

---

# Insurance Customer Risk Analysis Report

---

Project #2

Singh, Gurjeet

This report contains analysis done while building logistic regression models to determine if the person is a safer or riskier driver.

# Insurance Customer Risk Analysis Report

---

## Table of Contents

Introduction: .....	3
Section 1: Data Exploration.....	4
Section 1.1: Statistics .....	4
Section 1.1.1: Numerical Variables .....	4
Section 1.1.2: Categorical Variables.....	5
Section 1.2: Examining Distributions .....	6
Section 2: Data Preparation .....	13
Section 3: Model Building .....	15
Section 3.1: Model 1 .....	15
Section 3.2: Model 2 .....	18
Section 3.3: Model 3 .....	21
Section 4: Model Comparison and Selection .....	24
Section 5: Model Testing and Scoring.....	26
Conclusion:.....	27
Appendix I: Model Development R Code.....	28
Appendix II: Stand-Alone R Code .....	45

# Insurance Customer Risk Analysis Report

---

## Table of Figures

Figure 1: Table - Data Dictionary .....	3
Figure 2: Table - Statistical Values of Numerical Variables.....	4
Figure 3: Table - Statistical Values of Categorical Variables .....	5
Figure 4: Histogram: Review Distributions .....	6
Figure 5: Histogram: Review Distributions .....	7
Figure 6: Histogram: Review Distributions .....	8
Figure 7: Histogram: Review Distributions .....	9
Figure 8: Histogram: Review Distributions .....	10
Figure 9: Histogram: Review Distributions .....	11
Figure 10: Table – Variable Selection List .....	12
Figure 11: Table – Imputed values for INCOME JOB and other variables.....	13
Figure 12: Table – Indicator Variables .....	13
Figure 13: Table – Statistical values with Indicator variables .....	14
Figure 14: Table – Dropped Variables.....	14
Figure 15: Model 1: Output .....	15
Figure 16: Model 1: Odds Ratio .....	16
Figure 17: Model 1: ROC Curves .....	17
Figure 18: Model 1: Statistics.....	17
Figure 19: Model 2: Output .....	18
Figure 20: Model 2: Odds Ratio .....	19
Figure 21: Model 2: ROC Curves .....	20
Figure 22: Model 2: Statistics.....	20
Figure 23: Model 3: Output .....	21
Figure 24: Model 3: Odds Ratio .....	22
Figure 25: Model 3: ROC Curves .....	23
Figure 26: Model 2: Statistics.....	23
Figure 27: Model Selection: Model 1, Model 2, and Model 3 .....	24
Figure 27: Model Selection: Statistics.....	24
Figure 29: Models: ROC Curves.....	25
Figure 20: Auto Insurance Test Data Result Stats.....	26

# Insurance Customer Risk Analysis Report

## Introduction:

The purpose of this project is to build Logistic Regression models using the data from an auto insurance company to predict if the customer is a safe or risky driver.

For our purposes, we use the data set that contains customer information of the auto insurance company. Each record in the data represents a customer. Each record has two target variables. The first target variable, TARGET\_FLAG, is a “1” or a “0”. A “1” means that the person was in a car crash. A “0” (zero) means that the person was not in a car crash. The second target variable is TARGET\_AMT. This contains the value of zero (0) if the person did not crash their car. However, if they did crash their car, the number in this variable will be a value greater than zero (0).

The training dataset contains 8161 observations and 26 explanatory variables of which 13 numerical variables, 10 categorical variables, 3 target variables, and 1 index variable. The test dataset contains 2141 observations and 26 explanatory variables. In the test dataset, there are no “TARGET\_FLAG” and “TARGET\_AMT” values. We will be using our selected model to score the test data file to predict the probability of the risk factor of a customer (“TARGET\_FLAG”) and calculate the amount of payout (“TARGET\_AMT”) if involved in a car crash.

Figure 1 gives the basic descriptions of each field and how those affect the prediction.

Figure 1: Table - Data Dictionary

VARIABLE NAME	DEFINITION	THEORETICAL EFFECT
INDEX	Identification Variable (do not use)	None
TARGET_FLAG	Was Car in a crash? 1=YES 0=NO	None
TARGET_AMT	If car was in a crash, what was the cost	None
AGE	Age of Driver	Very young people tend to be risky. Maybe very old people also.
BLUEBOOK	Value of Vehicle	Unknown effect on probability of collision, but probably effect the payout if there is a crash
CAR_AGE	Vehicle Age	Unknown effect on probability of collision, but probably effect the payout if there is a crash
CAR_TYPE	Type of Car	Unknown effect on probability of collision, but probably effect the payout if there is a crash
CAR_USE	Vehicle Use	Commercial vehicles are driven more, so might increase probability of collision
CLM_FREQ	#Claims(Past 5 Years)	The more claims you filed in the past, the more you are likely to file in the future
EDUCATION	Max Education Level	Unknown effect, but in theory more educated people tend to drive more safely
HOMEKIDS	#Children @Home	Unknown effect
HOME_VAL	Home Value	In theory, home owners tend to drive more responsibly
INCOME	Income	In theory, rich people tend to get into fewer crashes
JOB	Job Category	In theory, white collar jobs tend to be safer
KIDSDRIV	#Driving Children	When teenagers drive your car, you are more likely to get into crashes
MSTATUS	Marital Status	In theory, married people drive more safely
MVR_PTS	Motor Vehicle Record Points	If you get lots of traffic tickets, you tend to get into more crashes
OLDCLAIM	Total Claims(Past 5 Years)	If your total payout over the past five years was high, this suggests future payouts will be high
PARENT1	Single Parent	Unknown effect
RED_CAR	A Red Car	Urban legend says that red cars (especially red sports cars) are more risky. Is that true?
REVOKED	License Revoked (Past 7 Years)	If your license was revoked in the past 7 years, you probably are a more risky driver.
SEX	Gender	Urban legend says that women have less crashes then men. Is that true?
TIF	Time in Force	People who have been customers for a long time are usually more safe.
TRAVTIME	Distance to Work	Long drives to work usually suggest greater risk
URBANICITY	Home/Work Area	Unknown
YOJ	Years on Job	People who stay at a job for a long time are usually more safe

Now that we have some context for our analysis and dataset, let's look at the results in the next section.

# Insurance Customer Risk Analysis Report

## Section 1: Data Exploration

The first step towards any modeling project is the Data Exploration i.e. Exploratory Data Analysis (EDA). This helps us to understand and analyze the data set to summarize the main characteristics of variables. For this purposes, we will look into the basic statistics to understand the data and examine the distributions of the variables.

### Section 1.1: Statistics

#### Section 1.1.1: Numerical Variables

Figure 2 shows the statistical values of the numerical variables in the Insurance training data set. The number of observations in the data is 8161 records. We can clearly see that there are missing values (NAs) in the data for variables, AGE, YOJ, INCOME, HOME\_VAL, and CAR\_AGE. In Table 2, we can also see outliers in various variables in the data. For instance, the variable CAR\_AGE has a minimum value of “-3” which to me seems incorrect. Hence, we convert it to “0” (zero) to make it less than 1 year old. We also notice that INCOME and HOME\_VAL are \$0. We will need to further investigate to make sure those values are correct or miscoded.

We explained how we fixed the missing values and outliers in Section 2: Data Preparation.

Figure 2: Table - Statistical Values of Numerical Variables

Variable Names	nobs	NAs	Minimum	Maximum	1. Quartile	3. Quartile	Mean	Median	Stdev	Skewness	Kurtosis
INDEX	8161	0	1	10302	2559	7745	5151.86766	5133	2978.89396	0.002004	-1.203421
TARGET_FLAG	8161	0	0	1	0	1	0.263816	0	0.440728	1.071661	-0.851646
TARGET_AMT	8161	0	0	107586.136	0	1036	1504.32465	0	4704.02693	8.706303	112.288439
KIDSDRIV	8161	0	0	4	0	0	0.171057	0	0.511534	3.351837	11.780192
AGE	8161	6	16	81	39	51	44.790313	45	8.627589	-0.028989	-0.061702
HOMEKIDS	8161	0	0	5	0	1	0.721235	0	1.116323	1.341127	0.648991
YOJ	8161	454	0	23	9	13	10.499286	11	4.092474	-1.202968	1.177341
INCOME	8161	445	0	367030.262	28096.9657	85986.21189	61898.0974	54028.1694	47572.6873	1.186316	2.129017
HOME_VAL	8161	464	0	885282.345	0	238724.4489	154867.29	161159.526	129123.777	0.488595	-0.016083
TRAVTIME	8161	0	5	142.1206	22.4517	43.80707	33.488797	32.87097	15.904747	0.447774	0.664152
BLUEBOOK	8161	0	1500	69740	9280	20850	15709.8995	14440	8419.73408	0.794214	0.791356
TIF	8161	0	1	25	1	7	5.351305	4	4.146635	0.890812	0.422494
OLDCLAIM	8161	0	0	57037	0	4636	4037.07622	0	8777.1391	3.11904	9.860658
CLM_FREQ	8161	0	0	5	0	2	0.798554	0	1.158453	1.208799	0.284289
MVR_PTS	8161	0	0	13	0	3	1.695503	1	2.147112	1.34784	1.37549
CAR_AGE	8161	510	-3	28	1	12	8.328323	8	5.700742	0.281953	-0.748976

# Insurance Customer Risk Analysis Report

## Section 1.1.2: Categorical Variables

Figure 3 shows the statistical values i.e. Frequency of the categorical variables in the Insurance training data set. We notice that JOB has 526 missing values (NAs) in the data. We will fix these in Section 2: Data Preparation.

Figure 3: Table - Statistical Values of Categorical Variables

Variable Name	Value	Frequency
PARENT1	No	7084
PARENT1	Yes	1077
MSTATUS	Yes	4894
MSTATUS	z_No	3267
SEX	M	3786
SEX	z_F	4375
EDUCATION	<High School	1203
EDUCATION	Bachelors	2242
EDUCATION	Masters	1658
EDUCATION	PhD	728
EDUCATION	z_High School	2330
JOB	Clerical	1271
JOB	Doctor	246
JOB	Home Maker	641
JOB	Lawyer	835
JOB	Manager	988
JOB	Professional	1117
JOB	Student	712
JOB	z_Blue Collar	1825
JOB	NA	526
CAR_USE	Commercial	3029
CAR_USE	Private	5132
CAR_TYPE	Minivan	2145
CAR_TYPE	Panel Truck	676
CAR_TYPE	Pickup	1389
CAR_TYPE	Sports Car	907
CAR_TYPE	Van	750
CAR_TYPE	z_SUV	2294
RED_CAR	no	5783
RED_CAR	yes	2378
REVOKED	No	7161
REVOKED	Yes	1000
URBANICITY	Highly Urban/ Urban	6492
URBANICITY	z_Highly Rural/ Rural	1669

# Insurance Customer Risk Analysis Report

## Section 1.2: Examining Distributions

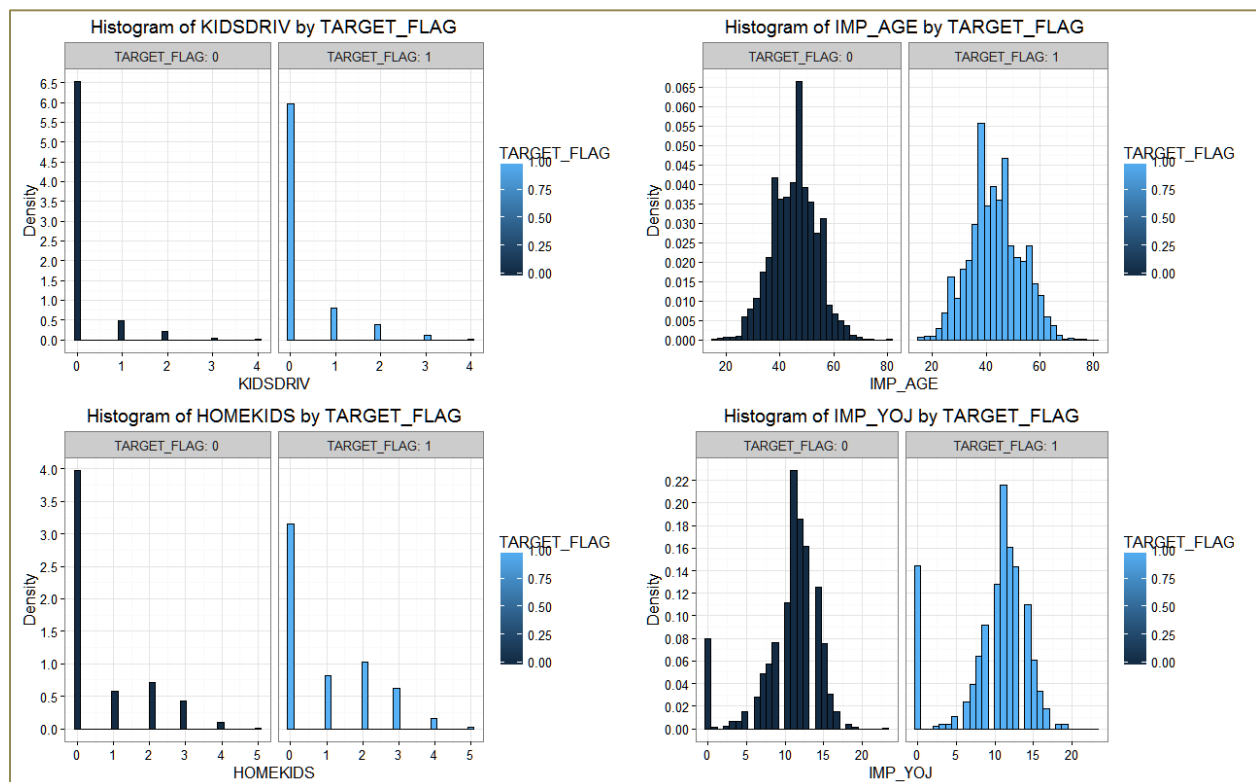
We continue our data exploration by examining the distribution after fixing the missing values to get an appropriate distribution. So, if you notice any variable name starting with “IMP\_”, it means that the variable is imputed.

Figure 4 shows the distributions of values for the variables, KIDSDRIV, IMP\_AGE, HOMEKIDS, and IMP\_YOJ, classified by TARGET\_FLAG i.e. by customers who crashed their car (TARGET\_FLAG = 1) color-coded in blue and customers who did not crash their car (TARGET\_FLAG = 0) color-coded in black.

In Figure 4, in each of the subgraphs, we notice that percentage of people would crash their car with an increase in the values of each variable. For instance, if a customer has a number of kids who drive, are more likely to crash their car. If a person who stays at a job longer are less like to crash their car.

From Figure 4, we believe that the variables, KIDSDRIV, IMP\_AGE, and HOMEKIDS, are predictable and should be included in the model. We would show the final selection of variables after reviewing all the variables.

