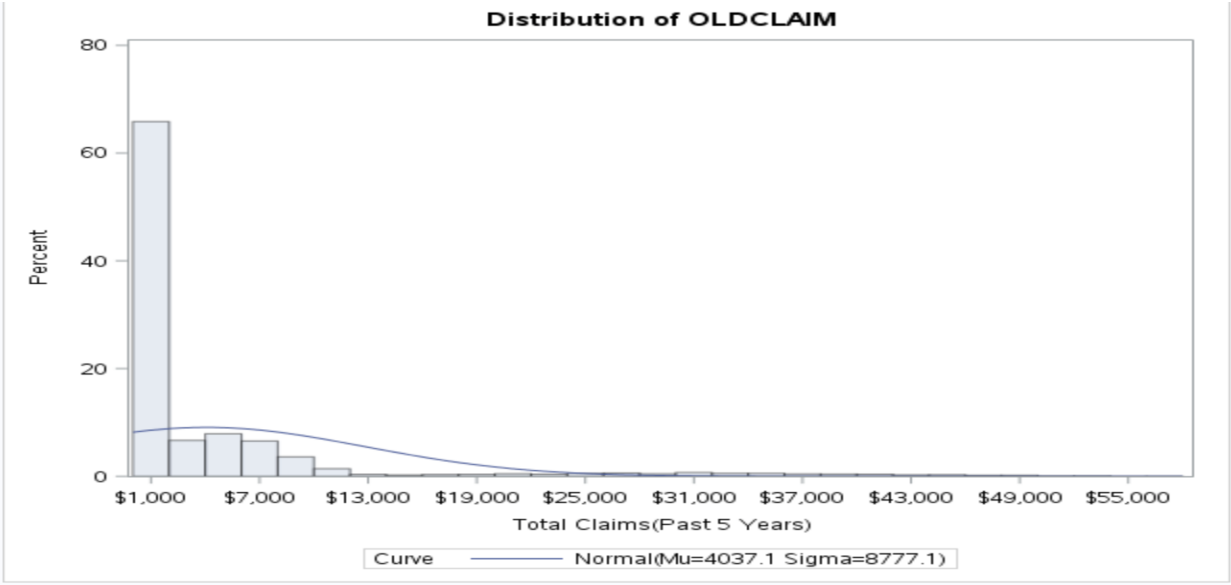


Fig:4 distributions of OLDCLAIM

HOMEVAL:

As we can observe from the below graph that the mean is around 154867.29 but there is long tail for this graph and the skewness is about 0.48878559 which is good.

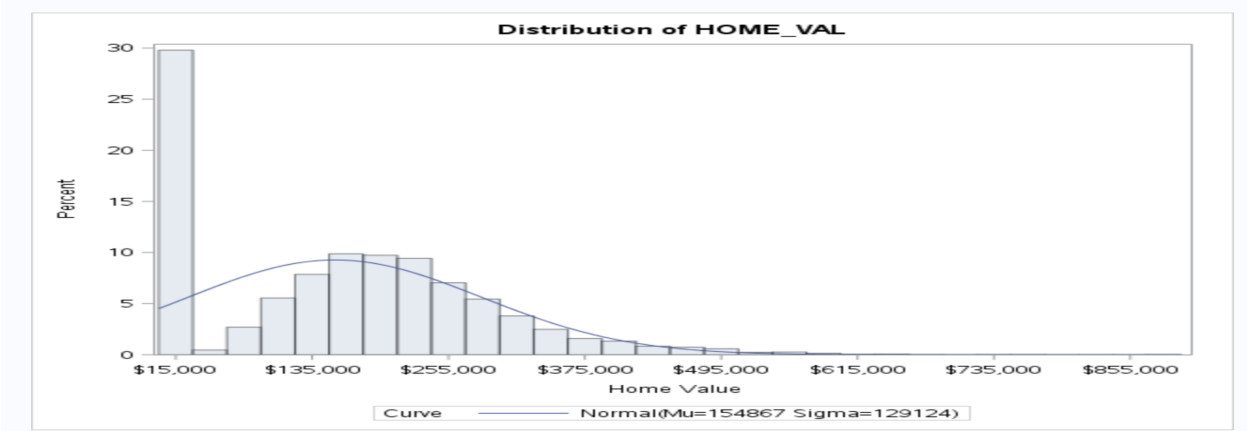


Fig:5 distributions of HOMEVAL

By observing all the distributions via a histogram as well as test statistics for normality that includes a series of goodness of fit test based on the empirical distribution function, we found that despite the way the variables appear on the histogram , our good-ness fit test indicated that we should not reject the null hypothesis, which means that our variables are normally distributed, but there are few very extreme values and asymmetrical distributions in few of the variables, these issues can be addressed by various techniques like deleting the extreme values, use bucketing, transformation.

2. Data Preparation:

In the above data exploration section, we identified few of the variables which had strong presence in left side of the histogram car_age, home_val and old_claim and also there were few variables which have missing values, thus needs to be transformed and imputed respectively, before we start modelling.

Missing values:

On reviewing the missing value chart, we can see that below variables have missing values:
CAR_AGE
HOME_VAL
INCOME
YOJ

Just to be on the safer side and we may use the other variables in our modelling later so we have imputed all the variables which have missing values with their mean.

Below new variables have been created:
IMP_AGE
I_IMP_AGE
IMP_CAR_AGE
IMP_HOME_VAL
I_IMP_HOME_VAL
IMP_INCOME
I_IMP_INCOME

IMP_YOJ
I_IMP_YOJ

Let's see the distribution of the imputed variables now:

IMP_HOME_VAL

As we can observe from the below graph that the mean is around 154867.29 but there is long tail for this graph and the skewness is about 0. 0.50329721 which is good.

