Insurance Logistic Regression

Introduction

In this document we explore logistic techniques to estimate the probability that a person will crash their car and a model to estimate the cost in the event of the crash. We will be using the Insurance data set to build logistic regression models to predict car crashes. This data set contains approximately 8100 records. Each record represents a customer at an auto insurance company and 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 zero means that the person was not in a car crash. 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. In order to create a model that most accurately predicts the number of car crashes, we will prepare the data using imputation, capping, and binning, and we will select the variables using stepwise automated variable selection. We will present several models, discuss the merits and shortcomings of each, and ultimately select one model that best predicts which customers are most likely to crash their cars.

Data Exploration

The dataset contains approximately 8161 records with 23 variables with the combination of categorical and the continuous variables. Dataset also contains 2 target variables and an index variable. First we will focus on predicting the TARGET_FLAG variable which signifies whether or not customer crashed his/her car. Based on the value of TARGET_FLAG then we calculate the TARGET_AMT which is a second target variable and signifies the cost of the crash.

Table1 shows the list of variables included in the dataset and proposed effect of each variable on the response variable is also shown in the table. Dictionary seems to be the right place to start as it has been prepared and the data matches with the description of these variables. Data contains both the categorical and continuous variables. It also makes sense to validate some of these urban legends and the theories that certain variables can have more effect on the response variable. It will also help to consider some of these proposed effects when we impute the data. Some of these variables may be highly correlated with each other or some sort of relationship between them. One variable of such that is the INCOME and the HOME_VAL as people with high income tend to have larger or expensive homes.

Table 1: Data Dictionary with Theoretical Effect

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	Cost of the crash	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

Next we will look at the descriptive statistics of the dataset. Some of the variables are showing high scales compared to most of the other variables in the dataset, so we may have to transform these variables during the preparation phase. All the variables in the dataset are continuous.

Table 2: Data Statistics

	count	mean	std	min	25%	50%	75%	max
index	8161.0	5151.867663	2978.893962	1.0	2559.0	5133.0	7745.0	10302.00000
target_flag	8161.0	0.263816	0.440728	0.0	0.0	0.0	1.0	1.00000
target_amt	8161.0	1504.324648	4704.026930	0.0	0.0	0.0	1036.0	107586.13616
kidsdriv	8161.0	0.171057	0.511534	0.0	0.0	0.0	0.0	4.00000
age	8155.0	44.790313	8.627589	16.0	39.0	45.0	51.0	81.00000
homekids	8161.0	0.721235	1.116323	0.0	0.0	0.0	1.0	5.00000
yoj	7707.0	10.499286	4.092474	0.0	9.0	11.0	13.0	23.00000
travtime	8161.0	33.485725	15.908333	5.0	22.0	33.0	44.0	142.00000
tif	8161.0	5.351305	4.146635	1.0	1.0	4.0	7.0	25.00000
clm_freq	8161.0	0.798554	1.158453	0.0	0.0	0.0	2.0	5.00000
mvr_pts	8161.0	1.695503	2.147112	0.0	0.0	1.0	3.0	13.00000
car_age	7651.0	8.328323	5.700742	-3.0	1.0	8.0	12.0	28.00000

Table 3 shows the summary of the missing values in the training dataset. Six of the variables, AGE, YOJ, INCOME, HOME_VAL, JOB and CAR_AGE, have missing data. We will create two new variables in this process; one with the IMP_* prefix for the imputed variable, leaving the original variable untouched, and one with the m_* prefix as an indicator variable for the imputed variable. Sometimes, the fact that a variable was missing can actually be predictive. This means the indicator variables might be entered into the predictive model. JOB variable is a categorical variable and has the missing data, instead of creating a new variable, we will replace the missing values with "*Missing Job Info*" new category. Remaining all other variables will be imputed with their median value as it less prone to the outliers.

Table 3: Missing Values

index	0
target_flag	0
target_amt	0
kidsdriv	0
age	6
homekids	0
yoj	454
income	445
parent1	0
home_val	464
mstatus	0
sex	0
education	0
job	526
travtime	0
car_use	0
bluebook	0
tif	0
car_type	0
red_car	0
oldclaim	0
clm_freq	0
revoked	0
mvr_pts	0
car_age	510
urbanicity	0
dtype: int64	

Next I have created the correlation matrix with the strongest correlated variables, shown in Figure 1. It is interesting to observe that MVR_PTS and CLM_FREQ higher correlation and it makes sense as the customer with high number of traffic tickets prone to more crashes or claims. Same applies to the HOMEKIDS and KIDSDRIV, the more kids are at home, the possibility of kids driving the vehicle is high. Rest of the continuous variables seems to be fairly independent and that further proves that this data might prepared well before.