Figure 4: Histogram: Review Distributions



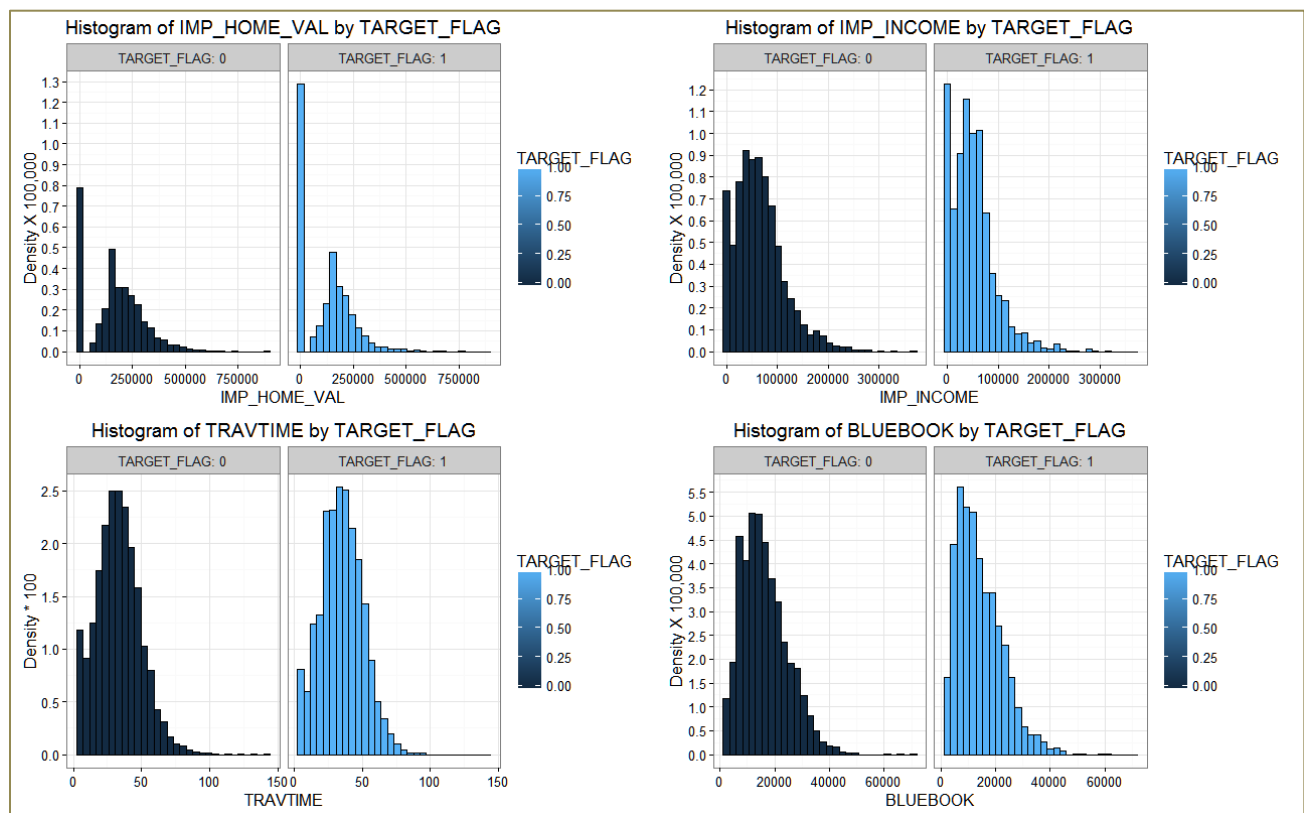
# Insurance Customer Risk Analysis Report

Figure 5 shows the distributions of values for the variables, IMP\_HOME, IMP\_INCOME, TRAVTIME, and BLUEBOOK, classified by TARGET\_FLAG i.e. by customers who crashed their car (TARGET\_FLAG = 1) color-coded in blue and customers who did not crash their car (TARGET\_FLAG = 0) color-coded in black.

In Figure 5, we notice that percentage of people who do not own a house i.e. renting or other (staying with family) are more likely to crash their car versus people who do own a house. Similarly, people with less income are likely to crash their car versus people with higher income. People with higher travel time are also considered to have more car crash. Figure 5 shows that people with lower bluebook value of the car would crash their car more. However, in order to keep a fair prediction regardless of car bluebook value, we would not include this variable in our model. To me, this variable seems to be more useful to calculate the payout if there is a crash.

Therefore, we will include these variables, IMP\_HOME\_VAL, IMP\_INCOME, and TRAVTIME in our list because these seem to be predictable.

Figure 5: Histogram: Review Distributions





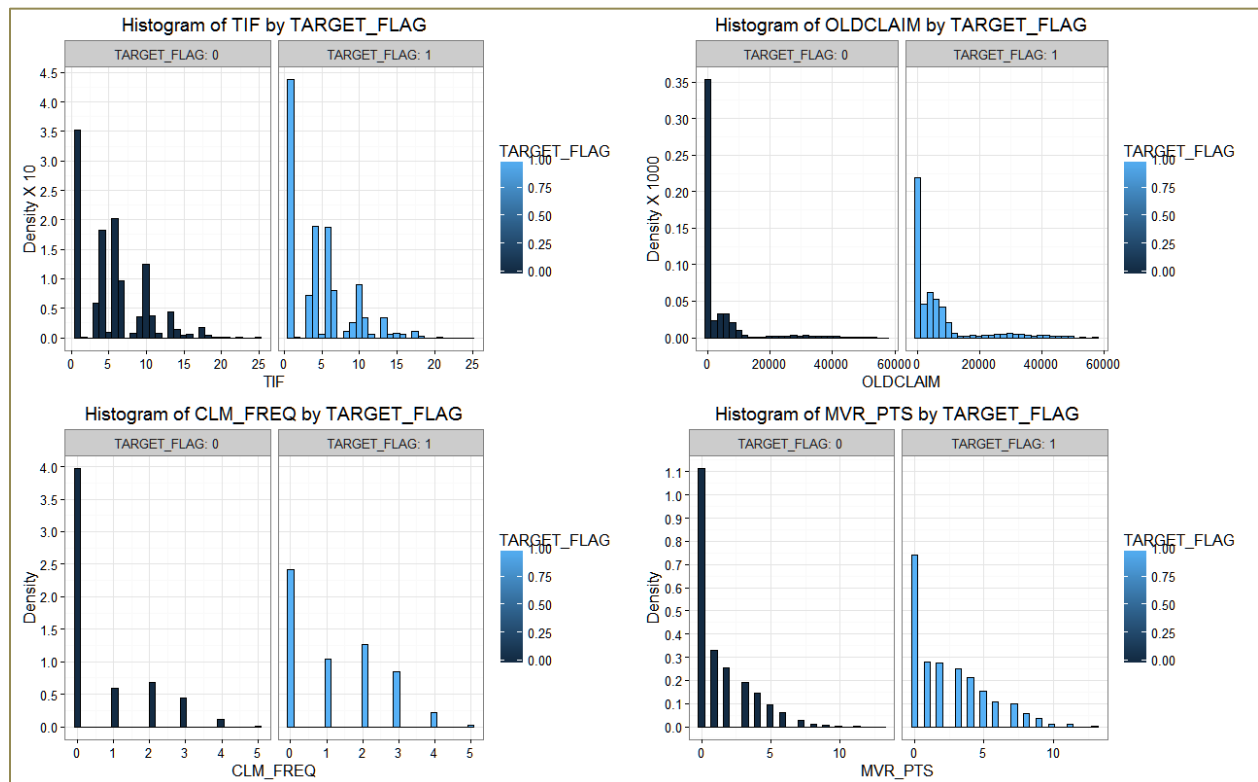
# Insurance Customer Risk Analysis Report

Figure 6 shows the distributions of values for the variables, TIF, OLDCLAIM, CLM\_FREQ, and MVR\_PTS, classified by TARGET\_FLAG i.e. by customers who crashed their car (TARGET\_FLAG = 1) color-coded in blue and customers who did not crash their car (TARGET\_FLAG = 0) color-coded in black.

In Figure 6, we notice that percentage of people who changed their insurance company pretty frequently had more car crash. Figure 6 shows that people with more claims filed and higher old claim payouts had more car crash. Lastly, the more traffic tickets a person received, the more numbers of car crash they had.

Therefore, we will include all of these variables in our list because these seem to be predictable.

Figure 6: Histogram: Review Distributions



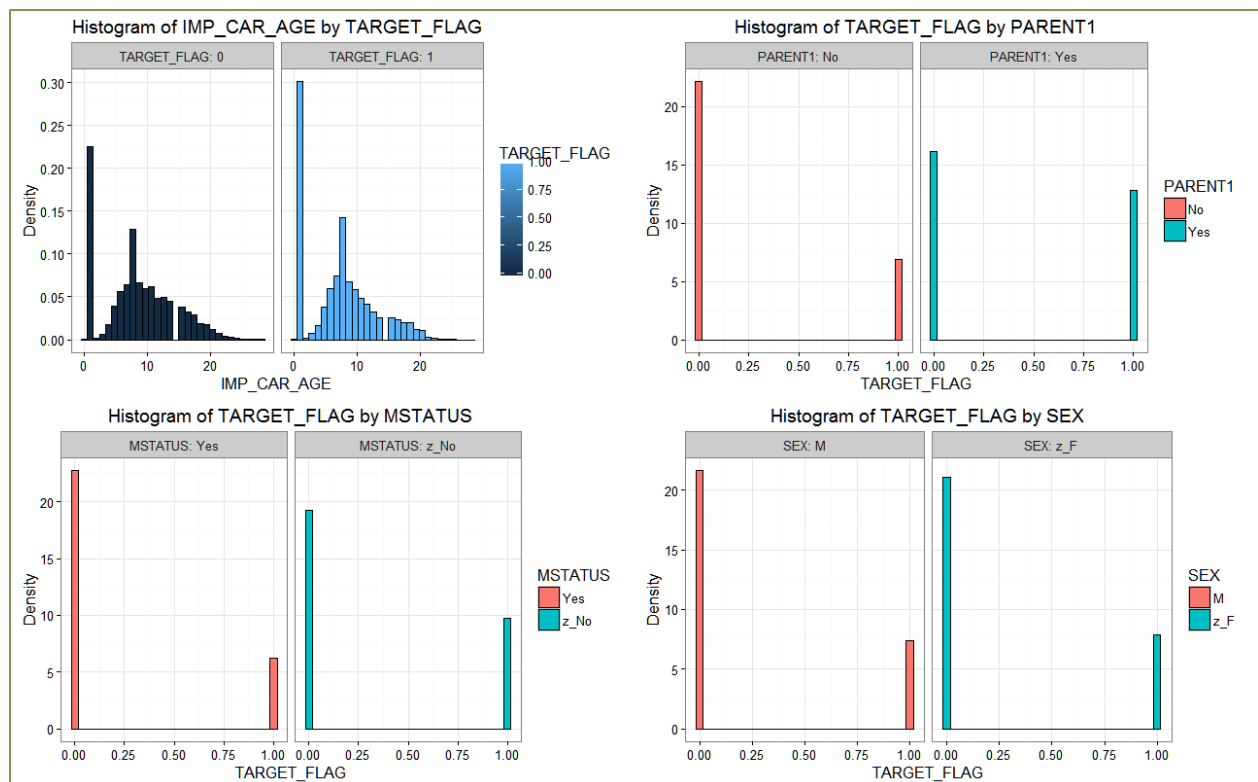
# Insurance Customer Risk Analysis Report

Figure 7 shows the distributions of values for the variable, IMP\_CAR\_AGE, classified by TARGET\_FLAG i.e. by customers who crashed their car (TARGET\_FLAG = 1) color-coded in blue and customers who did not crash their car (TARGET\_FLAG = 0) color-coded in black. Figure 7 also shows the TARGET\_FLAG distribution by categorical variables.

We notice in this figure that the people with the older car tend to have more crashes. Also, the percentage of single-parent people have met with more car crash than a non-single parent. Similarly, single people have more accidents than married people. Generally, people say that women are safer driver than men, however, our data doesn't hold true to that case. In Figure 7, you can see that females had more car crash than males.

Therefore, again we will include all of these variables in our list because these seem to be predictable.

Figure 7: Histogram: Review Distributions



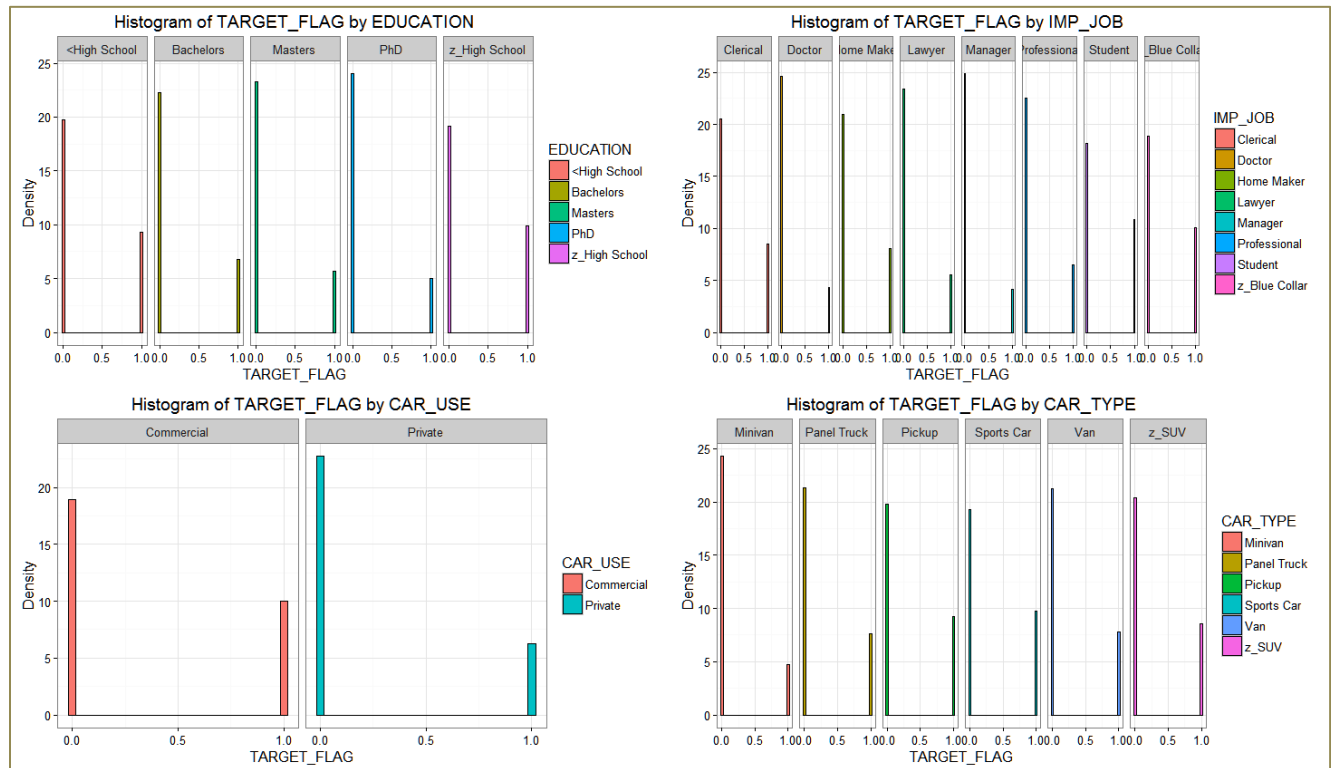
# Insurance Customer Risk Analysis Report

Figure 8 shows the TARGET\_FLAG distribution by categorical variables.

We notice that students with no degree (High School or below) have more car crash than people with a degree. Similarly, white collar people have less crash than blue collar and students. Figure 8 also confirms that people who used their car for commercial purposes had more car crash. Lastly, people with the sports car or heavy cars had more car crash.

Therefore, again we will include all of these variables in our list because these seem to be predictable.

Figure 8: Histogram: Review Distributions



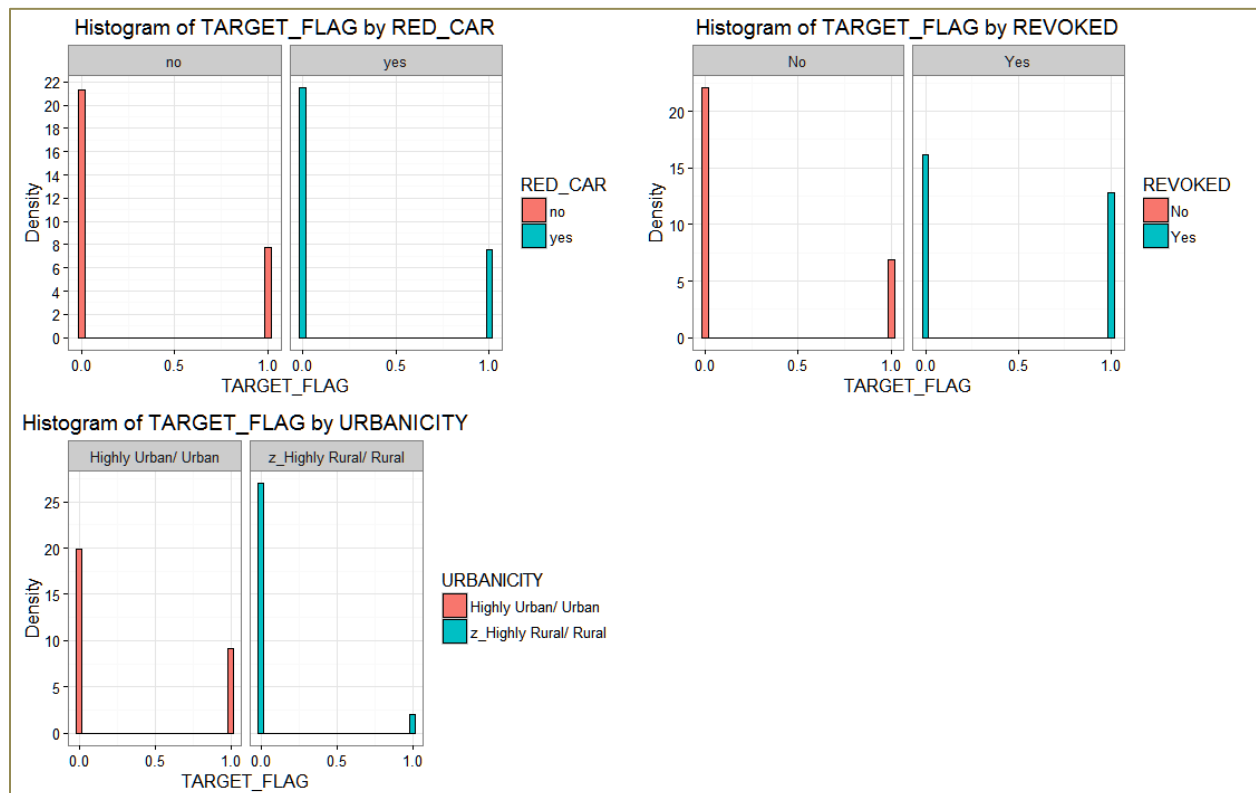
# Insurance Customer Risk Analysis Report

Figure 9 again shows the TARGET\_FLAG distribution by categorical variables.