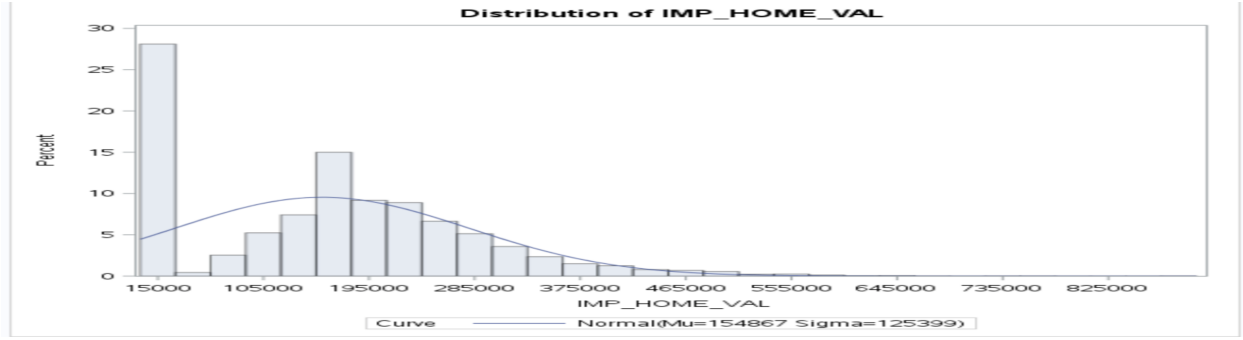


Fig:6 distributions of IMP_HOMEVAL

IMP_CAR_AGE

As we can observe from the below graph that the mean is around 8.3283231 but there is long tail for this graph and the skewness is about 0.29130939 which is good.

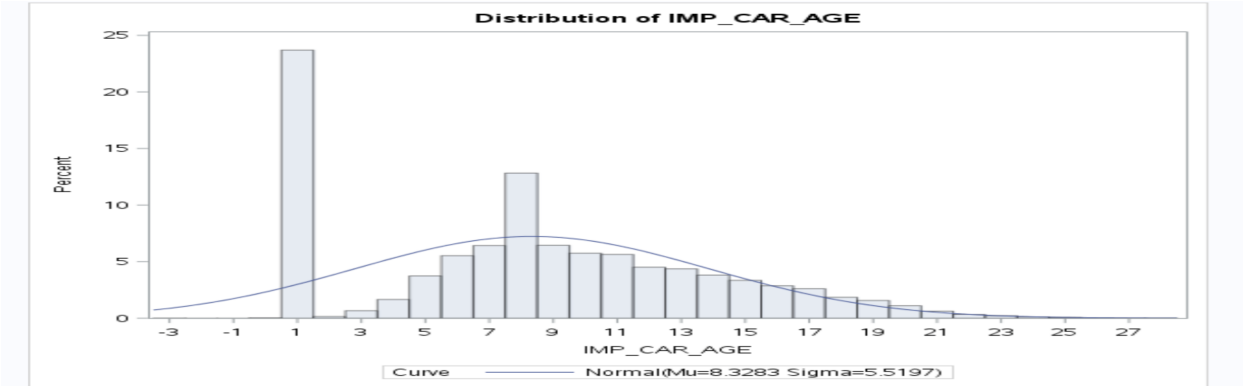


Fig:7 distributions of IMP_CAR_AGE

Below is the correlation table after the imputations:

Variable	Correlation
imp_age	-10313
bluebook	-0.10338
imp_car_age	-0.09734
imp_home_val	-0.17848
imp_income	-0.13824
oldclaim	0.13808
tif	-0.08237
travtime	0.04815
imp_yoj	-0.06849

From the above table we can see that we would take few of the variables which are exceeding 0.10, here we can see that the travel time is such low co-relation, also after seeing the above correlation table we can ignore the variables travtime and yoj for now and proceed further. As we have computed the cross correlation we also noticed some potential issue, for now we will reserve the observations until we are into model construction. Now we will examine the correlation of the continuous variables to the dependent variable by creating indicator variable families for each of the continuous variables.

Categorical variables:

Now let's have a look at the categorical variables and see how we can handle these variables, below are the categorical variables:

VARIABLE NAME	TYPE	DEFINITION
CAR_TYPE	categorical	Type of Car
CAR_USE	categorical	Vehicle Use
EDUCATION	categorical	Max Education Level
JOB	categorical	Job Category
KIDSDRIV	categorical	Driving Children
MSTATUS	categorical	Marital Status
PARENT1	categorical	Single Parent
RED_CAR	categorical	A Red Car
REVOKED	categorical	License Revoked (Past 7 Years)
SEX	categorical	Gender
URBANICITY	categorical	Home/Work Area

We went through all the categorical variables and could see that we have few of the data preprocessed such as the mstatus where we have Z-* which makes its easy for us to distinguish between the married and unmarried, we have dummy coded all the categorical variables with reference to one of the values. All the Yes are converted to 1 and No to 0.

For ex: The Car type we have different values in that variables for ex:
Minivan
z_SUV
Sports Car
Van
Panel Truck
Pickup

We have dummy coded the variables by creating new variables as below with variable of reference as Panel truck:

TYPE_MINI
TYPE_PICK
TYPE_SPOR
TYPE_VAN
TYPE_SUV

Below is the sample data after dummy coding of the categorical variables:
Similarly, all other categorical variables have been dummy coded.

TYPE_MINI	TYPE_PICK	TYPE_SPOR	TYPE_VAN	TYPE_SUV
1	0	0	0	0
1	0	0	0	0
0	0	0	0	1
1	0	0	0	0
0	0	0	0	1
0	0	1	0	0
0	0	0	0	1
0	0	0	1	0
0	0	0	0	1
0	0	0	1	0
0	0	1	0	0
1	0	0	0	0

BUILD MODELS:

We will be using logistics regression for our modelling and fisher default scoring method, we picked up of the top two models with a single variable through ten variables.