Figure 1: Variables with Strongest Correlation Matrix

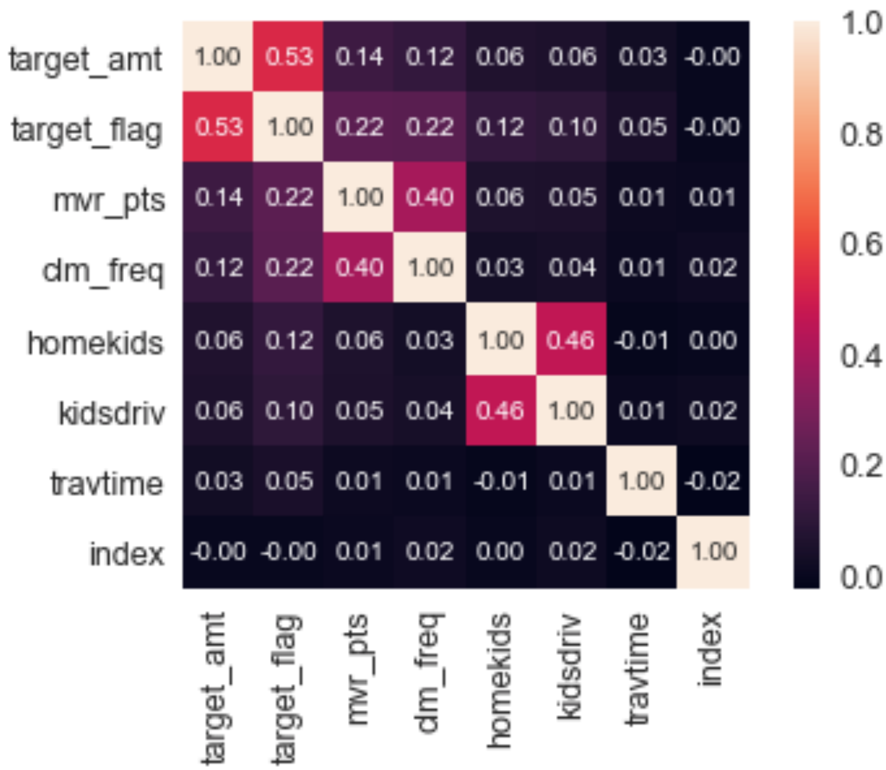
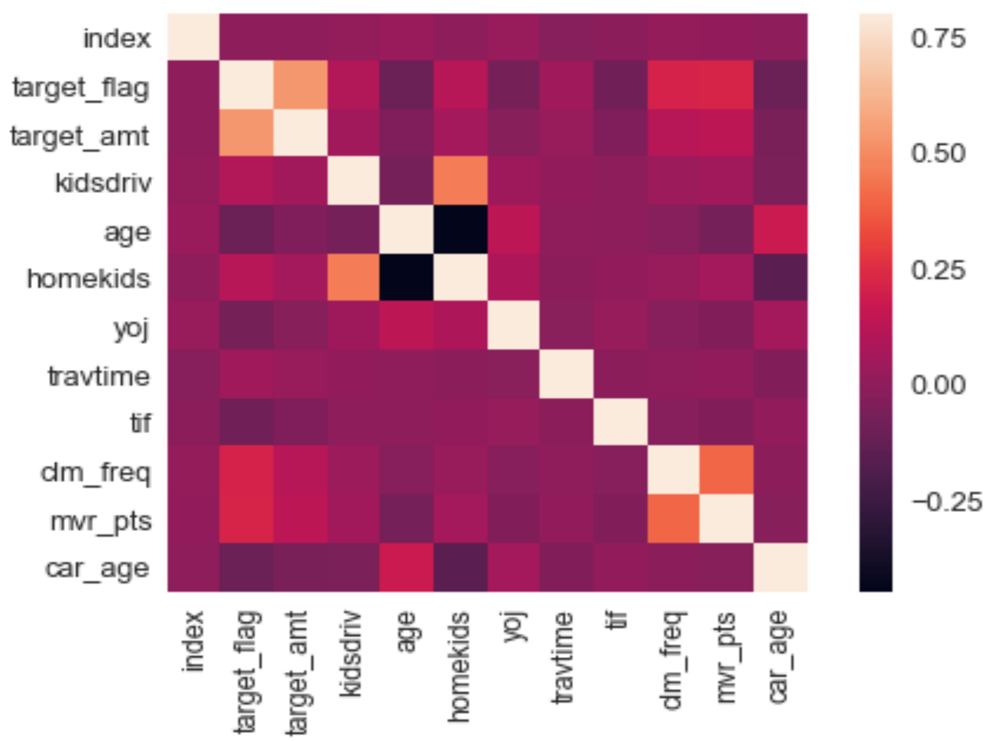
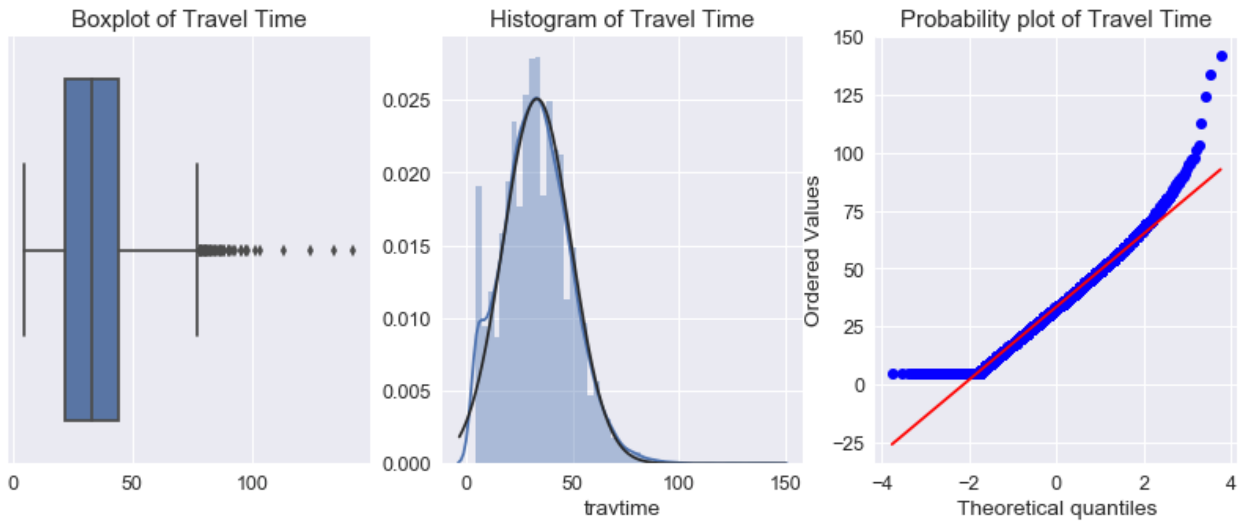


Figure 2: Correlation Matrix



So next, I have started with the continuous variables first. Some of these variables have an extreme values and resulting highly skewed distribution. Figure 3 shows the distribution of TRAVTIME for the customer. For this variable data seems to be fairly close to the normal distribution and the number of outliers that will imputed are shown in the figure as well. TRAVTIME had a reported commute time of over two hours and twenty minutes, which could be a mistake, but is otherwise skewing the data.