We noticed that people who have a red car, not necessarily meet with more car crash than people who do not own a red car. Figure 9 shows us that people who had their license revoked are more likely to crash their car. Lastly, people living in the city i.e. urban area have more car crash.

Therefore, we will include the variables, REVOKED and URBANCITY in our list because these seem to be predictable and exclude RED\_CAR.

Figure 9: Histogram: Review Distributions



# Insurance Customer Risk Analysis Report

Figure 10 shows the variables we have selected based on our above distribution analyses. It also includes the analysis done using the percentage (proportion value) of each variable to determine if those are higher than the average percentage (when TARGET\_FLAG = 1, the average percentage is: **0.264**) of people who had car crash i.e. (TARGET\_FLAG = 1). In the “FINAL SELECTION” column, “YES” means to include based on the criteria and “NO” means do not include in the final population for model building. We have also included short notes or comment for each of the variables that we would create indicator variables. Apart from these variables, we also have our response variable i.e. TARGET\_FLAG.

**Figure 10: Table – Variable Selection List**

Variables	Distribution Analysis	% higher than TARGET_FLAG = 1	FINAL SELECTION	Comments
KIDSDRIV	YES	YES	YES	
IMP_AGE	YES	YES	YES	
HOMEKIDS	YES	YES	YES	
IMP_YOJ	NO	NO	NO	
IMP_HOME_VAL	YES	YES	YES	Using a indicator variable for renters versus home owners
IMP_INCOME	YES	YES	YES	
TRAVTIME	YES	YES	YES	
BLUEBOOK	YES	NO	NO	
TIF	YES	YES	YES	
OLDCLAIM	YES	YES	YES	
CLM_FREQ	YES	YES	YES	
MVR_PTS	YES	YES	YES	
IMP_CAR_AGE	YES	NO	NO	
PARENT1	YES	YES	YES	Single Parents tends to be risky - Indicator Variable
MSTATUS	YES	YES	YES	Singles tends to be risky - Indicator Variable
SEX	YES	YES	YES	Female seems to be risky than Male - Indicator Variable
EDUCATION	YES	YES	YES	High School or below are risky - Indicator Variable Degree vs Non Degree
CAR_USE	YES	YES	YES	Commercial use have high risk - Indicator Variable
CAR_TYPE	YES	YES	YES	Sports and fast cars have high risk - Indicator Variables
RED_CAR	NO	NO	NO	
REVOKED	YES	YES	YES	License revoked people are more risky - Indicator Variable
URBANICITY	YES	YES	YES	People living in Urban area are risky - Indicator Variable
IMP_JOB	Yes	YES	YES	Blue Collar job people are risky - Indicator Variables

# Insurance Customer Risk Analysis Report

## Section 2: Data Preparation

In Section 1: Data Exploration, we noticed that we had some missing values and outliers in the data. This section explains the methodology we took to fix the missing values and transform data to eliminate outliers. First, we fixed the missing values by using the Decision Tree concepts for variable INCOME and JOB. Figure 11 shows the logic and value used to impute INCOME and JOB, as well as the Median values for the rest of the variables with missing values, were fixed using their Median values.

During our data exploration, we notice that income was 0 (zero) for Doctor, lawyers, Manager, etc. After further review, we noticed that Years in the job for each of those customers were 0 (zero) as well. This could be because a customer just started the job. We also notice certain students had very high income. This again was notified because we do not know if that income is miscoded or job category is miscoded. Hence, we would move forward with the data as is.

Figure 11: Table – Imputed values for INCOME JOB and other variables

Decision Tree logic for Missing values	INCOME Value	Decision Tree logic for Missing values	JOB Value	Variable	Median Values
INCOME is Empty and JOB is Empty	\$54,000	JOB is Empty and INCOME > 150000	Doctor	AGE	45
INCOME is Empty and JOB is Doctor	\$128,000	JOB is Empty and INCOME > 100000	Lawyer	YOJ	11
INCOME is Empty and JOB is Lawyer	\$88,000	JOB is Empty and INCOME > 85000	Manager	HOME_VAL	\$162,000
INCOME is Empty and JOB is Manager	\$87,000	JOB is Empty and INCOME > 75000	Professional	CAR_AGE	8
INCOME is Empty and JOB is Professional	\$76,000	JOB is Empty and INCOME > 60000	z_Blue Collar		
INCOME is Empty and JOB is Clerical	\$33,000	JOB is Empty and INCOME > 35000	Clerical		
INCOME is Empty and JOB is z_Blue Collar	\$58,000	JOB is Empty and INCOME > 12000	Home Maker		
INCOME is Empty and JOB is Home Maker	\$12,000	JOB is Empty and INCOME < 12000	Student		
INCOME is Empty and JOB is Student	\$6,300				

Figure 12 shows all the categorical variables that were converted into indicator variables.

Figure 12: Table – Indicator Variables

Variable	Indicator Variable
PARENT1	Single_Parent_Ind
MSTATUS	MSTATUS_Single_Ind
SEX	SEX_Female_Ind
EDUCATION - High School or below	Ed_Non_Degree_Ind
CAR_USE - Commercial	CU_Commercial_Ind
CAR_TYPE - Panel Truck	CT_Panel_Truck_Ind
CAR_TYPE - Pickup	CT_Pickup_Ind
CAR_TYPE - Sports Car	CT_Sports_Car_Ind
CAR_TYPE - Van	CT_Van_Ind
CAR_TYPE - z_SUV	CT_SUV_Ind
REVOKED - Yes	REVOKED_Ind
URBANICITY - Highly Urban/ Urban	UC_HUU_Ind
IMP_JOB - All except Blue Collar, Student, and Home Maker	JOB_White_Collar_Ind
IMP_JOB - z_Blue Collar	JOB_Blue_Collar_Ind
IMP_JOB - Student	JOB_Student_Ind

# Insurance Customer Risk Analysis Report

We have also created indicator variables for the categorical variables. Figure 13 shows all the variables including the imputed variables and indicator variables. The variables that are highlighted in yellow had missing values and we imputed those by creating corresponding variables highlighted in green.

Figure 13: Table – Statistical values with Indicator variables

Variable Names	nobs	NAs	Minimum	Maximum	1. Quartile	3. Quartile	Mean	Median	Stdev	Skewness	Kurtosis
INDEX	8161	0	1	10302	2559	7745	5151.86766	5133	2978.89396	0.002004	-1.203421
TARGET_FLAG	8161	0	0	1	0	1	0.263816	0	0.440728	1.071661	-0.851646
TARGET_AMT	8161	0	0	107586.136	0	1036	1504.32465	0	4704.02693	8.706303	112.288439
KIDSDRIV	8161	0	0	4	0	0	0.171057	0	0.511534	3.351837	11.780192
AGE	8161	6	16	81	39	51	44.790313	45	8.627589	-0.028989	-0.061702
HOMEKIDS	8161	0	0	5	0	1	0.721235	0	1.116323	1.341127	0.648991
YOJ	8161	454	0	23	9	13	10.499286	11	4.092474	-1.202968	1.177341
INCOME	8161	445	0	367030.262	28096.9657	85986.21189	61898.0974	54028.1694	47572.6873	1.186316	2.129017
HOME_VAL	8161	464	0	885282.345	0	238724.4489	154867.29	161159.526	129123.777	0.488595	-0.016083
TRAVTIME	8161	0	5	142.1206	22.4517	43.80707	33.488797	32.87097	15.904747	0.447774	0.664152
BLUEBOOK	8161	0	1500	69740	9280	20850	15709.8995	14440	8419.73408	0.794214	0.791356
TIF	8161	0	1	25	1	7	5.351305	4	4.146635	0.890812	0.422494
OLDCLAIM	8161	0	0	57037	0	4636	4037.07622	0	8777.1391	3.11904	9.860658
CLM_FREQ	8161	0	0	5	0	2	0.798554	0	1.158453	1.208799	0.284289
MVR_PTS	8161	0	0	13	0	3	1.695503	1	2.147112	1.34784	1.37549
CAR_AGE	8161	510	-3	28	1	12	8.328323	8	5.700742	0.281953	-0.748976
IMP_AGE	8161	0	16	81	39	51	44.790467	45	8.624419	-0.029053	-0.05954
M_AGE	8161	0	0	1	0	0	0.000735	0	0.027106	36.832982	1354.834579
IMP_YOJ	8161	0	0	23	9	13	10.527141	11	3.978654	-1.257251	1.451051
M_YOJ	8161	0	0	1	0	0	0.05563	0	0.229221	3.876745	13.030749
IMP_INCOME	8161	0	0	367030.262	28299.4384	85493.52994	61382.99471	54463.90395	46830.78647	1.194082	2.241705
M_INCOME	8161	0	0	1	0	0	0.054528	0	0.22707	3.92318	13.39298
IMP_HOME_VAL	8161	0	0	885282.345	0	233352.2675	155272.8254	162000	125409.7624	0.493292	0.156303
M_HOME_VAL	8161	0	0	1	0	0	0.056856	0	0.231581	3.826651	12.644811
IMP_CAR_AGE	8161	0	0	28	4	12	8.308173	8	5.519639	0.302791	-0.595943
M_CAR_AGE	8161	0	0	1	0	0	0.062615	0	0.242284	3.610076	11.034003
M_JOB	8161	0	0	1	0	0	0.064453	0	0.245573	3.546756	10.580774
Single_Parent_Ind	8161	0	0	1	0	0	0.131969	0	0.338478	2.174356	2.728159
MSTATUS_Single_Ind	8161	0	0	1	0	1	0.400319	0	0.489993	0.406819	-1.834723
SEX_Female_Ind	8161	0	0	1	0	1	0.536086	1	0.498727	-0.144696	-1.979306
Ed_Non_Degree_Ind	8161	0	0	1	0	1	0.432913	0	0.495509	0.270748	-1.926931
CU_Commercial_Ind	8161	0	0	1	0	1	0.371155	0	0.483144	0.533294	-1.715808
CT_Panel_Truck_Ind	8161	0	0	1	0	0	0.082833	0	0.275646	3.026455	7.160309
CT_Pickup_Ind	8161	0	0	1	0	0	0.1702	0	0.375831	1.75483	1.079559
CT_Sports_Car_Ind	8161	0	0	1	0	0	0.111138	0	0.314323	2.47398	4.121084
CT_Van_Ind	8161	0	0	1	0	0	0.091901	0	0.288903	2.824819	5.980333
CT_SUV_Ind	8161	0	0	1	0	1	0.281093	0	0.44956	0.973752	-1.051936
REVOKED_Ind	8161	0	0	1	0	0	0.122534	0	0.327922	2.30189	3.299101
UC_HUU_Ind	8161	0	0	1	1	1	0.795491	1	0.403367	-1.464941	0.146069
JOB_White_Collar_Ind	8161	0	0	1	0	1	0.6025	1	0.489411	-0.418818	-1.824815
JOB_Blue_Collar_Ind	8161	0	0	1	0	0	0.229751	0	0.420699	1.284604	-0.349834
JOB_Student_Ind	8161	0	0	1	0	0	0.087367	0	0.282389	2.922087	6.539395
Home_Owner_else_Renter_Ind	8161	0	0	1	0	1	0.718907	1	0.44956	-0.973752	-1.051936

Figure 14 shows all the variables that were dropped prior to model building.

Figure 14: Table – Dropped Variables

Variables Dropped			
INDEX	MSTATUS	RED_CAR	M_HOME_VAL
TARGET_AMT	SEX	REVOKED	IMP_CAR_AGE
AGE	EDUCATION	CAR_AGE	M_CAR_AGE
YOJ	JOB	URBANICITY	IMP_JOB
INCOME	CAR_USE	IMP_YOJ	M_JOB
PARENT1	BLUEBOOK	M_YOJ	
HOME_VAL	CAR_TYPE	IMP_HOME_VAL	

# Insurance Customer Risk Analysis Report

## Section 3: Model Building

For the model building, I took the automated variable selections approach as well manually selected the variables to include in the model. After trying different combinations (of course, combinations that made sense), we selected three models that gave us decent metrics and also logically made sense.

This section explains the three (3) models (Model 1, Model 2, and Model 3) that we have created to predict the probability of customer crashing the car. Out of these models, we have selected one (1) that fits the best and also logically correct to use for predicting the probability of a safer or riskier customer.

We have further split our Insurance training data into two sets, train and test. Seventy (70) percent of the data is split into train and 30 percent to test. This further split will help us better test the classification accuracy of our models and help us decide which one to pick based on various metrics.

### Section 3.1: Model 1

Figure 15 shows the summary output of the Model 1. Model 1 was created using the automated variable selection. We used the stepwise variable selection approach. In this approach, we start with one regressor in the model and continue to remove or add terms until removing or adding another term makes the criterion of interest worse i.e. increase AIC. We notice that variables, "SEX\_Female\_Ind" and "CT\_Panel\_Truck\_Ind" are not statistically significant. Hence, we would remove these variables in the next model. The AIC for this model is 5262 which seem to be fairly good. It would be interesting to see how other models turn out. Apart from the above variables not being statistically significant, this model seems to be counterintuitive for one variable, OLDCLAIM. For instance, it is more logical for a driver's risk to increase than decrease if they have more claims in the past 5 years.

Figure 15: Model 1: Output

```
Call:
glm(formula = TARGET_FLAG ~ CLM_FREQ + UC_HUU_Ind + IMP_INCOME +
    CU_Commercial_Ind + Single_Parent_Ind + MVR_PTS + REVOKED_Ind +
    TIF + Home_Owner_else_Renter_Ind + TRAVTIME + Ed_Non_Degree_Ind +
    KIDSDRIV + CT_Sports_Car_Ind + CT_SUV_Ind + MSTATUS_Single_Ind +
    CT_Pickup_Ind + CT_Van_Ind + OLDCLAIM + HOMEKIDS + SEX_Female_Ind +
    M_AGE + CT_Panel_Truck_Ind, family = binomial(link = "logit"),
    data = train.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5430  -0.7246  -0.4174   0.6416   3.1227

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.622598932  0.2252668591 -20.521 < 0.0000000000000002 ***
CLM_FREQ    0.1855597617  0.0336232548   5.519 0.00000003413372136 ***
UC_HUU_Ind   2.3331415518  0.1334947657  17.477 < 0.0000000000000002 ***
IMP_INCOME  -0.0000068421  0.0000009718  -7.041 0.00000000000191102 ***
CU_Commercial_Ind 0.8489134624  0.0833282928  10.188 < 0.0000000000000002 ***
Single_Parent_Ind 0.3021872930  0.1291519343   2.340 0.019295 ***
MVR_PTS      0.1229521748  0.0162909325   7.547 0.00000000000004445 ***
REVOKED_Ind  0.8510919112  0.1080916142   7.874 0.00000000000000344 ***
TIF          -0.0576556333  0.0087057154  -6.623 0.000000000003526163 ***
Home_Owner_else_Renter_Ind -0.2937104396  0.0891571286  -3.294 0.000987 ***
TRAVTIME     0.0148843784  0.002254993    6.688 0.000000000002260754 ***
Ed_Non_Degree_Ind 0.4875807551  0.0812224974   6.003 0.00000000193673719 ***
KIDSDRIV     0.3055667024  0.0720299907   4.242 0.00002213245730633 ***
CT_Sports_Car_Ind 1.1760249857  0.1410476149   8.338 < 0.0000000000000002 ***
CT_SUV_Ind    0.9419412540  0.1205439696   7.814 0.00000000000000554 ***
MSTATUS_Single_Ind 0.5462015607  0.0982280392   5.561 0.00000002689314821 ***
CT_Pickup_Ind 0.6444753820  0.1152655004   5.591 0.00000002254727910 ***
CT_Van_Ind    0.6225179115  0.1402811131   4.438 0.00000909480581901 ***
OLDCLAIM     -0.0000164601  0.0000046664  -3.527 0.000420 ***
HOMEKIDS     0.1137735140  0.0397550242   2.862 0.004212 **
SEX_Female_Ind -0.1588413649  0.1035221165  -1.534 0.124938
M_AGE        1.8963058536  1.3782157786   1.376 0.168848
CT_Panel_Truck_Ind 0.2321602986  0.1582527517   1.467 0.142370

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)


    Null deviance: 6629.2  on 5735  degrees of freedom
Residual deviance: 5216.3  on 5713  degrees of freedom
AIC: 5262.3
```



# Insurance Customer Risk Analysis Report

Figure 16 shows the Odds Ratio of the Model 1. The Model 1 tells us that, after controlling for the effects of all the variables, the odds of a driver being a high risk among people with claim frequency is 1.2 times the odds of a driver being a high risk among people without claim frequency. Similarly, after controlling for the effects of all the variables, the odds of a driver being a low risk among people with high income is 0.99 times the odds of a driver being a low risk among people with low income.