Number of variables	Chi-Square	Variables Included in Model
1	360.5259	MVR_PTS
1	349.4074	CLM_FREQ
2	508.7027	CLM_FREQ MVR_PTS
2	504.413	MVR_PTS REV_L
3	641.5337	CLM_FREQ MVR_PTS REV_L
3	639.9598	MVR_PTS USE_P REV_L
4	775.7635	IMP_INCOME MVR_PTS USE_P REV_L
4	763.1352	CLM_FREQ IMP_INCOME USE_P REV_L
5	888.9681	CLM_FREQ IMP_INCOME MVR_PTS USE_P REV_L
5	880.2277	IMP_INCOME MVR_PTS USE_P MARRIED_Y REV_L
6	983.9883	CLM_FREQ IMP_INCOME MVR_PTS USE_P MARRIED_Y REV_L
6	968.6079	CLM_FREQ IMP_INCOME MVR_PTS I_HOMEOWN USE_P REV_L
7	1044.1525	CLM_FREQ IMP_INCOME MVR_PTS TYPE_MINI USE_P MARRIED_Y REV_L
7	1028.586	CLM_FREQ IMP_INCOME MVR_PTS I_HOMEOWN TYPE_MINI USE_P REV_L
8	1082.5405	CLM_FREQ HOMEKIDS IMP_INCOME MVR_PTS TYPE_MINI USE_P MARRIED_Y REV_L
8	1078.1204	BLUEBOOK CLM_FREQ IMP_INCOME MVR_PTS TYPE_MINI USE_P MARRIED_Y REV_L
9	1113.2768	BLUEBOOK CLM_FREQ HOMEKIDS IMP_INCOME MVR_PTS TYPE_MINI USE_P MARRIED_Y REV_L
9	1106.0143	CLM_FREQ HOMEKIDS IMP_INCOME MVR_PTS TYPE_MINI USE_P JOB_M MARRIED_Y REV_L
10	1134.2002	BLUEBOOK CLM_FREQ HOMEKIDS IMP_INCOME MVR_PTS TYPE_MINI USE_P JOB_M MARRIED_Y REV_
10	1130.3332	BLUEBOOK CLM_FREQ HOMEKIDS IMP_INCOME MVR_PTS TYPE_MINI USE_P EDU_BA MARRIED_Y REV_L

Now we will pick up the variables from range 5 to 7 and build models of size five, six and seven variables in size. We chose on 5 as we can see the chi square has increased and more variables are included.

Model 1:

Now let's begin modelling our first model using the 5 variables and interpret the results:
Below are the variables used:

CLM_FREQ
IMP_INCOME
MVR_PTS
USE_P
REV_L

TARGET_FLAG	Crashes
CLM_FREQ	claims
IMP_INCOME	imputed income
MVR_PTS	Motor vehicle record points
USE_P	vehicle use
REV_L	Licensed Revoked

The model equation is:
Y=B0+B1X+ B2X2+B3X3+B4X4+B5X5 +E

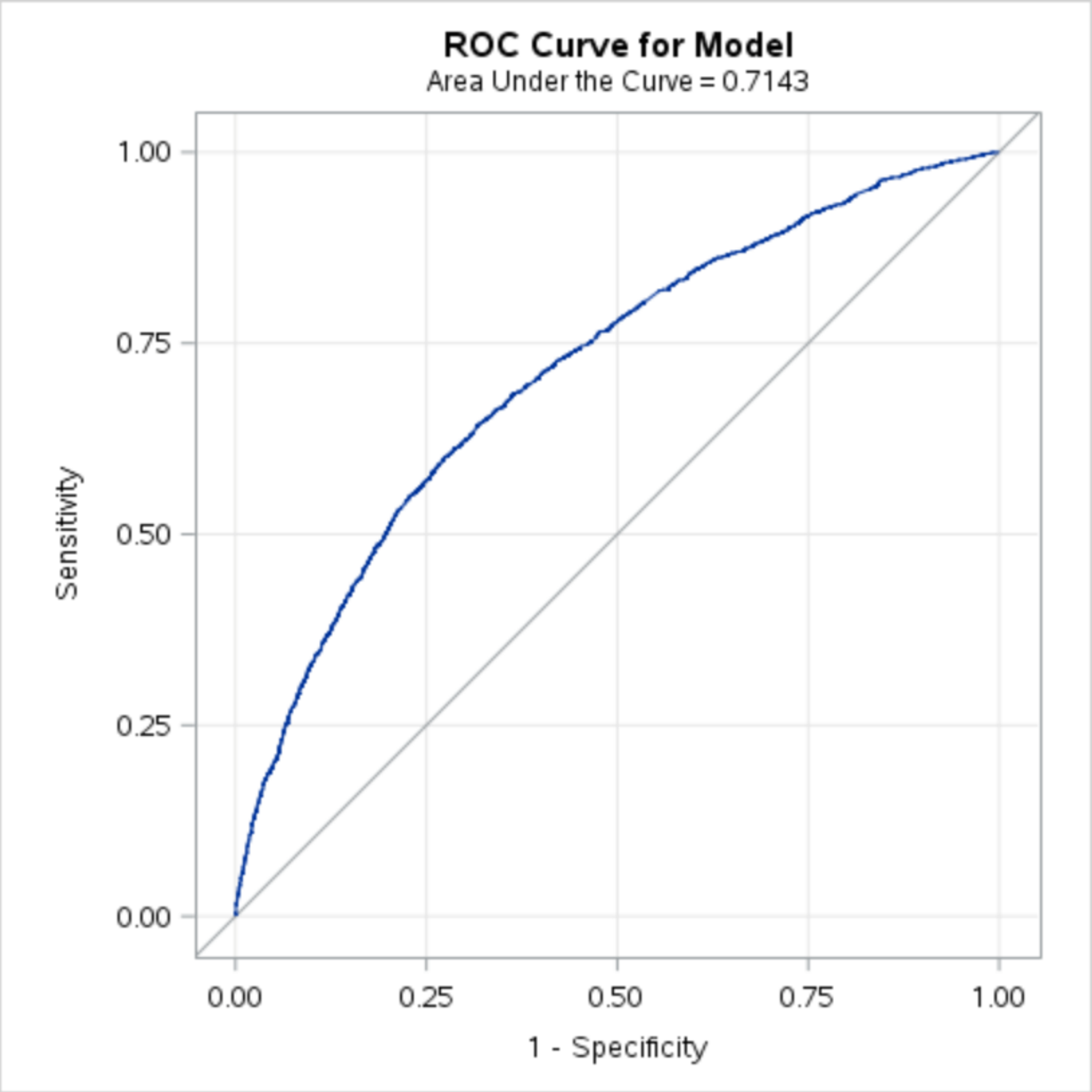
Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.8074	0.0646	156.3133	<.0001
CLM_FREQ	1	0.2709	0.0233	135.2113	<.0001
IMP_INCOME	1	-8.19E-6	6.654E-7	151.5979	<.0001
MVR_PTS	1	0.1463	0.0126	135.3884	<.0001
USE_P	1	-0.6694	0.0545	150.7663	<.0001
REV_L	1	0.8967	0.0739	147.2318	<.0001