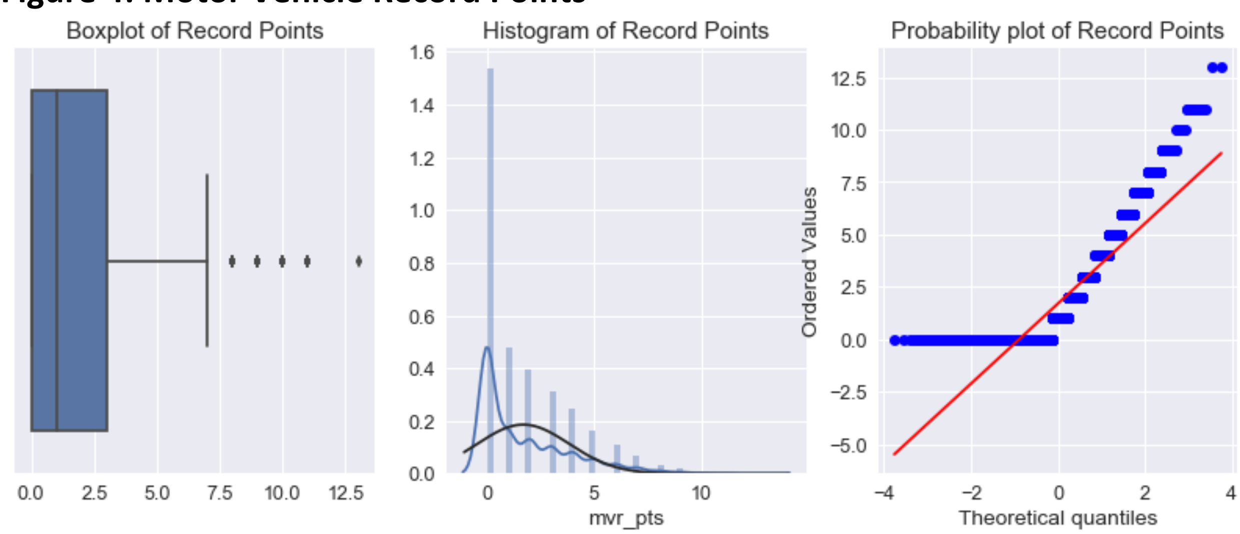
Figure 3: Travel Time




```
Outliers for Travel Time:  [113, 124, 142, 134]
```

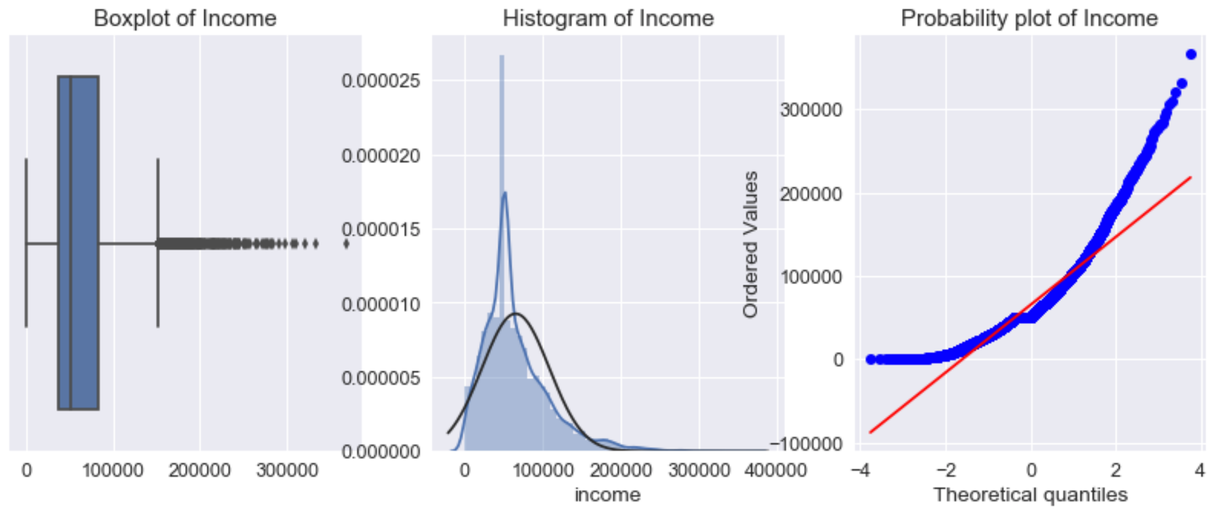
Figure 4 shows the distribution and the outliers for the motor vehicle record points. Some records seem to have about 13 points and that is even though quite high, is skewing the data.