Figure 16: Model 1: Odds Ratio

	Odds Ratio 
(Intercept)	0.009827222
CLM_FREQ	1.203892144
UC_HUU_Ind	10.310280991
IMP_INCOME	0.999993158
CU_Commercial_Ind	2.337106118
Single_Parent_Ind	1.352814576
MVR_PTS	1.130830337
REVOKED_Ind	2.342202934
TIF	0.943974965
Home_Owner_else_Renter_Ind	0.745492325
TRAVTIME	1.014995702
Ed_Non_Degree_Ind	1.628372020
KIDSDRIV	1.357394023
CT_Sports_Car_Ind	3.241463696
CT_SUV_Ind	2.564955819
MSTATUS_Single_Ind	1.726681849
CT_Pickup_Ind	1.904987376
CT_Van_Ind	1.863614556
OLDCLAIM	0.999983540
HOMEKIDS	1.120498319
SEX_Female_Ind	0.853131685
M_AGE	6.661241333
CT_Panel_Truck_Ind	1.261321901

# Insurance Customer Risk Analysis Report

Figure 17 shows the ROC Curves of Model 1 on train and test data. Both exhibit a big bow that has all the lift in the beginning showing good characteristics of a good model.

Figure 17: Model 1: ROC Curves

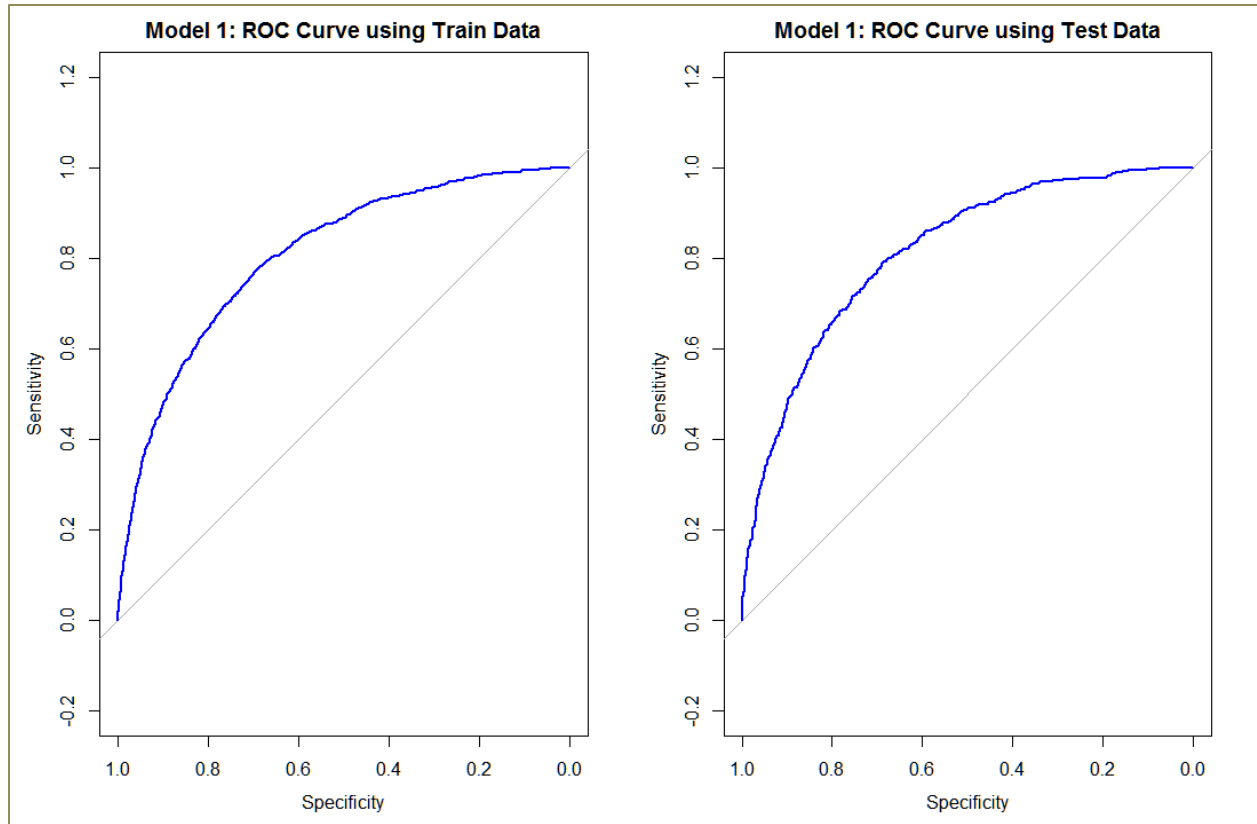


Figure 18 shows the statistics of the Model 1. We see the area under the curve (AUC) for the train data is 0.8061 which is roughly 80.61 % (approx.) and test data is 0.8126 which is roughly 81.26 % (approx.). Getting a better AUC for test data is a good sign that our model is not overfitting. KS Statistic for train and test data is 0.7354 and 0.7381, respectively. The classification accuracy for our model in train data set is 78.89% and test data set is 78.47%. This seems to be a really good model. However, it does have some counterintuitive effects as well as some variables are not statistically significant. Hence, we will decide on our final model in Section 4: Model Comparison and Selection.

Figure 18: Model 1: Statistics

Models	AIC	AUC - Train data	AUC - Test Data	KS Statistic - Train Data	KS Statistic - Test Data	Classification Accuracy - Train Data	Classification Accuracy - Test Data
Model_1	5262.302	0.8061	0.8126	0.7354	0.7381	0.7889	0.7847

# Insurance Customer Risk Analysis Report

## Section 3.2: Model 2

Figure 19 shows the summary output of the Model 2. Model 2 has pretty much all the variables except the ones that were not statistically significant and the one that was counterintuitive. The AIC for this model is 5283 which seem to slightly higher than Model 1. All the variables are statistically significant except one Flag variable, M\_INCOME. All the other variables intuitive as per the commonly held beliefs.

Figure 19: Model 2: Output

```
Call:
glm(formula = TARGET_FLAG ~ CLM_FREQ + UC_HUU_Ind + IMP_INCOME +
    M_INCOME + CU_Commercial_Ind + Single_Parent_Ind + REVOKED_Ind +
    MVR_PTS + TRAVTIME + Ed_Non_Degree_Ind + MSTATUS_Single_Ind +
    TIF + KIDSDRIV + CT_Sports_Car_Ind + CT_SUV_Ind + CT_Pickup_Ind +
    CT_Van_Ind + Home_Owner_else_Renter_Ind, family = binomial(link = "logit"),
    data = train.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5788  -0.7303  -0.4249   0.6553   3.0960

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    -4.5274604992  0.2184296194 -20.727 < 0.0000000000000002 ***
CLM_FREQ         0.1319755174  0.0299479544   4.407  0.000010489487369 ***
UC_HUU_Ind       2.3244910905  0.1333372092  17.433 < 0.0000000000000002 ***
IMP_INCOME      -0.0000068080  0.0000009624  -7.074  0.000000000001507 ***
M_INCOME         0.0256698028  0.1555013382   0.165    0.868883
CU_Commercial_Ind 0.8991351539  0.0768241378  11.704 < 0.0000000000000002 ***
Single_Parent_Ind 0.4860217949  0.1106816250   4.391  0.000011274251825 ***
REVOKED_Ind     0.6704261814  0.0946565132   7.083  0.000000000001413 ***
MVR_PTS         0.1193321951  0.0161484690   7.390  0.000000000000147 ***
TRAVTIME        0.0149330842  0.0022147017   6.743  0.0000000000015546 ***
Ed_Non_Degree_Ind 0.5013800591  0.0804923937   6.229  0.0000000000469685 ***
MSTATUS_Single_Ind 0.4664214992  0.0935587577   4.985  0.000000618554021 ***
TIF             -0.0570364238  0.0086777990  -6.573  0.0000000000049417 ***
KIDSDRIV        0.3897772420  0.0662036899   5.888  0.000000003919731 ***
CT_Sports_Car_Ind 1.0021040089  0.1178426659   8.504 < 0.0000000000000002 ***
CT_SUV_Ind       0.7847554245  0.0933286092   8.409 < 0.0000000000000002 ***
CT_Pickup_Ind    0.5595281011  0.1031261170   5.426  0.000000057738105 ***
CT_Van_Ind       0.5556656480  0.1268115434   4.382  0.000011769078638 ***
Home_Owner_else_Renter_Ind -0.3093577902  0.0889402170  -3.478    0.000505 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6629.2  on 5735  degrees of freedom
Residual deviance: 5245.0  on 5717  degrees of freedom
AIC: 5283
```

## Insurance Customer Risk Analysis Report

Figure 20 shows the Odds Ratio of the Model 2. The Model 2 tells us that, after controlling for the effects of all the variables, the odds of a driver being a high risk among people with commercial use vehicles is 2.46 times the odds of a driver being a high risk among people with private use vehicles. Similarly, after controlling for the effects of all the variables, the odds of a driver being a high risk among people with more traffic tickets is 1.23 times the odds of a driver being a high risk among people with fewer traffic tickets.

Figure 20: Model 2: Odds Ratio

	Odds Ratio <sup>a</sup>
(Intercept)	0.01080809
CLM_FREQ	1.14108038
UC_HUU_Ind	10.22147696
IMP_INCOME	0.99999319
M_INCOME	1.02600211
CU_Commercial_Ind	2.45747685
Single_Parent_Ind	1.62583543
REVOKED_Ind	1.95507036
MVR_PTS	1.12674416
TRAVTIME	1.01504514
Ed_Non_Degree_Ind	1.65099817
MSTATUS_Single_Ind	1.59427885
TIF	0.94455966
KIDSDRIV	1.47665182
CT_Sports_Car_Ind	2.72400714
CT_SUV_Ind	2.19187080
CT_Pickup_Ind	1.74984655
CT_Van_Ind	1.74310089
Home_Owner_else_Renter_Ind	0.73391813

# Insurance Customer Risk Analysis Report

Figure 21 shows the ROC Curves of Model 2 on train and test data. Both exhibit a big bow that has all the lift in the beginning showing good characteristics of a good model. ROC curves for Model 2 does not much differ from Model 1. In Section 4: Model Comparison and Selection, we will overlay the lines to better visually see and understand the difference in the ROC curves.

Figure 21: Model 2: ROC Curves

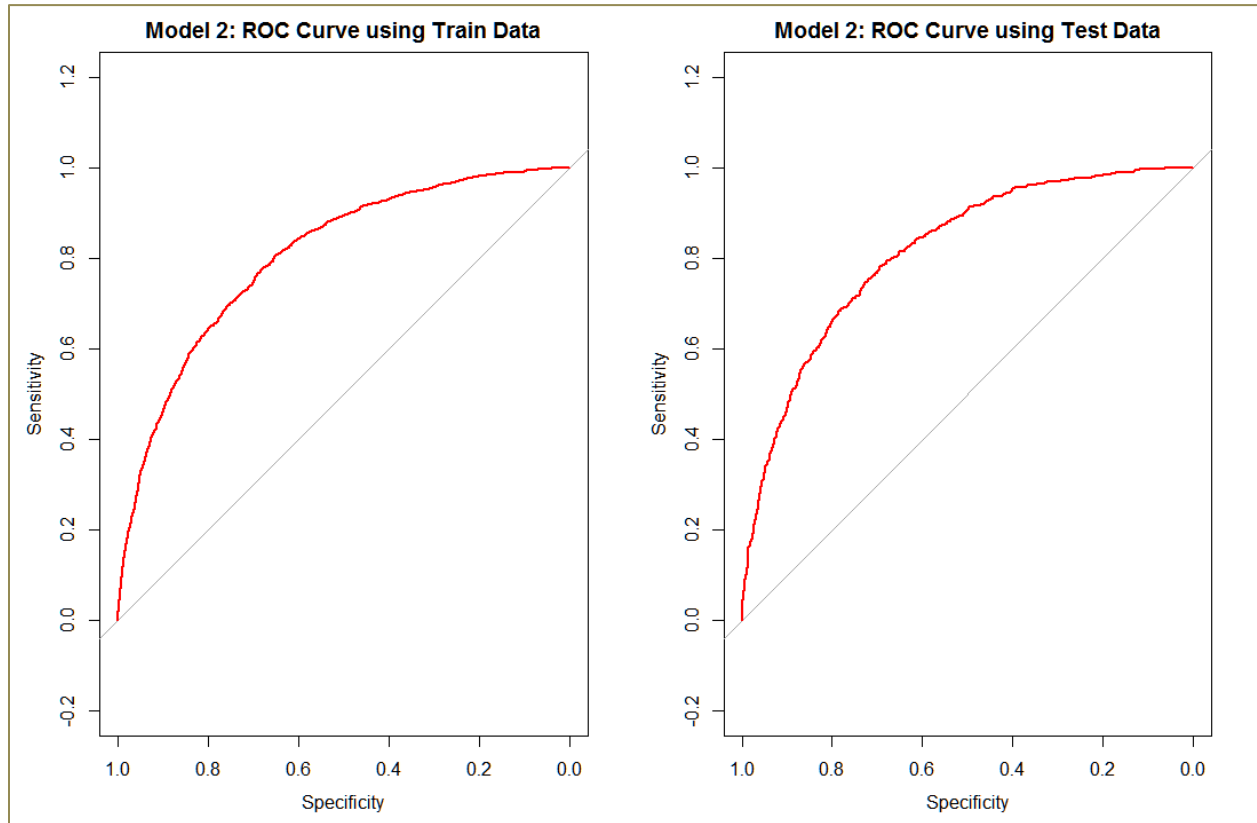


Figure 22 shows the statistics of the Model 2. We see the area under the curve (AUC) for the train data is 0.8033 which is roughly 80.33 % (approx.) and test data is 0.8127 which is roughly 81.27 % (approx.). Getting a better AUC for test data is a good sign that our model is not overfitting. KS Statistic for train and test data is 0.7354 and 0.7381, respectively. The classification accuracy for our model in train data set is 78.59% and test data set is 78.76%. This seems to be a really good model as well. So, this seems to be a model that could be selected as our final model.

Figure 22: Model 2: Statistics

Models	AIC	AUC - Train data	AUC - Test Data	KS Statistic - Train Data	KS Statistic - Test Data	Classification Accuracy - Train Data	Classification Accuracy - Test Data
Model_2	5283.038	0.8033	0.8127	0.7354	0.7381	0.7859	0.7876

# Insurance Customer Risk Analysis Report

## Section 3.3: Model 3

Figure 23 shows the summary output of the Model 3. Model 3 we have removed some of the variables from Model 2 and ran our model. The AIC for this model is 5322 which increased tremendously. This suggests that we did remove a statistically significant variable. All the variables are intuitive except one, OLDCLAIM. This model seems to be counterintuitive for one variable, OLDCLAIM. For instance, it is more logical for a driver's risk to increase than decrease if they have more claims in the past 5 years.