Interpretation of the above model1, is that one-unit increase in the claim frequency the likelihood of a crash increases by 31.1 %. (e^ (0.2709-1). Similarly, for one unit increase in income the likelihood of a crash decreases by 0.0008%.for one unit in increase of the motor vehicle record points the likelihood of a crash increase by 15.8%(e ^ 0.1463-1) Use of personal vehicles results in 48 % less likelihood of a

crash as opposed to a commercial vehicle, finally if the driver has their licensed revoked in the past then there is 145 % increase in the likelihood of a crash.

Below is the roc Curve for the model1and we can see that the area under the curve is 0.7143.



Model 2:

Below are the variables used:

CLM_FREQ
IMP_INCOME
MVR_PTS
USE_P
MARRIED_Y
REV_L

TARGET_FLAG	Crashes
CLM_FREQ	claims
IMP_INCOME	imputed income
MVR_PTS	Motor vehicle record points
USE_P	vehicle use
MARRIED_Y	Marital status
REV_L	Licensed Revoked

The model equation is:

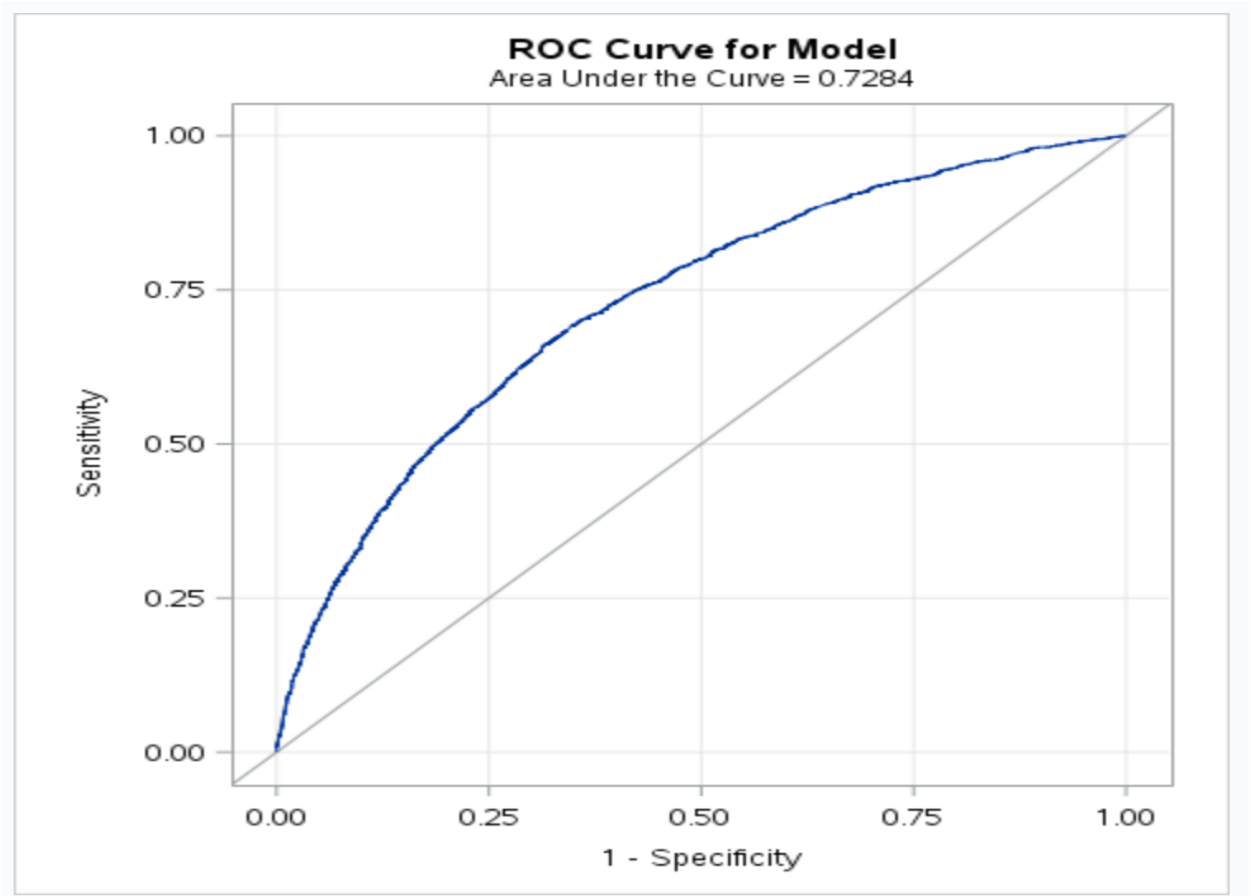
$$Y=B0+B1X+ B2X2+B3X3+B4X4+B5X5 + B6X6+E$$

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.4333	0.0727	35.5139	<.0001
CLM_FREQ	1	0.2631	0.0235	124.9811	<.0001
IMP_INCOME	1	-8.56E-6	6.705E-7	162.8418	<.0001
MVR_PTS	1	0.1449	0.0127	130.5962	<.0001
USE_P	1	-0.6733	0.0550	149.9436	<.0001
MARRIED_Y	1	-0.5995	0.0543	121.8612	<.0001
REV_L	1	0.8764	0.0746	138.1145	<.0001

Interpretation of the above model2, is that one-unit increase in the claim frequency the likelihood of a crash increases by 30.1 %. (e^ (0.2631-1). Similarly, for one-unit increase in income the likelihood of a crash decreases by 0.0008%.for one unit in increase of the motor vehicle record points the likelihood of a crash increase by 15.5%(e ^ 0.1449-1) Use of personal vehicles results in 48.9 % less likelihood of a crash as opposed to a commercial vehicle, If the person is married it results in 45 % less likelihood of a crash. Finally, if the driver has their licensed revoked in the past then there is 140 % increase in the likelihood of a crash.