Figure 4: Motor Vehicle Record Points



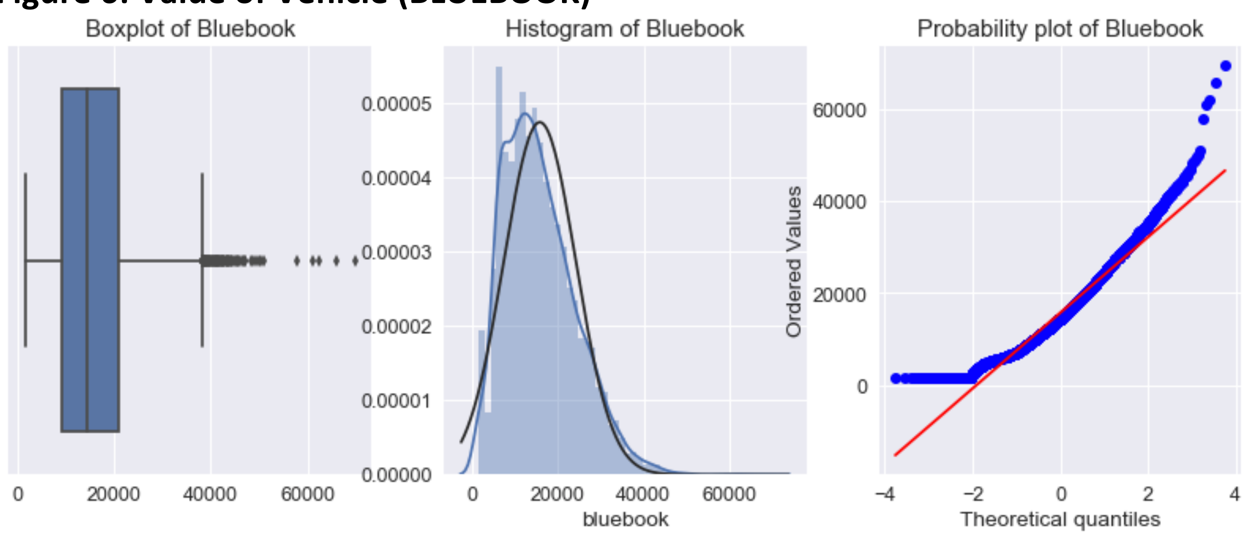

```
Outliers for Motor Vehicle Record Points:  [13, 13]
```

Figure 5: Income



Next, I have looked at the distribution for the INCOME. This one looks quite a bit like HOME_VAL. In case of INCOME, it has extreme values in the right tail. There are few income groups going up to 500,000. Figure 5 shows the outliers that are in the extreme group. Figure 6 shows the distribution and the outliers for the BLUEBOOK. As the figure shows there are some vehicles with value more than 60,000. It is possible that these records are correlated with the records that are in the high income group. People with high income might be more likely to buy expensive cars.

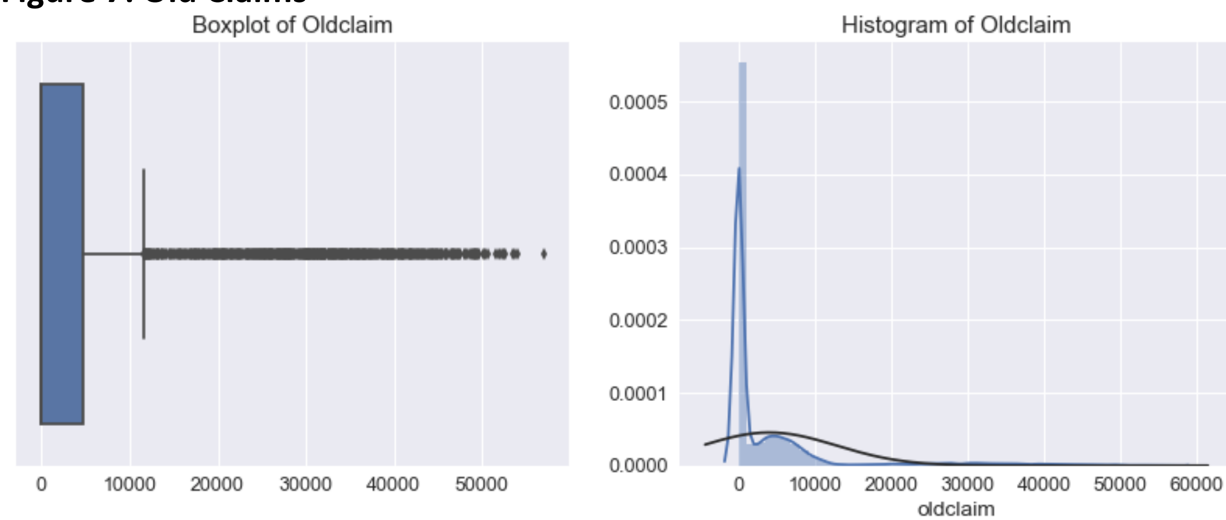
Figure 6: Value of Vehicle (BLUEBOOK)




```
Outliers for Bluebook:  [65970, 62240, 57970, 61050, 69740]
```

Next, I have looked at the OLDCLAIM as part of the final continuous variable. As shown in Figure 7, this variable has the quite interesting distribution compared to the others. There are many records with 0 claims meaning there are many people did not have past claims. Also data seems to be distributed into three different categories. First category is people without any old claims, second category is the people that are between 0 to 11000. And the final category is where people with claim that is worth more than the 11000. And the last category seems to go up all the way to 60000. We will consider this fact when we do the data extraction.