Figure 23: Model 3: Output

```
Call:
glm(formula = TARGET_FLAG ~ CLM_FREQ + UC_HUU_Ind + CU_Commercial_Ind +
  Single_Parent_Ind + REVOKED_Ind + MVR_PTS + TRAVTIME + Ed_Non_Degree_Ind +
  MSTATUS_Single_Ind + TIF + KIDSDRIV + CT_Sports_Car_Ind +
  CT_SUV_Ind + CT_Pickup_Ind + CT_Van_Ind + OLDCLAIM + Home_Owner_else_Renter_Ind,
  family = binomial(link = "logit"), data = train.df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5770  -0.7276  -0.4418   0.6421   3.1209

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.950131661  0.212440021 -23.301 < 0.0000000000000002 ***
CLM_FREQ    0.194854435  0.033399615   5.834 0.00000000541036294 ***
UC_HUU_Ind   2.241249302  0.132707428  16.889 < 0.0000000000000002 ***
CU_Commercial_Ind 0.830508532  0.075717500  10.969 < 0.0000000000000002 ***
Single_Parent_Ind 0.528878494  0.110298214   4.795 0.00000162685430026 ***
REVOKED_Ind  0.851714248  0.107138341   7.950 0.000000000000000187 ***
MVR_PTS      0.128683788  0.016161445   7.962 0.000000000000000169 ***
TRAVTIME     0.014742201  0.002214137   6.658 0.00000000002771728 ***
Ed_Non_Degree_Ind 0.790487891  0.071076924  11.122 < 0.0000000000000002 ***
MSTATUS_Single_Ind 0.399397423  0.093107210   4.290 0.00001789542330101 ***
TIF          -0.056554068  0.008649256  -6.539 0.00000000006209530 ***
KIDSDRIV     0.372785190  0.065901539   5.657 0.00000001543117741 ***
CT_Sports_Car_Ind 1.093623428  0.117141385   9.336 < 0.0000000000000002 ***
CT_SUV_Ind   0.850097481  0.092952750   9.145 < 0.0000000000000002 ***
CT_Pickup_Ind 0.623035321  0.102141996   6.100 0.00000000106269087 ***
CT_Van_Ind   0.525105296  0.125886886   4.171 0.00003029372650809 ***
OLDCLAIM     -0.000015827  0.000004641  -3.410 0.000649 ***
Home_Owner_else_Renter_Ind -0.370059599  0.088629290  -4.175 0.00002975091545165 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6629.2  on 5735  degrees of freedom
Residual deviance: 5286.4  on 5718  degrees of freedom
AIC: 5322.4
```

# Insurance Customer Risk Analysis Report

Figure 24 shows the Odds Ratio of the Model 3. The Model 3 tells us that, after controlling for the effects of all the variables, the odds of a driver being a high risk among people with a revoked license is 2.34 times the odds of a driver being a high risk among people with a non-revoked license. Similarly, after controlling for the effects of all the variables, the odds of a driver being a high risk among people with driving children is 1.45 times the odds of a driver being a high risk among people with less or no driving children.

Figure 24: Model 3: Odds Ratio

	Odds Ratio <sup>a</sup>
(Intercept)	0.007082476
CLM_FREQ	1.215134093
UC_HUU_Ind	9.405073727
CU_Commercial_Ind	2.294485263
Single_Parent_Ind	1.697028013
REVOKED_Ind	2.343661027
MVR_PTS	1.137330430
TRAVTIME	1.014851403
Ed_Non_Degree_Ind	2.204471706
MSTATUS_Single_Ind	1.490926030
TIF	0.945015388
KIDSDRIV	1.451772452
CT_Sports_Car_Ind	2.985070690
CT_SUV_Ind	2.339874935
CT_Pickup_Ind	1.864579057
CT_Van_Ind	1.690636857
OLDCLAIM	0.999984174
Home_Owner_else_Renter_Ind	0.690693164

# Insurance Customer Risk Analysis Report

Figure 25 shows the ROC Curves of Model 3 on train and test data. Both exhibit a big bow that has all the lift in the beginning showing good characteristics of a good model. ROC curves for Model 3 does not much differ from Model 1 and Model 2. In Section 4: Model Comparison and Selection, we will overlay the lines to better visually see and understand the difference in the ROC curves.

Figure 25: Model 3: ROC Curves

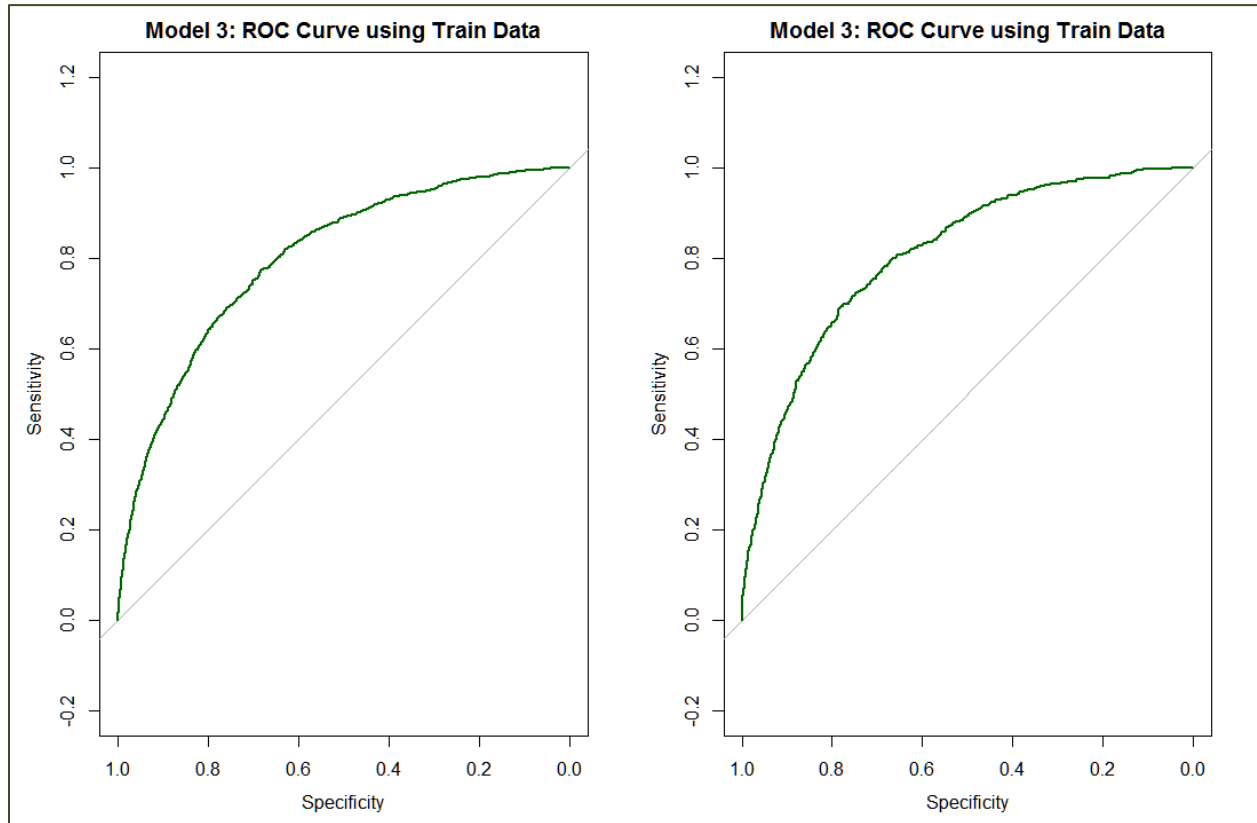


Figure 26 shows the statistics of the Model 3. We see the area under the curve (AUC) for the train data is 0.799 which is roughly 79.90% (approx.) and test data is 0.8069 which is roughly 80.69 % (approx.). Getting a better AUC for test data is a good sign that our model is not overfitting. KS Statistic for train and test data is 0.7354 and 0.7381, respectively. We have got the same KS-Statistic for all our model. The classification accuracy for our model in train data set is 78.33% and test data set is 78.76%.

Figure 26: Model 2: Statistics

Models	AIC	AUC - Train data	AUC - Test Data	KS Statistic - Train Data	KS Statistic - Test Data	Classification Accuracy - Train Data	Classification Accuracy - Test Data
Model_3	5322.351	0.799	0.8069	0.7354	0.7381	0.7833	0.7876



# Insurance Customer Risk Analysis Report

## Section 4: Model Comparison and Selection

Figure 27 shows each of our three models i.e. Model 1, Model 2, and Model 3. Any cell that is grayed out is not used in that model. The variables that are highlighted in Figure 27 means that those were not statistically significant.

Figure 27: Model Selection: Model 1, Model 2, and Model 3

Model 1	Model 2	Model 3
CLM_FREQ	CLM_FREQ	CLM_FREQ
CT_Panel_Truck_Ind		
CT_Pickup_Ind	CT_Pickup_Ind	CT_Pickup_Ind
CT_Sports_Car_Ind	CT_Sports_Car_Ind	CT_Sports_Car_Ind
CT_SUV_Ind	CT_SUV_Ind	CT_SUV_Ind
CT_Van_Ind	CT_Van_Ind	CT_Van_Ind
CU_Commercial_Ind	CU_Commercial_Ind	CU_Commercial_Ind
Ed_Non_Degree_Ind	Ed_Non_Degree_Ind	Ed_Non_Degree_Ind
Home_Owner_else_Renter_Ind	Home_Owner_else_Renter_Ind	Home_Owner_else_Renter_Ind
HOMEKIDS	HOMEKIDS	
IMP_INCOME	IMP_INCOME	
KIDSDRIV	KIDSDRIV	KIDSDRIV
M_AGE		
MSTATUS_Single_Ind	MSTATUS_Single_Ind	MSTATUS_Single_Ind
MVR_PTS	MVR_PTS	MVR_PTS
OLDCLAIM		OLDCLAIM
REVOKED_Ind	REVOKED_Ind	REVOKED_Ind
SEX_Female_Ind		
Single_Parent_Ind	Single_Parent_Ind	Single_Parent_Ind
TIF	TIF	TIF
TRAVTIME	TRAVTIME	TRAVTIME
UC_HUU_Ind	UC_HUU_Ind	UC_HUU_Ind
	M_INCOME	

Figure 27 shows the statistics of all three model. We notice that Model 1 has beaten the other two in the training dataset. However, it did not perform well in the test data set. Model 3 did not perform well in any of the metrics. Hence, it will not be considered for the final model. Between Model 1 and Model 2, I would select Model 2 as my final model because all the variables were intuitive and statistically significant as well as the AUC and classification accuracy for test data is better than Model 1.

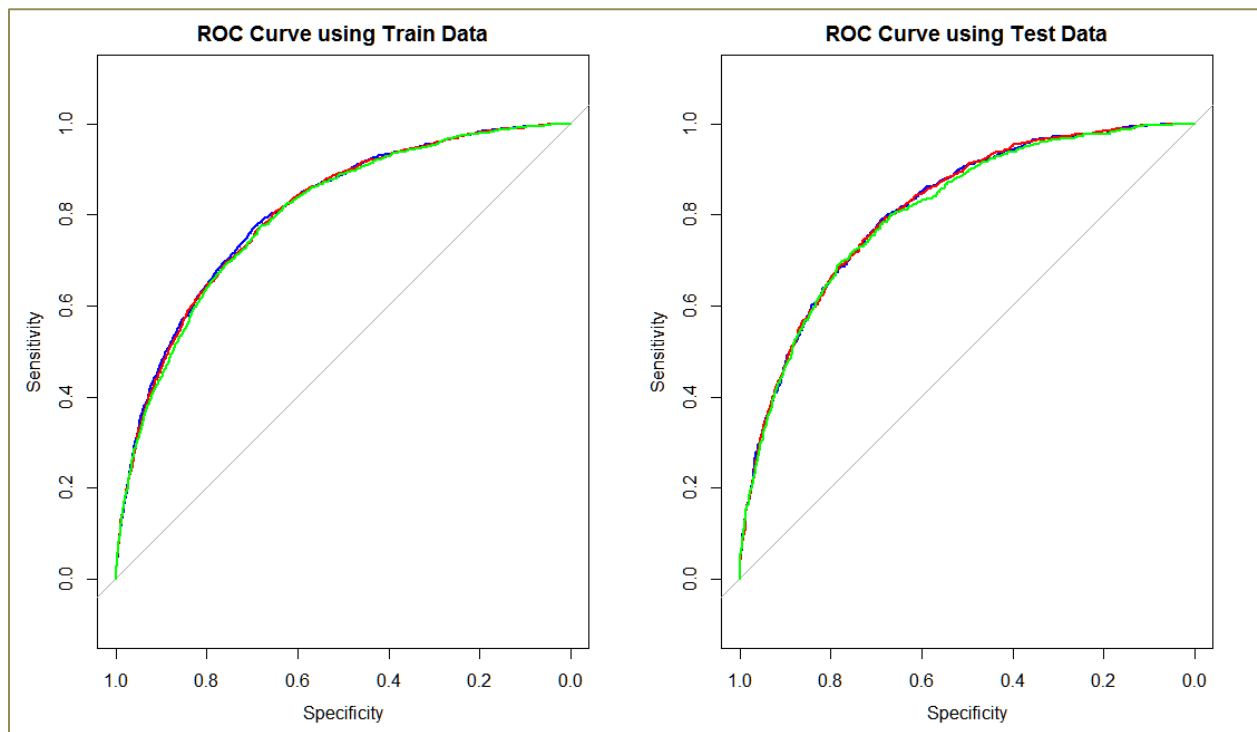
Figure 28: Model Selection: Statistics

Models	AIC	AUC - Train data	AUC - Test Data	KS Statistic - Train Data	KS Statistic - Test Data	Classification Accuracy - Train Data	Classification Accuracy - Test Data
Model_1	5262.302	0.8061	0.8126	0.7354	0.7381	0.7889	0.7847
Model_2	5283.038	0.8033	0.8127	0.7354	0.7381	0.7859	0.7876
Model_3	5322.351	0.799	0.8069	0.7354	0.7381	0.7833	0.7876

# Insurance Customer Risk Analysis Report

Figure 29 shows the ROC Curves for all three models on train and test data. Both exhibit a big bow that has all the lift in the beginning showing good characteristics of a good model. As it can be seen that all the model have a pretty much same shape. This is why the area under the curve was very close and we had our KS-Statistic identical for each model. Since Model 2 is logically sound and has better metrics in test data, we would go ahead and select this as our final model.

Figure 29: Models: ROC Curves



# Insurance Customer Risk Analysis Report

## Section 5: Model Testing and Scoring

Next, we use our Model 2 to score the INSURANCE Test data file ("test data"). Prior to running our model, we will clean the test data in a similar way as we cleaned our training dataset. We will fix missing values as well outliers in the data set. Note, a separate protocol document will accompany the stand-alone R program. This document will list the directions on how to import file and acquire results.

Figure 20 shows the summary statistics of the predicted probability and amount of payouts. We notice that we have all the observations i.e. 2141 observations, and there are no missing values. Our model predicted a minimum of 0 probability for a customer who would crash the car (this zero could be due to rounding of two digit) for a team and maximum of 0.95. In my opinion, the results seem to be OK. I would have been surprised or alarmed if our model predicted probability over 1.

Therefore, we will accept this score data result and use it as our final output.

Figure 30: Auto Insurance Test Data Result Stats

Variables	nobs	NAs	Minimum	Maximum	1. Quartile	3. Quartile	Mean	Median
P_TARGET_FLAG	2141	0	0	0.95	0.1	0.4	0.271966	0.22
P_TARGET_AMT	2141	0	4296.91	9633.54	5108.86	6450.73	5835.920906	5692.76

# Insurance Customer Risk Analysis Report

---

## Conclusion:

In conclusion, I would like to state that Model 1 and Model 2 were fairly close with the metrics. I doubt that we would have seen a major significant difference using that model. However, we picked a model that was logically sound as well provided us with best metrics.

We have also predicted the amounts for payouts if there was a car crash. We used a Linear Regression to determine the amount. Majority of the variables used did not help the model. Some variables were counterintuitive and others were not statistically significant. With the logical sense, we have used the variable, BLUEBOOK, to predict the amount to be paid out. It is my understanding that BLUEBOOK generally includes other factors such as car age, mileage, car type, etc. when determining the value of the car. Hence, using just that variable seems logically correct to me.