As we can see from the below roc curve that adding one of the extra variable that is MARRIED_Y increase the area under the curve from 0.7143 to 0.7284



Model 3:

CLM_FREQ
HOMEKIDS
IMP_INCOME
MVR_PTS
TYPE_MINI
USE_P
MARRIED_Y
REV_L

TARGET_FLAG	Crashes
CLM_FREQ	claims
Homekids	Home kids
IMP_INCOME	imputed income
MVR_PTS	Motor vehicle record points
TYPE_MINI	Type of car minivan
USE_P	vehicle use
MARRIED_Y	Marital status
REV_L	Licensed Revoked

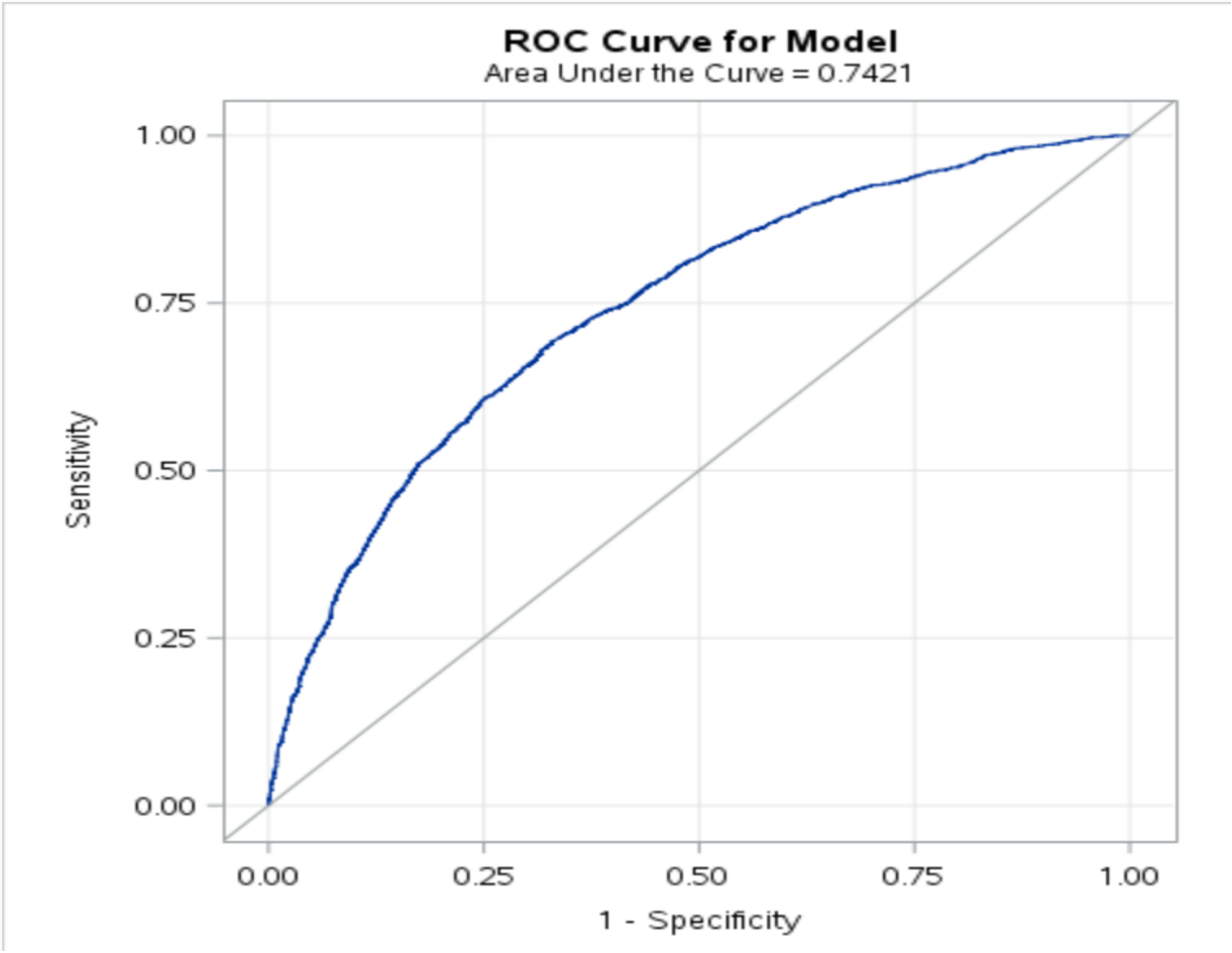
The model equation is:
Y=B0+B1X+ B2X2+B3X3+B4X4+B5X5 + B6X6+ B7X7 + B8X8 +E

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.5089	0.0768	43.9171	<.0001
CLM_FREQ	1	0.2620	0.0238	121.4630	<.0001
HOMEKIDS	1	0.1813	0.0235	59.6181	<.0001
IMP_INCOME	1	-7.64E-6	6.73E-7	128.8806	<.0001
MVR_PTS	1	0.1395	0.0128	118.9466	<.0001
TYPE_MINI	1	-0.6053	0.0700	74.8048	<.0001
USE_P	1	-0.5915	0.0562	110.6927	<.0001
MARRIED_Y	1	-0.6336	0.0550	132.6777	<.0001
REV_L	1	0.8521	0.0753	127.9962	<.0001

Interpretation of the above model3, is that one-unit increase in the claim frequency the likelihood of a crash increases by 29.9 %. (e^ (0.2620-1). If the no of home kids increase by 1 than the likelihood of crash is 19%,Now for Similarly, for one-unit increase in income the likelihood of a crash decreases by 0.00076%.for one unit in increase of the motor vehicle record points the likelihood of a crash increase by 14.9%(e ^ 0.1395-1), use of minicar results in 45 % less likelihood of a crash, Use of personal vehicles results in 44.65 % less likelihood of a crash as opposed to a commercial vehicle, If the person is married it results in 46% less likelihood of a crash. Finally, if the driver has their licensed revoked in the past then there is 134% increase in the likelihood of a crash.