Figure 7: Old Claims

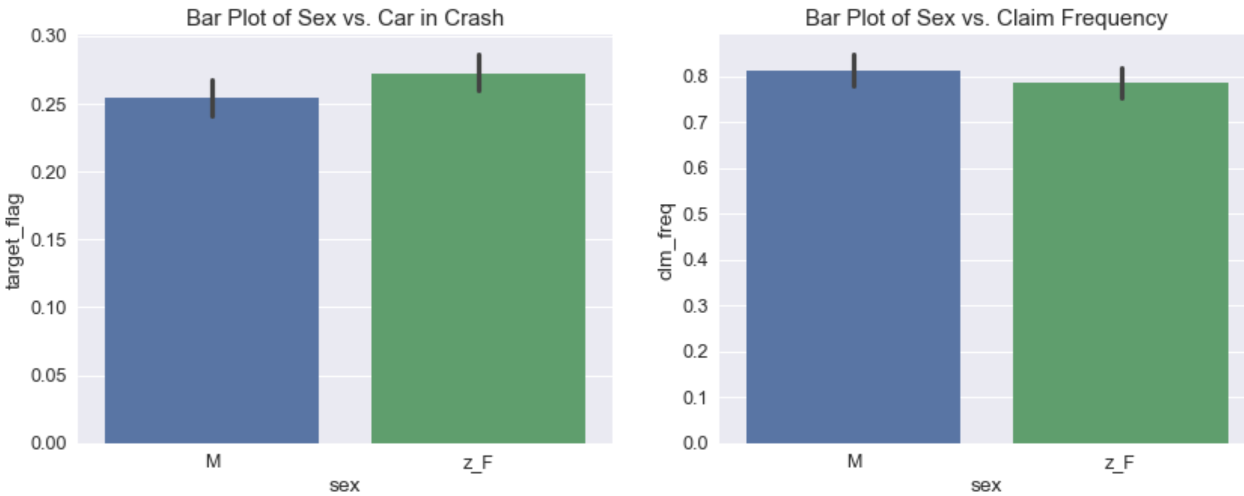


For categorical, first I have looked at the Red Car vs. car in the crash and the claim frequency. From the Figure 8, it is evident that there is no effect of the car color in number of crashes reported. Although there seems to be a slight increase in the claims for red color car, but that may be just due to the number of red color cars within the dataset.

Figure 8: Red Car

Figure 9 shows the comparison for Sex vs. car in crash and the claim frequency. As shown in the figure females may have more crashes reported, but the males seem to be high in the claim frequency.

Figure 9: Male vs. Female



Next in the categorical variable is driving children in a household. From the figure 10, it shows that as the number of kids increase in the household, the number of crashes seem to be increasing. However, claim frequency seems to be evenly distributed between the families without kids and the families with the kids.

Figure 10: Driving Children

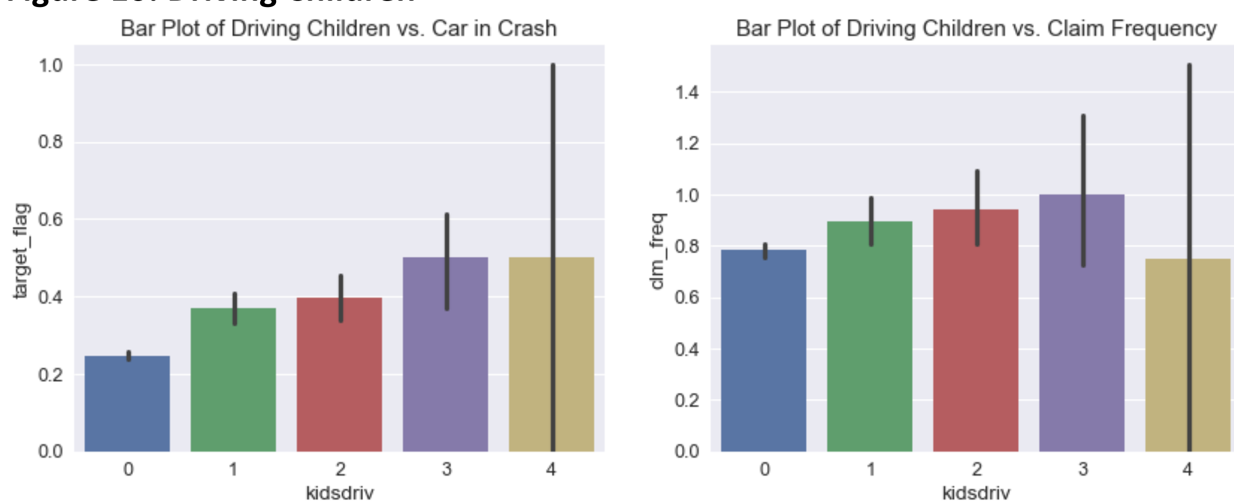
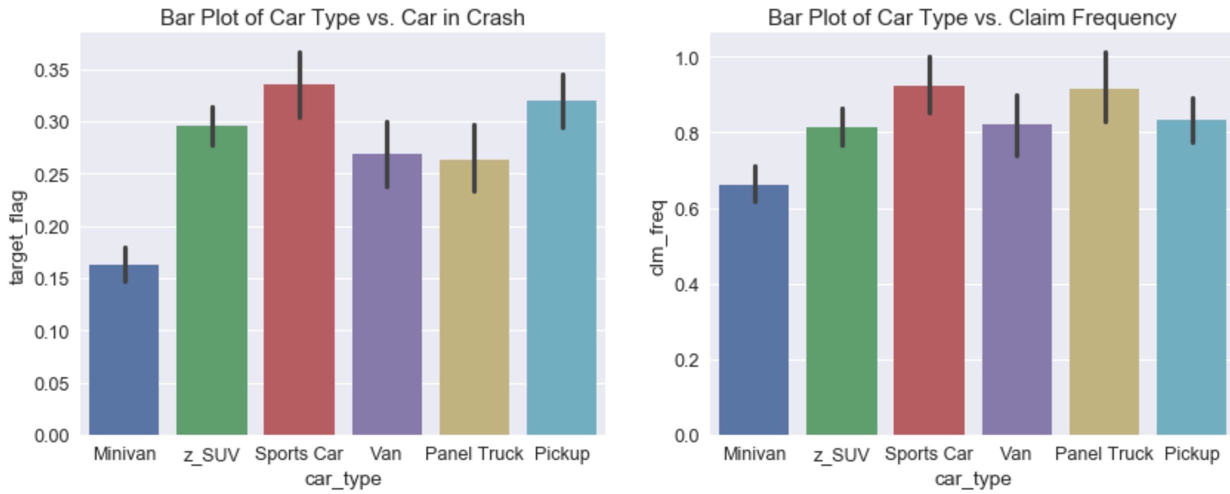