# Insurance Customer Risk Analysis Report

---

## Appendix I: Model Development R Code

```
#-----
# Auto Insurance Customer Risk Analysis
# Singh, Gurjeet
#-----

library(readr)
library(car)
library(fBasics)
library(ggplot2)
library(corrplot)
library(plyr)
library(gmodels)
library(MASS)
library(gridExtra)
library(pROC)

options(scipen = 999)
#-----
## 1 - DATA EXPLORATION
#-----

colnames(logit_insurance)[1] <- "INDEX"

View(logit_insurance)
#Understand the stats and summary
str(logit_insurance)
summary(logit_insurance)

#Getting the Frequency of the categorical variables
char_Freq <- lapply(logit_insurance[sapply(logit_insurance,
                                           is.character)], FUN = count)
char_stats <- ldply(char_Freq, data.frame)
names(char_stats) <- c("Variable Name", "Value",
                      "Frequency")

#reviewing the stats
View(t(basicStats(logit_insurance[sapply(logit_insurance,
                                           is.numeric)])))
View(char_stats)

#-----
##clean missing values with median values
#-----

summary(logit_insurance)
##clean missing values with median values

logit_insurance$IMP_AGE <- ifelse(is.na(logit_insurance$AGE),
                                45,
                                logit_insurance$AGE)
logit_insurance$M_AGE <- ifelse(is.na(logit_insurance$AGE),
                                1, 0)
```

# Insurance Customer Risk Analysis Report

---

```
logit_insurance$IMP_YOJ <-ifelse(is.na(logit_insurance$YOJ),
                                11,
                                logit_insurance$YOJ)
logit_insurance$M_YOJ <- ifelse(is.na(logit_insurance$YOJ),
                                1, 0)

logit_insurance$IMP_INCOME <-
  ifelse(is.na(logit_insurance$INCOME) &
        is.na(logit_insurance$JOB), 54000,
        ifelse(is.na(logit_insurance$INCOME) &
              (logit_insurance$JOB == "Doctor"), 128000,
              ifelse(is.na(logit_insurance$INCOME) &
                    (logit_insurance$JOB == "Lawyer"), 88000,
                    ifelse(is.na(logit_insurance$INCOME) &
                          (logit_insurance$JOB == "Manager"), 87000,
                          ifelse(is.na(logit_insurance$INCOME) &
                                (logit_insurance$JOB == "Professional"), 76000,
                                ifelse(is.na(logit_insurance$INCOME) &
                                      (logit_insurance$JOB == "Clerical"), 33000,
                                      ifelse(is.na(logit_insurance$INCOME) &
                                            (logit_insurance$JOB == "z_Blue Collar"), 58000,
                                            ifelse(is.na(logit_insurance$INCOME) &
                                                  (logit_insurance$JOB == "Home Maker"), 12000,
                                                  ifelse(is.na(logit_insurance$INCOME) &
                                                        (logit_insurance$JOB == "Student"), 6300,
                                                        logit_insurance$INCOME))))))))))

logit_insurance$M_INCOME <-
  ifelse(is.na(logit_insurance$INCOME),
        1, 0)

logit_insurance$IMP_HOME_VAL <-
  ifelse(is.na(logit_insurance$HOME_VAL),
        162000, logit_insurance$HOME_VAL)
logit_insurance$M_HOME_VAL <-
  ifelse(is.na(logit_insurance$HOME_VAL),
        1, 0)

logit_insurance$IMP_CAR_AGE <-
  ifelse(is.na(logit_insurance$CAR_AGE), 8,
  ifelse(logit_insurance$CAR_AGE < 0, 0,
  logit_insurance$CAR_AGE))

logit_insurance$M_CAR_AGE <-
  ifelse(is.na(logit_insurance$CAR_AGE), 1,
  ifelse(logit_insurance$CAR_AGE < 0, 1,
  0))

logit_insurance$IMP_JOB <-
  ifelse(is.na(logit_insurance$JOB) &
        logit_insurance$IMP_INCOME > 150000, "Doctor",
  ifelse(is.na(logit_insurance$JOB) &
```

# Insurance Customer Risk Analysis Report

---

```
logit_insurance$IMP_INCOME > 100000, "Lawyer",
ifelse(is.na(logit_insurance$JOB) &
logit_insurance$IMP_INCOME > 85000, "Manager",
ifelse(is.na(logit_insurance$JOB) &
logit_insurance$IMP_INCOME > 75000, "Professional",
ifelse(is.na(logit_insurance$JOB) &
logit_insurance$IMP_INCOME > 60000, "z_Blue Collar",
ifelse(is.na(logit_insurance$JOB) &
logit_insurance$IMP_INCOME > 35000, "Clerical",
ifelse(is.na(logit_insurance$JOB) &
logit_insurance$IMP_INCOME >= 12000, "Home Maker",
ifelse(is.na(logit_insurance$JOB) &
logit_insurance$IMP_INCOME < 12000, "Student",
logit_insurance$JOB))))))

logit_insurance$M_JOB <- ifelse(is.na(logit_insurance$JOB),
                                1, 0)

#Print After fixing the values
char_Freq2 <- lapply(logit_insurance[is.character], FUN = count)
char_stats2 <- ldply(char_Freq2, data.frame)
names(char_stats2) <- c("Variable Name", "Value", "Frequency")

options(scipen = 999)
View(t(basicStats(logit_insurance[is.numeric])))
View(char_stats2)

logit_insurance[which(logit_insurance$IMP_JOB == "Doctor" &
logit_insurance$IMP_INCOME ==0),
c("IMP_JOB", "IMP_INCOME", "IMP_YOJ", "M_YOJ")]
logit_insurance[which(logit_insurance$IMP_JOB == "Lawyer" &
logit_insurance$IMP_INCOME ==0),
c("IMP_JOB", "IMP_INCOME", "IMP_YOJ", "M_YOJ")]
logit_insurance[which(logit_insurance$IMP_JOB == "Manager" &
logit_insurance$IMP_INCOME ==0),
c("IMP_JOB", "IMP_INCOME", "IMP_YOJ", "M_YOJ")]
logit_insurance[which(logit_insurance$IMP_JOB == "Professional" &
logit_insurance$IMP_INCOME ==0),
c("IMP_JOB", "IMP_INCOME", "IMP_YOJ", "M_YOJ")]
logit_insurance[which(logit_insurance$IMP_JOB == "z_Blue Collar" &
logit_insurance$IMP_INCOME ==0),
c("IMP_JOB", "IMP_INCOME", "IMP_YOJ", "M_YOJ")]
logit_insurance[which(logit_insurance$IMP_JOB == "Home Maker" &
logit_insurance$IMP_INCOME ==0),
c("IMP_JOB", "IMP_INCOME", "IMP_YOJ", "M_YOJ")]
logit_insurance[which(logit_insurance$IMP_JOB == "Student" &
logit_insurance$IMP_INCOME ==0),
c("IMP_JOB", "IMP_INCOME", "IMP_YOJ", "M_YOJ")]

#-----
##Histograms - Exploration
#-----
```

# Insurance Customer Risk Analysis Report

---

```
##-----
## Numerical Value by TARGET_FLAG
##-----

#Predictable - yes use it
plot1 <-
  ggplot(logit_insurance,mapping = aes(x=KIDSDRIV,
                                         y = (..density..) ,fill=TARGET_FLAG))+
  geom_histogram(colour="black") +
  facet_grid(~TARGET_FLAG, labeller = label_both )+theme_bw() +
  labs(title = "Histogram of KIDSDRIV by TARGET_FLAG",
        x = "KIDSDRIV", y = "Density") +
  scale_y_continuous(breaks = seq(0, 10,by = 0.5))

#Predictable - yes use it
plot2 <- ggplot(logit_insurance,mapping = aes(x=IMP_AGE,
                                              y = (..density..) ,fill=TARGET_FLAG))+
  geom_histogram(colour="black") +
  facet_grid(~TARGET_FLAG, labeller = label_both )+theme_bw() +
  labs(title = "Histogram of IMP_AGE by TARGET_FLAG",
        x = "IMP_AGE", y = "Density") +
  scale_y_continuous(breaks = seq(0, 1,by = 0.005))

#Predictable - yes use it
plot3 <- ggplot(logit_insurance,mapping = aes(x=HOMEKIDS,
                                              y = (..density..) ,fill=TARGET_FLAG))+
  geom_histogram(colour="black") +
  facet_grid(~TARGET_FLAG, labeller = label_both )+theme_bw() +
  labs(title = "Histogram of HOMEKIDS by TARGET_FLAG",
        x = "HOMEKIDS", y = "Density") +
  scale_y_continuous(breaks = seq(0, 6,by = 0.5))

#Predictable - NO
plot4 <- ggplot(logit_insurance,mapping = aes(x=IMP_YOJ,
                                              y = (..density..) ,fill=TARGET_FLAG))+
  geom_histogram(colour="black") +
  facet_grid(~TARGET_FLAG, labeller = label_both )+theme_bw() +
  labs(title = "Histogram of IMP_YOJ by TARGET_FLAG",
        x = "IMP_YOJ", y = "Density") +
  scale_y_continuous(breaks = seq(0, 6,by = 0.02))

grid.arrange(plot1, plot2,plot3, plot4, nrow = 2)

#Predictable - Could be but those people might be just renters
plot5 <- ggplot(logit_insurance,mapping = aes(x=IMP_HOME_VAL,
                                              y = (..density..) * 100000 ,fill=TARGET_FLAG))+
  geom_histogram(colour="black") +
  facet_grid(~TARGET_FLAG, labeller = label_both )+theme_bw() +
  labs(title = "Histogram of IMP_HOME_VAL by TARGET_FLAG",
        x = "IMP_HOME_VAL", y = "Density X 100,000") +
  scale_y_continuous(breaks = seq(0, 2,by = 0.1))

#Predictable - yes use it
plot6 <- ggplot(logit_insurance,mapping = aes(x=IMP_INCOME,
                                              y = (..density..) * 100000 ,fill=TARGET_FLAG))+
```



# Insurance Customer Risk Analysis Report

---

```
geom_histogram(colour="black") +
facet_grid(~TARGET_FLAG, labeller = label_both )+theme_bw() +
labs(title = "Histogram of IMP_INCOME by TARGET_FLAG",
      x = "IMP_INCOME", y = "Density X 100,000") +
scale_y_continuous(breaks = seq(0, 2,by = 0.1))

#Predictable - yes use it
plot7 <- ggplot(logit_insurance,mapping = aes(x=TRAVTIME,
      y = (..density..) * 100 ,fill=TARGET_FLAG))+
geom_histogram(colour="black") +
facet_grid(~TARGET_FLAG, labeller = label_both )+theme_bw() +
labs(title = "Histogram of TRAVTIME by TARGET_FLAG",
      x = "TRAVTIME", y = "Density * 100") +
scale_y_continuous(breaks = seq(0, 5,by = 0.5))

#Predictable - yes use it
plot8 <- ggplot(logit_insurance,mapping = aes(x=BLUEBOOK,
      y = (..density..) * 100000 ,fill=TARGET_FLAG))+
geom_histogram(colour="black") +
facet_grid(~TARGET_FLAG, labeller = label_both )+theme_bw() +
labs(title = "Histogram of BLUEBOOK by TARGET_FLAG",
      x = "BLUEBOOK", y = "Density X 100,000") +
scale_y_continuous(breaks = seq(0, 10,by = 0.5))

grid.arrange(plot5, plot6, plot7, plot8, nrow = 2)

#Predictable - yes use it
plot9 <- ggplot(logit_insurance,mapping = aes(x=TIF,
      y = (..density..) * 10 ,fill=TARGET_FLAG))+
geom_histogram(colour="black") +
facet_grid(~TARGET_FLAG, labeller = label_both )+theme_bw() +
labs(title = "Histogram of TIF by TARGET_FLAG",
      x = "TIF", y = "Density X 10") +
scale_y_continuous(breaks = seq(0, 5,by = 0.5))

#Predictable - yes use it
plot10 <- ggplot(logit_insurance,mapping = aes(x=OLDCLAIM,
      y = (..density..) * 1000 ,fill=TARGET_FLAG))+
geom_histogram(colour="black") +
facet_grid(~TARGET_FLAG, labeller = label_both )+theme_bw() +
labs(title = "Histogram of OLDCLAIM by TARGET_FLAG",
      x = "OLDCLAIM", y = "Density X 1000") +
scale_y_continuous(breaks = seq(0, 1,by = 0.05))

#Predictable - yes use it
plot11 <- ggplot(logit_insurance,mapping = aes(x=CLM_FREQ,
      y = (..density..) ,fill=TARGET_FLAG))+
geom_histogram(colour="black") +
facet_grid(~TARGET_FLAG, labeller = label_both )+theme_bw() +
labs(title = "Histogram of CLM_FREQ by TARGET_FLAG",
      x = "CLM_FREQ", y = "Density") +
scale_y_continuous(breaks = seq(0, 10,by = 0.5))

#Predictable - yes use it
plot12 <- ggplot(logit_insurance,mapping = aes(x=MVR_PTS,
```

# Insurance Customer Risk Analysis Report

---

```
      y = (..density..) ,fill=TARGET_FLAG))+
geom_histogram(colour="black") +
facet_grid(~TARGET_FLAG, labeller = label_both )+theme_bw() +
labs(title = "Histogram of MVR_PTS by TARGET_FLAG",
      x = "MVR_PTS", y = "Density") +
scale_y_continuous(breaks = seq(0, 2,by = 0.1))

grid.arrange(plot9, plot10, plot11, plot12, nrow = 2)

#Predictable - yes use it
plot13 <- ggplot(logit_insurance,mapping = aes(x=IMP_CAR_AGE,
      y = (..density..) ,fill=TARGET_FLAG))+
geom_histogram(colour="black") +
facet_grid(~TARGET_FLAG, labeller = label_both )+theme_bw() +
labs(title = "Histogram of IMP_CAR_AGE by TARGET_FLAG",
      x = "IMP_CAR_AGE", y = "Density") +
scale_y_continuous(breaks = seq(0, 1,by = 0.05))

##-----
## Categorical Value
##-----

# Predictable - % higher for single parent
plot14 <- ggplot(logit_insurance,
      mapping = aes(x=TARGET_FLAG,
      y = (..density..), group = PARENT1,fill=PARENT1))+
geom_histogram(colour="black") +
facet_grid(~PARENT1, labeller = label_both )+theme_bw() +
labs(title = "Histogram of TARGET_FLAG by PARENT1",
      x = "TARGET_FLAG", y = "Density") +
scale_y_continuous(breaks = seq(0, 30,by =5))

# Predictable - % higher for Not married people
plot15 <- ggplot(logit_insurance,
      mapping = aes(x=TARGET_FLAG,
      y = (..density..), group = MSTATUS,fill=MSTATUS))+
geom_histogram(colour="black") +
facet_grid(~MSTATUS, labeller = label_both )+theme_bw() +
labs(title = "Histogram of TARGET_FLAG by MSTATUS",
      x = "TARGET_FLAG", y = "Density") +
scale_y_continuous(breaks = seq(0, 30,by =5))

# Predictable - Could be - CounterIntitutive - Female at higher risk
plot16 <- ggplot(logit_insurance,
      mapping = aes(x=TARGET_FLAG,
      y = (..density..), group = SEX,fill=SEX))+
geom_histogram(colour="black") +
facet_grid(~SEX, labeller = label_both )+theme_bw() +
labs(title = "Histogram of TARGET_FLAG by SEX",
      x = "TARGET_FLAG", y = "Density") +
scale_y_continuous(breaks = seq(0, 30,by =5))

grid.arrange(plot13, plot14, plot15, plot16, nrow = 2)

# Predictable - % is higher. Use Indicator variable Degree vs Non-degree
plot17 <- ggplot(logit_insurance,
```