As we can see from the below roc curve that adding two of the extra variable that is home kids typemini and increase the area under the curve from 0.7284 to 0.7421



MODEL COMPARISON

We have three models which we have built, Model 1, Model 2 and Model 3. For model 1 and 2 we selected the variables 5 and 6 respectively based on the chi square values.

Below is the roc coverage value for each of the model

	Model1	Model2	Model3
R-square	0.7143	0.7284	0.7421

By comparing the models, we can see that the ROC value increase by increase in the independent variable, for model 3 we see that we have the highest ROC coverage and also the interpretation of the results sounds realistic. We will go ahead with the model 3 for deployment.

Below is the equation of the model3:

The equation of the model selected is below:

P_TARGET_FLAG = -0.5089 +
+ 0.2620 * CLM_FREQ
+ 0.1813 * HOMEKIDS
-0.000000764 * IMP_INCOME
+ 0.1395 * MVR_PTS
-0.6053 * TYPEMINI
-0.5915 * USE_P
- 0.6336 * MARRIED_Y
+ 0.8521 * REV_L

Model for the TARGET_AMOUNT

Below are the variables used for this modelling:
BLUEBOOK
IMP_CAR_AGE
CLM_FREQ
OLDCLAIM
IMP_INCOME;

I did not spend much time on getting an efficient model for this target amount
We will be using the below model for deploying as we can see the r square and the r adjusted values are not that great but we will stick to this model:

Root MSE	4661.66531	**R-Square**	0.0185
Dependent Mean	1504.32465	**Adj R-Sq**	0.0179
Coeff Var	309.88426		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	1483.32938	132.12356	11.23	<.0001
BLUEBOOK	Value of Vehicle	1	0.01480	0.00675	2.19	0.0284
IMP_CAR_AGE		1	-37.24954	10.16102	-3.67	0.0002
CLM_FREQ	#Claims(Past 5 Years)	1	433.80442	51.30125	8.46	<.0001
OLDCLAIM	Total Claims(Past 5 Years)	1	0.00873	0.00677	1.29	0.1973
IMP_INCOME		1	-0.00457	0.00131	-3.48	0.0005

The model equation is:

TARGET_AMT= 1483.329 + 0.01480 * BLUEBOOK – 37.24954 * IMP_CAR_AGE + 433.80 * CLM_FREQ - IMP_INCOME * 0.00457

MODEL DEPLOYMENT CODE:

```
*///DEPLOYMENT///*;

libname mydata "/sscc/home/n/ngg135/assignment2/" access=readonly;

proc datasets library=mydata;
run;
 quit;

data testing;
set mydata.logit_insurance_test;
proc contents data=testing;
run;

data testing_new;
    set testing;
    IMP_CAR_AGE = CAR_AGE;
    I_IMP_CAR_AGE = 0;
    if missing(IMP_CAR_AGE) then do;
        IMP_CAR_AGE = 8.3283231;
        I_IMP_CAR_AGE = 1;
    end;


    IMP_INCOME = INCOME;
    I_IMP_INCOME = 0;
    if missing(IMP_INCOME) then do;
        IMP_INCOME = 61898.10;
        I_IMP_INCOME = 1;
    end;

    if CAR_USE in ('Commercial' 'Private') then do;
        USE_P = (car_use eq 'Private');
    end;
```

```
        if CAR_TYPE in ('Minivan' 'Panel Truck' 'Pickup' 'Sports Car' 'Van'
'z_SUV') then do;
        TYPE_MINI = (CAR_TYPE eq 'Minivan');
        TYPE_PICK = (CAR_TYPE eq 'Pickup');
        TYPE_SPOR = (CAR_TYPE eq 'Sports Car');
        TYPE_VAN = (CAR_TYPE eq 'Van');
        TYPE_SUV = (CAR_TYPE eq 'z_SUV');
    end;
    if MSTATUS in ('Yes' 'z_No') then do;
        MARRIED_Y = (MSTATUS eq 'Yes');
    end;
    if REVOKED in ('No' 'Yes') then do;
        REV_L = (REVOKED eq 'Yes');
    end;

data testing_score;
    set testing_new;

    Target_flag = -0.5089 + 0.2620 * CLM_FREQ + 0.1813 * HOMEKIDS
- 0.00000764 * IMP_INCOME + 0.1395 * MVR_PTS - 0.6053 *
TYPE_MINI- 0.5915 * USE_P - 0.6336 * MARRIED_Y + REV_L *0.8521 ;
/*  pi = exp(Target_flag) / (1+exp(Target_flag)); */
/*  P_TARGET_FLAG = (pi gt 0.50); */
    P_TARGET_FLAG = exp(Target_flag) / (1+exp(Target_flag));
    P_TARGET_AMT = 1483.329 + 0.01480 * BLUEBOOK - 37.24954 *
IMP_CAR_AGE + 433.80 * CLM_FREQ - IMP_INCOME * 0.00457;
    keep index P_TARGET_FLAG P_TARGET_AMT

 data home.final_prediction_score;
 set testing_score;
 run;
```

SCORED DATA FILE:

Scored data file is attached with the name final_prediction_score sas7bdat.
This file will have three columns one is the index and other are p_target_flag and p_target_amt

Conclusion:

We developed several models for this project using the data of the auto insurance, we chose the variables by using the automated selected variables created many logistic regression models and out of which 3 models have been reported in this report, we also observed that the dummy variables were selected by the automated selection process. We also built the ols model for the target amount. But overall this was a good project, where we were able to build models/deploy and make the model re-usable so that people can re-use our model. Overall we can say that the logistic model is more interpretable when compared to the OLS methods.