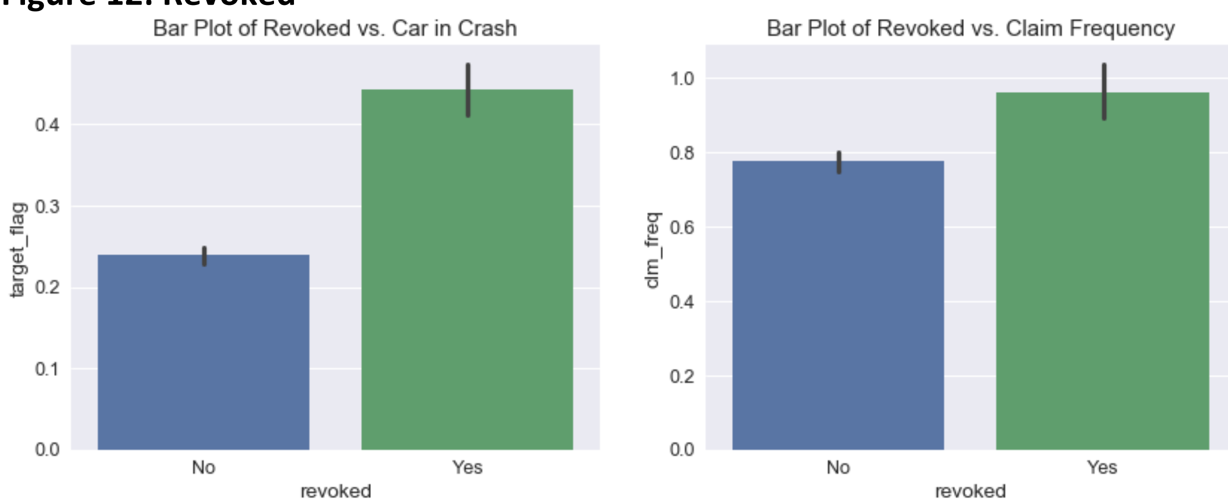
Figure 11 shows the comparison for the car types. From the chart it is evident that the sports car stands out for the number of crashes and followed by the pickup and the SUV. Claim frequency seems to even out across all the car types.

Figure 11: Car types



Finally, Figure 12 shows the comparison for the Revoked variable. From the chart it is clear that the people with revoked license seems to be more involved in the crash and the they had high claim frequency than the other group of people. It definitely makes sense to consider this variable when we fit the models.