# Insurance Customer Risk Analysis Report

---

```
        mapping = aes(x=TARGET_FLAG,
                      y = (..density..), group = EDUCATION,fill=EDUCATION))+
geom_histogram(colour="black") +
facet_grid(~EDUCATION )+theme_bw() +
#facet_grid(~EDUCATION, labeller = label_both )+theme_bw() +
labs(title = "Histogram of TARGET_FLAG by EDUCATION",
      x = "TARGET_FLAG", y = "Density") +
scale_y_continuous(breaks = seq(0, 100,by =5)) +
scale_x_continuous(breaks = seq(0, 1,by = 0.5))

# Predictable - % is higher. Use Indicator variable White collar vs Blue
collar
plot18 <- ggplot(logit_insurance,
                mapping = aes(x=TARGET_FLAG,
                              y = (..density..), group = IMP_JOB,fill=IMP_JOB))+
geom_histogram(colour="black") +
facet_grid(~IMP_JOB )+theme_bw() +
labs(title = "Histogram of TARGET_FLAG by IMP_JOB",
      x = "TARGET_FLAG", y = "Density") +
scale_y_continuous(breaks = seq(0, 30,by =5)) +
scale_x_continuous(breaks = seq(0, 1,by = 0.5))

# Predictable - % is higher for commercial
plot19 <- ggplot(logit_insurance,
                mapping = aes(x=TARGET_FLAG,
                              y = (..density..), group = CAR_USE,fill=CAR_USE))+
geom_histogram(colour="black") +
facet_grid(~CAR_USE)+theme_bw() +
labs(title = "Histogram of TARGET_FLAG by CAR_USE",
      x = "TARGET_FLAG", y = "Density") +
scale_y_continuous(breaks = seq(0, 30,by =5)) +
scale_x_continuous(breaks = seq(0, 1,by = 0.5))

# Predictable - check with average Target flag == 1 and determine.
plot20 <- ggplot(logit_insurance,
                mapping = aes(x=TARGET_FLAG,
                              y = (..density..), group = CAR_TYPE,fill=CAR_TYPE))+
geom_histogram(colour="black") +
facet_grid(~CAR_TYPE )+theme_bw() +
labs(title = "Histogram of TARGET_FLAG by CAR_TYPE",
      x = "TARGET_FLAG", y = "Density") +
scale_y_continuous(breaks = seq(0, 30,by =5)) +
scale_x_continuous(breaks = seq(0, 1,by = 0.5))

grid.arrange(plot17, plot18, plot19, plot20, nrow = 2)

# Predictable - NO
plot21 <- ggplot(logit_insurance,
                mapping = aes(x=TARGET_FLAG,
                              y = (..density..), group = RED_CAR,fill=RED_CAR))+
geom_histogram(colour="black") +
facet_grid(~RED_CAR )+theme_bw() +
labs(title = "Histogram of TARGET_FLAG by RED_CAR",
      x = "TARGET_FLAG", y = "Density") +
```

# Insurance Customer Risk Analysis Report

---

```
scale_y_continuous(breaks = seq(0, 30, by = 2)) +
scale_x_continuous(breaks = seq(0, 1, by = 0.5))

# Predictable - YES
plot22 <- ggplot(logit_insurance,
  mapping = aes(x=TARGET_FLAG,
    y = (..density..), group = REVOKED, fill=REVOKED)) +
  geom_histogram(colour="black") +
  facet_grid(~REVOKED) + theme_bw() +
  labs(title = "Histogram of TARGET_FLAG by REVOKED",
    x = "TARGET_FLAG", y = "Density") +
  scale_y_continuous(breaks = seq(0, 30, by = 5)) +
  scale_x_continuous(breaks = seq(0, 1, by = 0.5))

# Predictable - YES - Urban city is higher for urban
plot23 <- ggplot(logit_insurance,
  mapping = aes(x=TARGET_FLAG,
    y = (..density..), group = URBANICITY, fill=URBANICITY)) +
  geom_histogram(colour="black") +
  facet_grid(~URBANICITY) + theme_bw() +
  labs(title = "Histogram of TARGET_FLAG by URBANICITY",
    x = "TARGET_FLAG", y = "Density") +
  scale_y_continuous(breaks = seq(0, 30, by = 5)) +
  scale_x_continuous(breaks = seq(0, 1, by = 0.5))

grid.arrange(plot21, plot22, plot23, nrow = 2)

#-----
##Variable selection Exploration
#-----

#This provides the total Accident versus non accident splits with prop
CrossTable('Total' = logit_insurance$JOB, logit_insurance$TARGET_FLAG)

CrossTable(logit_insurance$KIDSDRIV,
  logit_insurance$TARGET_FLAG,
  prop.r = TRUE, prop.c = FALSE,
  prop.t = FALSE, prop.chisq = FALSE,
  row.labels = TRUE,
  dnn = c("KIDSDRIV", "TARGET_FLAG"))

TRUECrossTable(logit_insurance$HOMEKIDS,
  logit_insurance$TARGET_FLAG,
  prop.r = TRUE, prop.c = FALSE,
  prop.t = FALSE, prop.chisq = FALSE,
  row.labels = TRUE,
  dnn = c("HOMEKIDS", "TARGET_FLAG"))

CrossTable(logit_insurance$PARENT1,
  logit_insurance$TARGET_FLAG,
  prop.r = TRUE, prop.c = FALSE,
```

# Insurance Customer Risk Analysis Report

---

```
prop.t = FALSE, prop.chisq = FALSE,  
row.labels = TRUE,  
dnn = c("PARENT1", "TARGET_FLAG"))
```

```
CrossTable(logit_insurance$MSTATUS,  
  logit_insurance$TARGET_FLAG,  
  prop.r = TRUE, prop.c = FALSE,  
  prop.t = FALSE, prop.chisq = FALSE,  
  row.labels = TRUE,  
  dnn = c("MSTATUS", "TARGET_FLAG"))
```

```
CrossTable(logit_insurance$SEX,  
  logit_insurance$TARGET_FLAG,  
  prop.r = TRUE, prop.c = FALSE,  
  prop.t = FALSE, prop.chisq = FALSE,  
  row.labels = TRUE,  
  dnn = c("SEX", "TARGET_FLAG"))
```

```
CrossTable(logit_insurance$EDUCATION,  
  logit_insurance$TARGET_FLAG,  
  prop.r = TRUE, prop.c = FALSE,  
  prop.t = FALSE, prop.chisq = FALSE,  
  row.labels = TRUE,  
  dnn = c("EDUCATION", "TARGET_FLAG"))
```

```
CrossTable(logit_insurance$JOB,  
  logit_insurance$TARGET_FLAG,  
  prop.r = TRUE, prop.c = FALSE,  
  prop.t = FALSE, prop.chisq = FALSE,  
  row.labels = TRUE,  
  dnn = c("JOB", "TARGET_FLAG"))
```

```
CrossTable(logit_insurance$CAR_USE,  
  logit_insurance$TARGET_FLAG,  
  prop.r = TRUE, prop.c = FALSE,  
  prop.t = FALSE, prop.chisq = FALSE,  
  row.labels = TRUE,  
  dnn = c("CAR_USE", "TARGET_FLAG"))
```

```
CrossTable(logit_insurance$CAR_TYPE,  
  logit_insurance$TARGET_FLAG, prop.r = TRUE,  
  prop.c = FALSE,  
  prop.t = FALSE, prop.chisq = FALSE,  
  row.labels = TRUE,  
  dnn = c("CAR_TYPE", "TARGET_FLAG"))
```

```
CrossTable(logit_insurance$RED_CAR,  
  logit_insurance$TARGET_FLAG,  
  prop.r = TRUE, prop.c = FALSE,  
  prop.t = FALSE, prop.chisq = FALSE,
```

# Insurance Customer Risk Analysis Report

---

```
row.labels = TRUE,
dnn = c("RED_CAR", "TARGET_FLAG"))

CrossTable(logit_insurance$REVOKED,
  logit_insurance$TARGET_FLAG,
  prop.r = TRUE, prop.c = FALSE,
  prop.t = FALSE, prop.chisq = FALSE, row.labels = TRUE,
  dnn = c("REVOKED", "TARGET_FLAG"))

CrossTable(logit_insurance$URBANICITY,
  logit_insurance$TARGET_FLAG, prop.r = TRUE,
  prop.c = FALSE,
  prop.t = FALSE, prop.chisq = FALSE,
  row.labels = TRUE,
  dnn = c("URBANICITY", "TARGET_FLAG"))

CrossTable(logit_insurance$IMP_AGE,
  logit_insurance$TARGET_FLAG,
  prop.r = TRUE, prop.c = FALSE,
  prop.t = FALSE, prop.chisq = FALSE,
  row.labels = TRUE,
  dnn = c("IMP_AGE", "TARGET_FLAG"))

CrossTable(logit_insurance$IMP_YOJ,
  logit_insurance$TARGET_FLAG,
  prop.r = TRUE, prop.c = FALSE,
  prop.t = FALSE, prop.chisq = FALSE, row.labels = TRUE,
  dnn = c("IMP_YOJ", "TARGET_FLAG"))

CrossTable(logit_insurance$IMP_CAR_AGE,
  logit_insurance$TARGET_FLAG,
  prop.r = TRUE, prop.c = FALSE,
  prop.t = FALSE, prop.chisq = FALSE, row.labels = TRUE,
  dnn = c("IMP_CAR_AGE", "TARGET_FLAG"))

#-----
## 2 - DATA PREPARATION
#-----

#-----
##Indicator Variables
#-----

logit_insurance$Single_Parent_Ind <- ifelse(logit_insurance$PARENT1
=='Yes',1,0);
logit_insurance$MSTATUS_Single_Ind <- ifelse(logit_insurance$MSTATUS
=='Yes',0,1);
logit_insurance$SEX_Female_Ind <- ifelse(logit_insurance$SEX =='M',0,1);
logit_insurance$Ed_Non_Degree_Ind <- ifelse(logit_insurance$EDUCATION
=='<High School' |
```

# Insurance Customer Risk Analysis Report

---

```
logit_insurance$EDUCATION
=='z_High School', 1,0)
logit_insurance$CU_Commercial_Ind <- ifelse(logit_insurance$CAR_USE
=='Commercial',1,0);

logit_insurance$CT_Panel_Truck_Ind <- ifelse(logit_insurance$CAR_TYPE
=='Panel Truck',1,0);
logit_insurance$CT_Pickup_Ind <- ifelse(logit_insurance$CAR_TYPE
=='Pickup',1,0);
logit_insurance$CT_Sports_Car_Ind <- ifelse(logit_insurance$CAR_TYPE
=='Sports Car',1,0);
logit_insurance$CT_Van_Ind <- ifelse(logit_insurance$CAR_TYPE == 'Van',1,0);
logit_insurance$CT_SUV_Ind <- ifelse(logit_insurance$CAR_TYPE == 'z_SUV',1,0);

logit_insurance$REVOKED_Ind <- ifelse(logit_insurance$REVOKED == 'Yes',1,0);

logit_insurance$UC_HUU_Ind <- ifelse(logit_insurance$URBANICITY == 'Highly
Urban/ Urban',1,0);

logit_insurance$JOB_White_Collar_Ind <- ifelse(logit_insurance$IMP_JOB
=='Clerical' |
logit_insurance$IMP_JOB
=='Doctor' |
logit_insurance$IMP_JOB
=='Lawyer' |
logit_insurance$IMP_JOB
=='Manager' |
logit_insurance$IMP_JOB
=='Professional', 1,0)
logit_insurance$JOB_Blue_Collar_Ind <- ifelse(logit_insurance$IMP_JOB
=='z_Blue Collar',1,0);
logit_insurance$JOB_Student_Ind <- ifelse(logit_insurance$IMP_JOB
=='Student',1,0);

logit_insurance$Home_Owner_else_Renter_Ind <-
ifelse(logit_insurance$IMP_HOME_VAL != 0 ,1,0);

View(t(basicStats(logit_insurance[sapply(logit_insurance,is.numeric)])))

#-----
##creating a drop list
#-----
names(logit_insurance)

logit_insurance_LinearModel <- logit_insurance
# logit_insurance <- logit_insurance2

#creating a drop list to remove not required variables.
drop.list <- c('INDEX', 'TARGET_AMT', 'AGE', 'YOJ',
              'INCOME', 'PARENT1',
              'HOME_VAL', 'MSTATUS', 'SEX', 'EDUCATION', 'JOB',
              'CAR_USE', 'BLUEBOOK', 'CAR_TYPE',
              'RED_CAR', 'REVOKED', 'CAR_AGE', 'URBANICITY',
```

# Insurance Customer Risk Analysis Report

---

```
'M_CAR_AGE', 'IMP_YOJ', 'M_YOJ', 'IMP_HOME_VAL', 'M_HOME_VAL', 'IMP_CAR_AGE',
'M_CAR_AGE',
'IMP_JOB', 'M_JOB'
)

#dropping the variables
logit_insurance <- logit_insurance[,!(names(logit_insurance) %in% drop.list
)]

names(logit_insurance)
summary(logit_insurance)

#-----
#Requirement # Add a train/test flag to split the sample
#-----

logit_insurance$u <- runif(n=dim(logit_insurance)[1],min=0,max=1);

logit_insurance$train <- ifelse(logit_insurance$u<0.70,1,0);

# Save the R data frame as an .RData object
saveRDS(logit_insurance,file= );

# Check the counts on the train/test split
table(logit_insurance$train)

# Check the train/test split as a percentage of whole
table(logit_insurance$train)/dim(logit_insurance)[1]

#-----
## 3- Build Models
#-----
# Read (or reload) the .RData object as an R data frame
logit_insurance <- readRDS(file= "logit_insurance_train.RData");

names(logit_insurance)

# Create train/test split;
train.df <- subset(logit_insurance, train==1);
test.df <- subset(logit_insurance, train==0);

cutoff <- 0.50

# #-----
# ### ##Model_1_GLM - STEP AIC
# #-----
# Define the upper model as the FULL model
```



# Insurance Customer Risk Analysis Report

---

```
upper.glm <- glm(TARGET_FLAG ~ . -train -u
                 ,family = binomial(link = "logit")
                 ,data = train.df )
summary(upper.glm)

# Define the lower model as the Intercept model
lower.glm <- glm(TARGET_FLAG ~ 1
                 ,family = binomial(link = "logit")
                 ,data = train.df );
summary(lower.glm)

# Need a SLR to initialize stepwise selection
start.glm <- glm(TARGET_FLAG ~ CLM_FREQ
                 ,family = binomial(link = "logit")
                 ,data = train.df );
summary(start.glm)

Model_1_GLM <-
stepAIC(object=start.glm,scope=list(upper=formula(upper.glm),lower=~1),
        direction=c('both'));
summary(Model_1_GLM)

#ODDS Ratio
Odds_Ratio_Mod_1 <- as.data.frame(exp(Model_1_GLM$coefficients))
names(Odds_Ratio_Mod_1) <- "Odds Ratio"
View(Odds_Ratio_Mod_1)

## ROC Curve and AUC based on Train data set
Model_1_GLM_pred_train <- predict(Model_1_GLM, type = "response")
ROC_Model_1_GLM_train <- roc(train.df$TARGET_FLAG, Model_1_GLM_pred_train )

##test data
Model_1_GLM_pred_test <- predict(Model_1_GLM, newdata = test.df, type =
"response")
ROC_Model_1_GLM_test <- roc(test.df$TARGET_FLAG, Model_1_GLM_pred_test )

#confusion Matrix work
class_pred_Model_1_train <- ifelse(Model_1_GLM_pred_train > cutoff,1,0)
class_pred_Model_1_test <- ifelse(Model_1_GLM_pred_test > cutoff,1,0)

Conf_Matrix_Model_1_train <-
table(train.df$TARGET_FLAG,class_pred_Model_1_train)
Conf_Matrix_Model_1_test <-
table(test.df$TARGET_FLAG,class_pred_Model_1_test)

#compute the classification accuraxy
acc_Model_1_train <- sum(diag(Conf_Matrix_Model_1_train))/nrow(train.df)
acc_Model_1_test <- sum(diag(Conf_Matrix_Model_1_test))/nrow(test.df)

par(mfrow=c(1,2))
## ROC Curve for Test and Train data set
plot.roc(ROC_Model_1_GLM_train, col = "blue", main = "Model 1: ROC Curve
using Train Data")
plot.roc(ROC_Model_1_GLM_test, col = "blue", main = "Model 1: ROC Curve using
Test Data")
```