SAS CODE:

```
libname mydata "/sscc/home/n/ngg135/assignment2/" access=readonly;

proc datasets library=mydata;
run;
 quit;


data training;
set mydata.logit_insurance;
proc contents data=training;
run;

*///Exploratory data analysis///;

proc print data=training (obs=10);
run;

proc corr data=training;
with target_FLAG;
run;

proc means data=training ;
run;

proc means data=training NMISS N;
run;

proc corr data=training;
with target_FLAG;
run;

proc univariate data=training;
histogram INCOME /normal;
run;

proc univariate data=training;
histogram MVR_PTS /normal;
```

```
run;

proc univariate data=training;
histogram HOMEKIDS /normal;
run;


proc univariate data=training;
histogram OLDCLAIM /normal;
run;




proc univariate data=training;
histogram HOME_VAL /normal;
run;



*///IMPUTATION////*;

data imp_training;
  set training;

   IMP_AGE = AGE;
   I_IMP_AGE = 0;
   if missing(IMP_AGE) then do;
      IMP_AGE = 44.7903127;
      I_IMP_AGE = 1;
   end;

   IMP_CAR_AGE = CAR_AGE;
   I_IMP_CAR_AGE = 0;
   if missing(IMP_CAR_AGE) then do;
      IMP_CAR_AGE = 8.3283231;
      I_IMP_CAR_AGE = 1;
   end;
```

```
   IMP_HOME_VAL = HOME_VAL;
   I_IMP_HOME_VAL = 0;
   if missing(IMP_HOME_VAL) then do;
      IMP_HOME_VAL = 154867.29;
      I_IMP_HOME_VAL = 1;
   end;

   IMP_INCOME = INCOME;
   I_IMP_INCOME = 0;
   if missing(IMP_INCOME) then do;
      IMP_INCOME = 61898.10;
      I_IMP_INCOME = 1;
   end;
   log_IMP_INCOME = log(IMP_INCOME);
   t_IMP_INCOME = IMP_INCOME / 10000;

   IMP_YOJ = YOJ;
   I_IMP_YOJ = 0;
   if missing(IMP_YOJ) then do;
      IMP_YOJ = 10.4992864;
      I_IMP_YOJ = 1;
   end;

   if IMP_HOME_VAL = 0 then I_HOMEOWN = 0;
      else I_HOMEOWN = 1;

proc means data=imp_training NMISS N;
run;



proc univariate data=IMP_TRAINING;
histogram IMP_HOME_VAL /normal;
run;

proc univariate data=IMP_TRAINING;
histogram IMP_CAR_AGE /normal;
run;
```

```
proc corr data=imp_training;
with TARGET_FLAG;
  var TARGET_FLAG IMP_AGE BLUEBOOK IMP_CAR_AGE CLM_FREQ
HOMEKIDS IMP_HOME_VAL IMP_INCOME MVR_PTS OLDCLAIM TIF
TRAVTIME IMP_YOJ;


*///catergorical variables////*;

/* Categorical */

proc freq data=imp_training;
   tables CAR_TYPE CAR_USE EDUCATION JOB KIDSDRIV MSTATUS
PARENT1 RED_CAR REVOKED SEX URBANICITY;

proc freq data=imp_training;
table TARGET_FLAG*CAR_TYPE;
proc freq data=imp_training;
   table TARGET_FLAG*CAR_USE;
proc freq data=imp_training;
   table TARGET_FLAG*CLM_FREQ;
proc freq data=imp_training;
   table TARGET_FLAG*EDUCATION;
proc freq data=imp_training;
  table TARGET_FLAG*HOMEKIDS;
proc freq data=imp_training;
   table TARGET_FLAG*JOB;
proc freq data=imp_training;
   table TARGET_FLAG*KIDSDRIV;
proc freq data=imp_training;
   table TARGET_FLAG*MSTATUS;
proc freq data=imp_training;
   table TARGET_FLAG*MVR_PTS;
proc freq data=imp_training;
   table TARGET_FLAG*PARENT1;
proc freq data=imp_training;
   table TARGET_FLAG*RED_CAR;
proc freq data=imp_training;
   table TARGET_FLAG*REVOKED;
```

```
proc freq data=imp_training;
   table TARGET_FLAG*SEX;
proc freq data=imp_training;
   table TARGET_FLAG*URBANICITY;

data imp_ref_training;
   set imp_training;

   * Variable of reference: Panel Truck;
   if CAR_TYPE in ('Minivan' 'Panel Truck' 'Pickup' 'Sports Car' 'Van' 'z_SUV')
then do;
      TYPE_MINI = (CAR_TYPE eq 'Minivan');
      TYPE_PICK = (CAR_TYPE eq 'Pickup');
      TYPE_SPOR = (CAR_TYPE eq 'Sports Car');
      TYPE_VAN = (CAR_TYPE eq 'Van');
      TYPE_SUV = (CAR_TYPE eq 'z_SUV');
   end;

   * Variable of reference: Commercial;
   if CAR_USE in ('Commercial' 'Private') then do;
      USE_P = (car_use eq 'Private');
   end;

   * Variable of reference: PhD;
   if EDUCATION in ('<High School' 'Bachelors' 'Masters' 'PhD' 'z_High School')
then do;
      EDU_HS = (EDUCATION eq '<High School');
      EDU_BA = (EDUCATION eq 'Bachelors');
      EDU_MA = (EDUCATION eq 'Masters');
      EDU_ZHS = (EDUCATION eq 'z_High School');
   end;

   * Variable of reference: Doctor;
   if JOB in ('Clerical' 'Home Maker' 'Lawyer' 'Manager' 'Professional' 'Student'
'z_Blue Collar') then do;
      JOB_C = (JOB eq 'Clerical');
      JOB_HM = (JOB eq 'Home Maker');
      JOB_L = (JOB eq 'Lawyer');
      JOB_M = (JOB eq 'Manager');