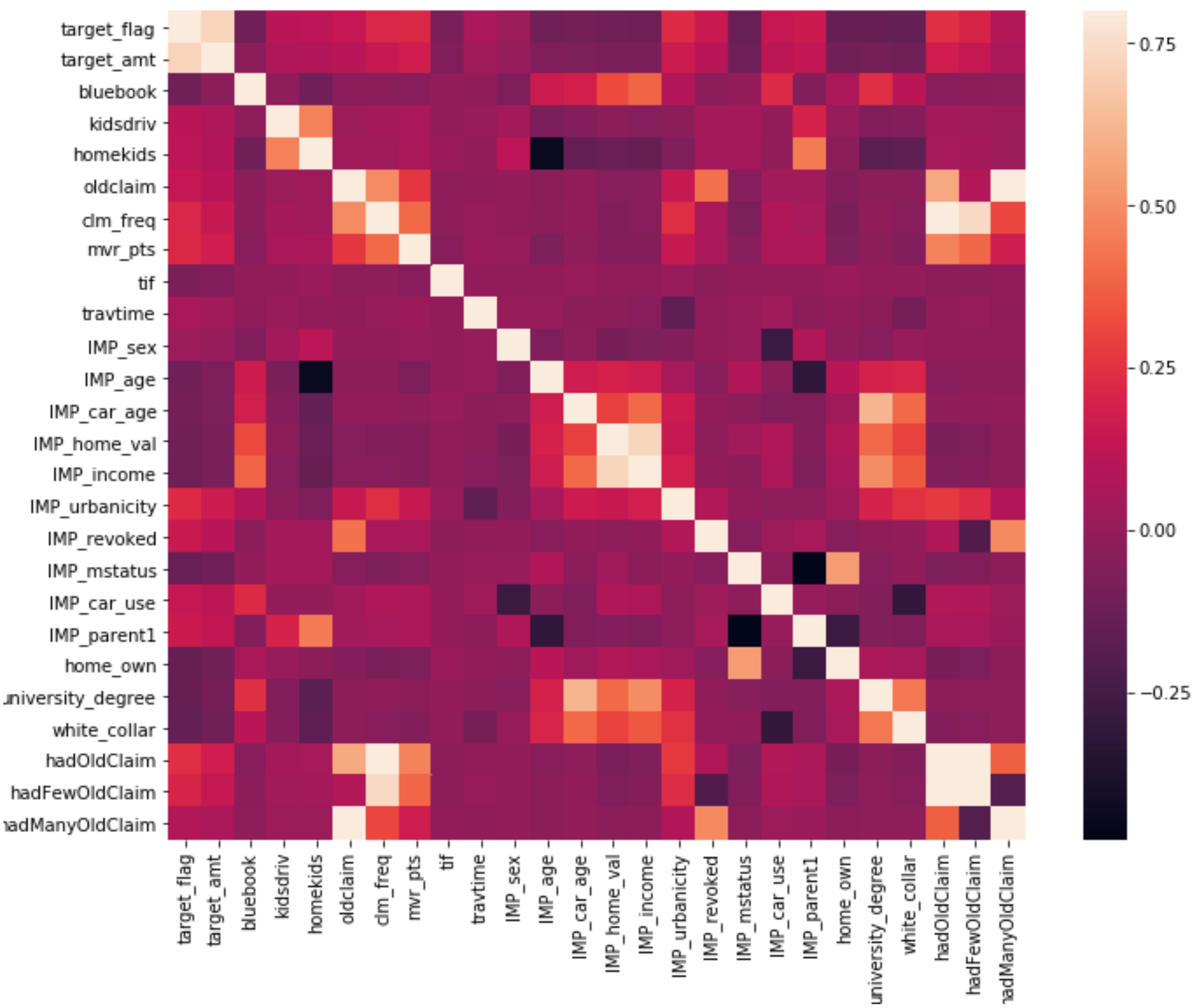
Figure 12: Revoked



Data Preparation

As part of the data preparation, first thing I have performed is to fix the missing values. I have decided to use the median value to replace all the missing values for each variable. For all the outliers, instead of dropping them I chose to use the truncating strategy based off of quantiles. For the variables listed above, if any values exceeded the 99th percentile, then they were replaced with the value of the 99th percentile. Likewise, for values less than the 1th percentile. Instead of replacing the data in the existing variables, I chose to create new variables with IMP_* for the imputed value and m_* to represent the existence of the data. Finally, with missing values imputed and outliers mostly fixed, it was important to remember to do these same actions with the test data. As such, I imputed missing data with the medians from the training data and truncating the variables using the original 99th and 1th percentiles of the same variables from the training set.

Figure 13: Correlation Matrix with Imputed data



As discussed above, I have binned OLDCLAIM into three different buckets. First bucket representing people without any claims, second bucket representing with claims from 0 to 11000 and the last bin containing the claims more than 11000.

Models

I have created several different models with different features set. However, I am listing the three models that seems to be interesting in terms of their performance.

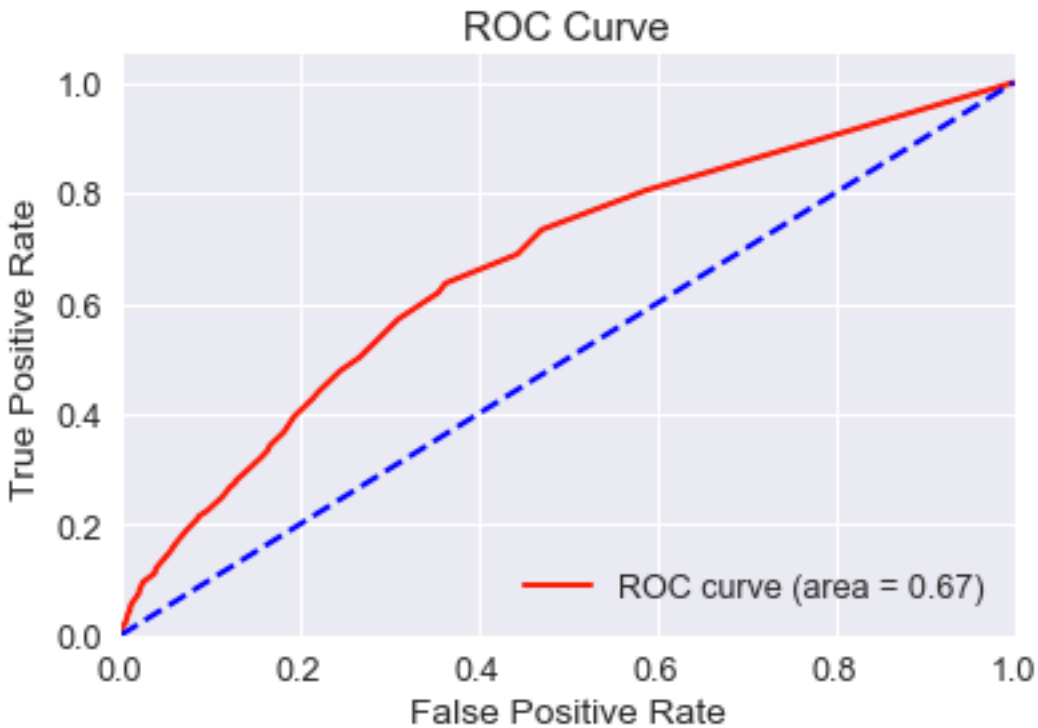
Model 1: clm_freq & mvr_pts

The first logistic regression I have created is solely using two features, claim frequency and the motor vehicle record points. The results from the logistic regression are shown below. Although this model considers only two features, it is performing quite well. ROC curve area of 0.67 is also pretty decent.