# Insurance Customer Risk Analysis Report

---

```
## AUC value of Train and Test data
auc(ROC_Model_1_GLM_train)
auc(ROC_Model_1_GLM_test)

# #-----
##Model_2_GLM
# #-----
Model_2_GLM <-glm(TARGET_FLAG ~ CLM_FREQ + UC_HUU_Ind + IMP_INCOME + M_INCOME
+
                CU_Commercial_Ind + Single_Parent_Ind + REVOKED_Ind +
MVR_PTS +
                TRAVTIME + Ed_Non_Degree_Ind + MSTATUS_Single_Ind + TIF +
                KIDSDRIV + CT_Sports_Car_Ind + CT_SUV_Ind +
                CT_Pickup_Ind + CT_Van_Ind +
                Home_Owner_else_Renter_Ind
                ,family = binomial(link = "logit"),data = train.df )
summary(Model_2_GLM)
sort(vif(Model_2_GLM), decreasing = TRUE)

#ODDS Ratio
Odds_Ratio_Mod_2 <- as.data.frame(exp(Model_2_GLM$coefficients))
names(Odds_Ratio_Mod_2) <- "Odds Ratio"
View(Odds_Ratio_Mod_2)

## ROC Curve and AUC based on Train data set
Model_2_GLM_pred_train <- predict(Model_2_GLM, type = "response")
ROC_Model_2_GLM_train <- roc(train.df$TARGET_FLAG, Model_2_GLM_pred_train )

##test data
Model_2_GLM_pred_test <- predict(Model_2_GLM, newdata = test.df, type =
"response")
ROC_Model_2_GLM_test <- roc(test.df$TARGET_FLAG, Model_2_GLM_pred_test )

#confusion Matrix work
class_pred_Model_2_train <- ifelse(Model_2_GLM_pred_train > cutoff,1,0)
class_pred_Model_2_test <- ifelse(Model_2_GLM_pred_test > cutoff,1,0)

Conf_Matrix_Model_2_train <-
table(train.df$TARGET_FLAG,class_pred_Model_2_train)
Conf_Matrix_Model_2_test <-
table(test.df$TARGET_FLAG,class_pred_Model_2_test)

#compute the classification accuraxy
acc_Model_2_train <- sum(diag(Conf_Matrix_Model_2_train))/nrow(train.df)
acc_Model_2_test <- sum(diag(Conf_Matrix_Model_2_test))/nrow(test.df)

## ROC Curve for Test and Train data set
plot.roc(ROC_Model_2_GLM_train, col = "red", main = "Model 2: ROC Curve using
Train Data")
plot.roc(ROC_Model_2_GLM_test, col = "red", main = "Model 2: ROC Curve using
Test Data")
```

# Insurance Customer Risk Analysis Report

---

```
## AUC value of Train and Test data
auc(ROC_Model_2_GLM_train)
auc(ROC_Model_2_GLM_test)

# #-----
##Model_3_GLM
# #-----

Model_3_GLM <-glm(TARGET_FLAG ~ CLM_FREQ + UC_HUU_Ind +
                  CU_Commercial_Ind + Single_Parent_Ind + REVOKED_Ind +
MVR_PTS +
                  TRAVTIME + Ed_Non_Degree_Ind + MSTATUS_Single_Ind + TIF +
                  KIDSDRIV + CT_Sports_Car_Ind + CT_SUV_Ind +
                  CT_Pickup_Ind + CT_Van_Ind +
                  OLDCLAIM + Home_Owner_else_Renter_Ind
, family = binomial(link = "logit"), data = train.df )
summary(Model_3_GLM)

names(logit_insurance)

#ODDS Ratio
Odds_Ratio_Mod_3 <- as.data.frame(exp(Model_3_GLM$coefficients))
names(Odds_Ratio_Mod_3) <- "Odds Ratio"
View(Odds_Ratio_Mod_3)

## ROC Curve and AUC based on Train data set
Model_3_GLM_pred_train <- predict(Model_3_GLM, type = "response")
ROC_Model_3_GLM_train <- roc(train.df$TARGET_FLAG, Model_3_GLM_pred_train )

##test data
Model_3_GLM_pred_test <- predict(Model_3_GLM, newdata = test.df, type =
"response")
ROC_Model_3_GLM_test <- roc(test.df$TARGET_FLAG, Model_3_GLM_pred_test )

#confusion Matrix work
class_pred_Model_3_train <- ifelse(Model_3_GLM_pred_train > cutoff,1,0)
class_pred_Model_3_test <- ifelse(Model_3_GLM_pred_test > cutoff,1,0)

Conf_Matrix_Model_3_train <-
table(train.df$TARGET_FLAG,class_pred_Model_3_train)
Conf_Matrix_Model_3_test <-
table(test.df$TARGET_FLAG,class_pred_Model_3_test)

#compute the classification accuraxy
acc_Model_3_train <- sum(diag(Conf_Matrix_Model_3_train))/nrow(train.df)
acc_Model_3_test <- sum(diag(Conf_Matrix_Model_3_test))/nrow(test.df)

## ROC Curve for Test and Train data set
plot.roc(ROC_Model_3_GLM_train, col = "dark green", main = "Model 3: ROC
Curve using Train Data")
plot.roc(ROC_Model_3_GLM_test, col = "dark green", main = "Model 3: ROC Curve
using Train Data")

## AUC value of Train and Test data
```

# Insurance Customer Risk Analysis Report

---

```
auc(ROC_Model_3_GLM_train)
auc(ROC_Model_3_GLM_test)

#-----
# ## 4- SELECT MODELS - Predictive Accuracy
#-----
#
summary(Model_1_GLM)
summary(Model_2_GLM)
summary(Model_3_GLM)

AIC(Model_1_GLM)
AIC(Model_2_GLM)
AIC(Model_3_GLM)

## AUC value of Train and Test data
auc(ROC_Model_1_GLM_train)
auc(ROC_Model_1_GLM_test)

auc(ROC_Model_2_GLM_train)
auc(ROC_Model_2_GLM_test)

auc(ROC_Model_3_GLM_train)
auc(ROC_Model_3_GLM_test)

options(scipen = 999)
ks.test(Model_1_GLM_pred_train, train.df$TARGET_FLAG)$statistic
ks.test(Model_1_GLM_pred_test, test.df$TARGET_FLAG)$statistic

ks.test(Model_2_GLM_pred_train, train.df$TARGET_FLAG)$statistic
ks.test(Model_2_GLM_pred_test, test.df$TARGET_FLAG)$statistic

ks.test(Model_3_GLM_pred_train, train.df$TARGET_FLAG)$statistic
ks.test(Model_3_GLM_pred_test, test.df$TARGET_FLAG)$statistic

##Classification Accuracy test

acc_Model_1_train
acc_Model_1_test

acc_Model_2_train
acc_Model_2_test

acc_Model_3_train
acc_Model_3_test

par(mfrow=c(1,2))

## ROC Curve for Test and Train data set
plot.roc(ROC_Model_1_GLM_train, col = "blue", main = "ROC Curve using Train
Data")
```

# Insurance Customer Risk Analysis Report

---

```
lines.roc(ROC_Model_2_GLM_train, col = "red")
lines.roc(ROC_Model_3_GLM_train, col = "green")

plot.roc(ROC_Model_1_GLM_test, col = "blue", main = "ROC Curve using Test
Data")
lines.roc(ROC_Model_2_GLM_test, col = "red")
lines.roc(ROC_Model_3_GLM_test, col = "green")

#
#-----
# ## 5- Linear Regression Model
#-----

names(logit_insurance_LinearModel)

y_amount <- ifelse(logit_insurance_LinearModel$TARGET_AMT > 0,
                   logit_insurance_LinearModel$TARGET_AMT, NA)

logit_insurance_LinearModel$y_amount <- y_amount

Model_4_lm <- lm(y_amount ~ BLUEBOOK +
                + CT_Panel_Truck_Ind + CT_Pickup_Ind + CT_Sports_Car_Ind
                +
                CT_Van_Ind + CT_SUV_Ind + OLDCLAIM
                , data = logit_insurance_LinearModel)

Model_4_lm <- lm(y_amount ~ BLUEBOOK
                , data = logit_insurance_LinearModel)

summary(Model_4_lm)
```

# Insurance Customer Risk Analysis Report

---

## Appendix II: Stand-Alone R Code

```
#-----
# Auto Insurance Customer Risk Analysis Project
# Singh, Gurjeet
# Stand-Alone program
#-----

library(readr)
library(car)
library(fBasics)
library(ggplot2)
library(corrplot)
library(plyr)
library(gmodels)
library(MASS)
library(gridExtra)
library(pROC)

options(scipen = 999)

#-----
## 1 - Importing a Test File and check import
#-----
summary(logit_insurance_test)
str(logit_insurance_test)
colnames(logit_insurance_test)[1] <- "INDEX"

#-----
## 2 - DATA PREPARATION
#-----
#-----
##clean missing values with median values
#-----

summary(logit_insurance_test)
##clean missing values with median values

logit_insurance_test$IMP_AGE <-ifelse(is.na(logit_insurance_test$AGE),
                                     45,
                                     logit_insurance_test$AGE)
logit_insurance_test$M_AGE <-ifelse(is.na(logit_insurance_test$AGE),
                                    1, 0)

logit_insurance_test$IMP_YOJ <-ifelse(is.na(logit_insurance_test$YOJ),
                                     11,
                                     logit_insurance_test$YOJ)
logit_insurance_test$M_YOJ <- ifelse(is.na(logit_insurance_test$YOJ),
                                    1, 0)

logit_insurance_test$IMP_INCOME <-
ifelse(is.na(logit_insurance_test$INCOME) & is.na(logit_insurance_test$JOB),
54000,
       ifelse(is.na(logit_insurance_test$INCOME) & (logit_insurance_test$JOB ==
"Doctor"), 128000,
       ifelse(is.na(logit_insurance_test$INCOME) & (logit_insurance_test$JOB ==
"Lawyer"), 88000,
```

# Insurance Customer Risk Analysis Report

---

```
    ifelse(is.na(logit_insurance_test$INCOME) & (logit_insurance_test$JOB ==
"Manager"), 87000,
    ifelse(is.na(logit_insurance_test$INCOME) & (logit_insurance_test$JOB ==
"Professional"), 76000,
    ifelse(is.na(logit_insurance_test$INCOME) & (logit_insurance_test$JOB ==
"Clerical"), 33000,
    ifelse(is.na(logit_insurance_test$INCOME) & (logit_insurance_test$JOB ==
"z_Blue Collar"), 58000,
    ifelse(is.na(logit_insurance_test$INCOME) & (logit_insurance_test$JOB ==
"Home Maker"), 12000,
    ifelse(is.na(logit_insurance_test$INCOME) & (logit_insurance_test$JOB ==
"Student"), 6300,
    logit_insurance_test$INCOME)))))))))

logit_insurance_test$M_INCOME <- ifelse(is.na(logit_insurance_test$INCOME),
1, 0)

logit_insurance_test$IMP_HOME_VAL <-
ifelse(is.na(logit_insurance_test$HOME_VAL),
162000,
logit_insurance_test$HOME_VAL)

logit_insurance_test$M_HOME_VAL <-
ifelse(is.na(logit_insurance_test$HOME_VAL),
1, 0)

logit_insurance_test$IMP_CAR_AGE <-
ifelse(is.na(logit_insurance_test$CAR_AGE), 8,
ifelse(logit_insurance_test$CAR_AGE
< 0, 0,

logit_insurance_test$CAR_AGE))

logit_insurance_test$M_CAR_AGE <- ifelse(is.na(logit_insurance_test$CAR_AGE),
1,
ifelse(logit_insurance_test$CAR_AGE < 0,
1,
0))

logit_insurance_test$IMP_JOB <- ifelse(is.na(logit_insurance_test$JOB) &
logit_insurance_test$IMP_INCOME > 150000, "Doctor",
ifelse(is.na(logit_insurance_test$JOB) &
logit_insurance_test$IMP_INCOME > 100000, "Lawyer",
ifelse(is.na(logit_insurance_test$JOB) &
logit_insurance_test$IMP_INCOME > 85000, "Manager",
ifelse(is.na(logit_insurance_test$JOB) &
logit_insurance_test$IMP_INCOME > 75000, "Professional",
ifelse(is.na(logit_insurance_test$JOB) &
logit_insurance_test$IMP_INCOME > 60000, "z_Blue Collar",
ifelse(is.na(logit_insurance_test$JOB) &
logit_insurance_test$IMP_INCOME > 35000, "Clerical",
ifelse(is.na(logit_insurance_test$JOB) &
logit_insurance_test$IMP_INCOME >= 12000, "Home Maker",
ifelse(is.na(logit_insurance_test$JOB) &
logit_insurance_test$IMP_INCOME < 12000, "Student",
logit_insurance_test$JOB)))))))))
```

# Insurance Customer Risk Analysis Report

---

```
logit_insurance_test$M_JOB <- ifelse(is.na(logit_insurance_test$JOB),
                                     1, 0)

#-----
##Add Indicator variables
#-----
logit_insurance_test$Single_Parent_Ind <- ifelse(logit_insurance_test$PARENT1
== 'Yes', 1, 0);
logit_insurance_test$MSTATUS_Single_Ind <-
ifelse(logit_insurance_test$MSTATUS == 'Yes', 0, 1);
logit_insurance_test$SEX_Female_Ind <- ifelse(logit_insurance_test$SEX
== 'M', 0, 1);
logit_insurance_test$Ed_Non_Degree_Ind <-
ifelse(logit_insurance_test$EDUCATION == '<High School' |
                                              logit_insurance_test$EDUCATION
== 'z_High School', 1, 0)
logit_insurance_test$CU_Commercial_Ind <- ifelse(logit_insurance_test$CAR_USE
== 'Commercial', 1, 0);

logit_insurance_test$CT_Panel_Truck_Ind <-
ifelse(logit_insurance_test$CAR_TYPE == 'Panel Truck', 1, 0);
logit_insurance_test$CT_Pickup_Ind <- ifelse(logit_insurance_test$CAR_TYPE
== 'Pickup', 1, 0);
logit_insurance_test$CT_Sports_Car_Ind <-
ifelse(logit_insurance_test$CAR_TYPE == 'Sports Car', 1, 0);
logit_insurance_test$CT_Van_Ind <- ifelse(logit_insurance_test$CAR_TYPE
== 'Van', 1, 0);
logit_insurance_test$CT_SUV_Ind <- ifelse(logit_insurance_test$CAR_TYPE
== 'z_SUV', 1, 0);

logit_insurance_test$REVOKED_Ind <- ifelse(logit_insurance_test$REVOKED
== 'Yes', 1, 0);

logit_insurance_test$UC_HUU_Ind <- ifelse(logit_insurance_test$URBANICITY
== 'Highly Urban/ Urban', 1, 0);

logit_insurance_test$JOB_White_Collar_Ind <-
ifelse(logit_insurance_test$IMP_JOB == 'Clerical' |
                                              logit_insurance_test$IMP_JOB
== 'Doctor' |
                                              logit_insurance_test$IMP_JOB
== 'Lawyer' |
                                              logit_insurance_test$IMP_JOB
== 'Manager' |
                                              logit_insurance_test$IMP_JOB
== 'Professional', 1, 0)
logit_insurance_test$JOB_Blue_Collar_Ind <-
ifelse(logit_insurance_test$IMP_JOB == 'z_Blue Collar', 1, 0);
logit_insurance_test$JOB_Student_Ind <- ifelse(logit_insurance_test$IMP_JOB
== 'Student', 1, 0);

logit_insurance_test$Home_Owner_else_Renter_Ind <-
ifelse(logit_insurance_test$IMP_HOME_VAL != 0, 1, 0);
#-----
--
```



# Insurance Customer Risk Analysis Report

---

```
## 3- MODEL Deployment
#-----
#-----
## Exporting Model - Logistic Model
#-----
log_ODDS_TARGET_FLAG <- with(logit_insurance_test, -4.5274604992
+ 0.1319755174 * CLM_FREQ
+ 2.3244910905 * UC_HUU_Ind
- 0.0000068080 * IMP_INCOME
+ 0.0256698028 * M_INCOME
+ 0.8991351539 * CU_Commercial_Ind
+ 0.4860217949 * Single_Parent_Ind
+ 0.6704261814 * REVOKED_Ind
+ 0.1193321951 * MVR_PTS
+ 0.0149330842 * TRAVTIME
+ 0.5013800591 * Ed_Non_Degree_Ind
+ 0.4664214992 * MSTATUS_Single_Ind
- 0.0570364238 * TIF
+ 0.3897772420 * KIDSDRIV
+ 1.0021040089 * CT_Sports_Car_Ind
+ 0.7847554245 * CT_SUV_Ind
+ 0.5595281011 * CT_Pickup_Ind
+ 0.5556656480 * CT_Van_Ind
- 0.3093577902 * Home_Owner_else_Renter_Ind)

ODDS_TARGET_FLAG <- exp(log_ODDS_TARGET_FLAG)
P_TARGET_FLAG <- ODDS_TARGET_FLAG/(1 + ODDS_TARGET_FLAG)

logit_insurance_test$P_TARGET_FLAG <- P_TARGET_FLAG

#-----
## Exporting Model - Linear Model
#-----

P_TARGET_AMT <- with(logit_insurance_test, 4131.65436
+ 0.11017 * BLUEBOOK)

#-----
## Creating Scoring Ouput file
#-----

FINAL_Submission <- with(logit_insurance_test,
cbind.data.frame(INDEX,
round(P_TARGET_FLAG,2),
round(P_TARGET_AMT,2)))

colnames(FINAL_Submission) <- c("INDEX", "P_TARGET_FLAG", "P_TARGET_AMT")
write.csv(FINAL_Submission, "Singh_Gurjeet_Insurance_Test_Score.csv")
```