      JOB_P = (JOB eq 'Professional');
```

```
      JOB_S = (JOB eq 'Student');
      JOB_BC = (JOB eq 'z_Blue Collar');
   end;

   if MSTATUS in ('Yes' 'z_No') then do;
      MARRIED_Y = (MSTATUS eq 'Yes');
   end;

   if PARENT1 in ('No' 'Yes') then do;
      PARTENT_S = (PARENT1 eq 'YES');
   end;

   if RED_CAR in ('no' 'yes') then do;
      RED_C = (RED_CAR eq 'yes');
   end;

   if REVOKED in ('No' 'Yes') then do;
      REV_L = (REVOKED eq 'Yes');
   end;

   * Variable of reference: Male;
   if SEX in ('M' 'z_F') then do;
      SEX_F = (SEX eq 'z_F');
  end;


proc print data=imp_ref_training (obs=100);
run;


proc means data=imp_ref_training NMISS N;
run;




*///MODELS////*;
```

```
proc logistic data=imp_ref_training descending plots(only)=roc(id=prob);
   model TARGET_FLAG = IMP_AGE BLUEBOOK IMP_CAR_AGE
CLM_FREQ HOMEKIDS IMP_INCOME MVR_PTS OLDCLAIM TRAVTIME
                  I_HOMEOWN
                  TYPE_MINI TYPE_PICK TYPE_SPOR TYPE_VAN TYPE_SUV
USE_P EDU_HS EDU_BA EDU_MA EDU_ZHS
                  JOB_C JOB_HM JOB_L JOB_M JOB_P JOB_S JOB_BC
MARRIED_Y PARTENT_S RED_C REV_L SEX_F /
                  selection=score outroc=roc_model rsq lackfit;
   output out=model_data pred=yhat_model;



*///MODELS-1 ////*;

proc logistic data=imp_ref_training DESCENDING PLOTS=EFFECT
PLOTS=ROC;
   model TARGET_FLAG = CLM_FREQ IMP_INCOME MVR_PTS USE_P
REV_L / rsq lackfit;
   output out=model_5_data pred=model_5_yhat;



proc logistic data=imp_ref_training DESCENDING PLOTS=EFFECT
PLOTS=ROC;
   model TARGET_FLAG =CLM_FREQ IMP_INCOME MVR_PTS USE_P
MARRIED_Y REV_L / rsq lackfit;
   output out=model_6_data pred=model_6_yhat;


proc logistic data=imp_ref_training DESCENDING PLOTS=EFFECT
PLOTS=ROC;
   model TARGET_FLAG =CLM_FREQ HOMEKIDS IMP_INCOME
MVR_PTS TYPE_MINI USE_P MARRIED_Y REV_L / rsq lackfit;
   output out=model_7_data pred=model_7_yhat;


proc logistic data=imp_ref_training DESCENDING PLOTS=EFFECT
PLOTS=ROC;
```

```
   model TARGET_FLAG =CLM_FREQ HOMEKIDS IMP_INCOME
MVR_PTS TYPE_MINI USE_P MARRIED_Y REV_L / rsq lackfit;
   output out=model_8_data pred=model_8_yhat;
```

```
/* Modeling TARGET_AMT */


proc reg data=imp_ref_training;
   model TARGET_AMT = BLUEBOOK IMP_HOME_VAL OLDCLAIM
CAR_AGE CLM_FREQ HOMEKIDS IMP_INCOME MVR_PTS MARRIED_Y/
   selection=adjrsq aic bic cp best=5;

proc reg data=imp_ref_training;
   model TARGET_AMT = BLUEBOOK IMP_CAR_AGE CLM_FREQ
OLDCLAIM IMP_INCOME;
  run;
```

```
*///DEPLOYMENT///*;

libname mydata "/sscc/home/n/ngg135/assignment2/" access=readonly;

proc datasets library=mydata;
run;
 quit;


data testing;
set mydata.logit_insurance_test;
```

```
proc contents data=testing;
run;


data testing_new;
   set testing;
   IMP_CAR_AGE = CAR_AGE;
   I_IMP_CAR_AGE = 0;
   if missing(IMP_CAR_AGE) then do;
      IMP_CAR_AGE = 8.3283231;
      I_IMP_CAR_AGE = 1;
   end;


   IMP_INCOME = INCOME;
   I_IMP_INCOME = 0;
   if missing(IMP_INCOME) then do;
      IMP_INCOME = 61898.10;
      I_IMP_INCOME = 1;
   end;

   if CAR_USE in ('Commercial' 'Private') then do;
      USE_P = (car_use eq 'Private');
   end;
      if CAR_TYPE in ('Minivan' 'Panel Truck' 'Pickup' 'Sports Car' 'Van' 'z_SUV')
then do;
      TYPE_MINI = (CAR_TYPE eq 'Minivan');
      TYPE_PICK = (CAR_TYPE eq 'Pickup');
      TYPE_SPOR = (CAR_TYPE eq 'Sports Car');
      TYPE_VAN = (CAR_TYPE eq 'Van');
      TYPE_SUV = (CAR_TYPE eq 'z_SUV');
   end;
   if MSTATUS in ('Yes' 'z_No') then do;
      MARRIED_Y = (MSTATUS eq 'Yes');
   end;
   if REVOKED in ('No' 'Yes') then do;
      REV_L = (REVOKED eq 'Yes');
   end;

data testing_score;
```

```
   set testing_new;

   Target_flag = -0.5089 + 0.2620 * CLM_FREQ + 0.1813 * HOMEKIDS -
0.00000764 * IMP_INCOME + 0.1395 * MVR_PTS - 0.6053 * TYPE_MINI-
0.5915 * USE_P - 0.6336 * MARRIED_Y + REV_L *0.8521 ;
/*  pi = exp(Target_flag) / (1+exp(Target_flag)); */
/*  P_TARGET_FLAG = (pi gt 0.50); */
   P_TARGET_FLAG = exp(Target_flag) / (1+exp(Target_flag));
   P_TARGET_AMT = 1483.329 + 0.01480 * BLUEBOOK - 37.24954 *
IMP_CAR_AGE + 433.80 * CLM_FREQ - IMP_INCOME * 0.00457;
   keep index P_TARGET_FLAG P_TARGET_AMT

 data home.final_prediction_score;
 set testing_score;
 run;
```