Table 4: Model 1 Summary

```
                           Logit Regression Results
==============================================================================
Dep. Variable:            target_flag   No. Observations:                 8161
Model:                          Logit   Df Residuals:                     8158
Method:                           MLE   Df Model:                            2
Date:                Sat, 17 Feb 2018   Pseudo R-squ.:                 0.05582
Time:                        01:16:27   Log-Likelihood:                -4446.1
converged:                       True   LL-Null:                       -4709.0
                                        LLR p-value:                 6.853e-115
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -1.5849      0.038    -42.165      0.000      -1.659      -1.511
clm_freq       0.2865      0.023     12.695      0.000       0.242       0.331
mvr_pts        0.1565      0.012     12.880      0.000       0.133       0.180
==============================================================================
                          Results: Logit
==================================================================
Model:              Logit            Pseudo R-squared: 0.056
Dependent Variable: target_flag      AIC:              8898.2170
Date:               2018-02-17 01:16 BIC:              8919.2384
No. Observations:   8161             Log-Likelihood:   -4446.1
Df Model:           2                LL-Null:          -4709.0
Df Residuals:       8158             LLR p-value:      6.8527e-115
Converged:          1.0000           Scale:            1.0000
No. Iterations:     5.0000
-------------------------------------------------------------------
              Coef.  Std.Err.    z     P>|z|   [0.025   0.975]
-------------------------------------------------------------------
Intercept   -1.5849   0.0376  -42.1654 0.0000  -1.6586  -1.5113
clm_freq     0.2865   0.0226   12.6953 0.0000   0.2423   0.3308
mvr_pts      0.1565   0.0122   12.8800 0.0000   0.1327   0.1804
==================================================================
```

Figure 14: Model 1 ROC Curve



Model 2: Top Five

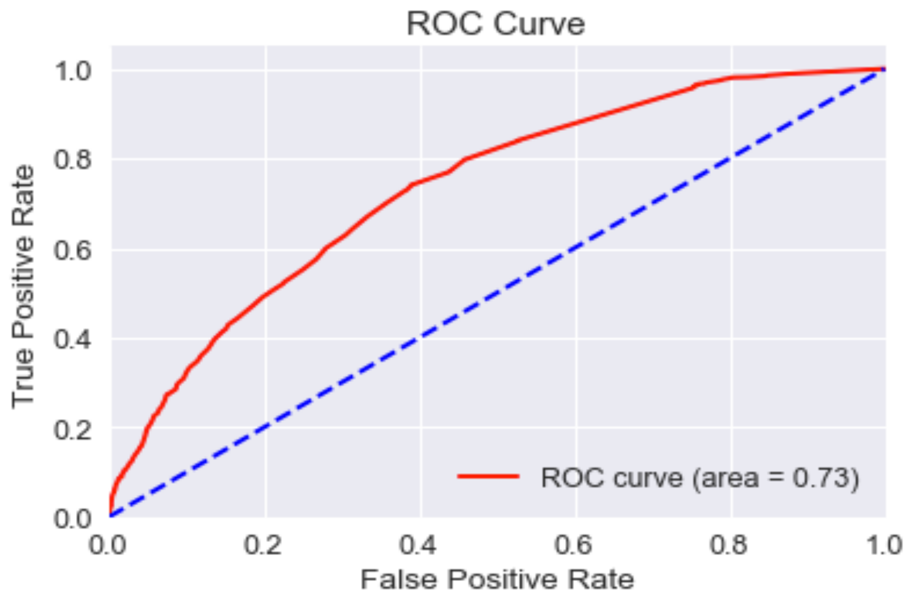
For the next model, I have considered the top five highly correlated variables with the response variable. The variables include, Claim Frequency, Motor Vehicle Record Points, Urban or City, Revoked and Single Parent. These are selected because they needed very little imputation. This model performed much better than the previous model. ROC curve area spiked to 0.73 and the AIC & BIC values are improved as well and the Pseudo R-square value is fairly low. All the five variables are statistically significant. The coefficient values match the expected theoretical effect and all the selected features should increase one's probability of getting into crash. Model 2 results and the ROC curve is listed below.

Table 5: Model 2 Summary

```
                           Logit Regression Results
==============================================================================
Dep. Variable:            target_flag   No. Observations:                 8161
Model:                          Logit   Df Residuals:                     8155
Method:                           MLE   Df Model:                            5
Date:                Sat, 17 Feb 2018   Pseudo R-squ.:                  0.1224
Time:                        01:16:27   Log-Likelihood:                -4132.8
converged:                       True   LL-Null:                       -4709.0
                                        LLR p-value:                6.228e-247
==================================================================================
                     coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------
Intercept         -3.1590      0.103    -30.547      0.000      -3.362      -2.956
clm_freq           0.2115      0.024      8.976      0.000       0.165       0.258
mvr_pts            0.1390      0.013     11.045      0.000       0.114       0.164
IMP_urbanicity     1.6238      0.104     15.620      0.000       1.420       1.828
IMP_revoked        0.7919      0.074     10.711      0.000       0.647       0.937
IMP_parent1        0.9764      0.074     13.253      0.000       0.832       1.121
==================================================================================
```

```
                         Results: Logit
=================================================================
Model:              Logit            Pseudo R-squared: 0.122
Dependent Variable: target_flag      AIC:              8277.6384
Date:               2018-02-17 01:16 BIC:              8319.6812
No. Observations:   8161             Log-Likelihood:   -4132.8
Df Model:           5                LL-Null:          -4709.0
Df Residuals:       8155             LLR p-value:      6.2282e-247
Converged:          1.0000           Scale:            1.0000
No. Iterations:     7.0000
-----------------------------------------------------------------
                   Coef.  Std.Err.    z     P>|z|   [0.025 0.975]
-----------------------------------------------------------------
Intercept         -3.1590   0.1034 -30.5474 0.0000 -3.3616 -2.9563
clm_freq           0.2115   0.0236   8.9761 0.0000  0.1653  0.2577
mvr_pts            0.1390   0.0126  11.0449 0.0000  0.1143  0.1637
IMP_urbanicity     1.6238   0.1040  15.6196 0.0000  1.4200  1.8275
IMP_revoked        0.7919   0.0739  10.7107 0.0000  0.6470  0.9368
IMP_parent1        0.9764   0.0737  13.2532 0.0000  0.8320  1.1208
-----------------------------------------------------------------
```

Figure 15: Model 2 ROC Curve



Model 3: Statistically Significant

For the last model, I have considered all the imputed and original variables that are statistically significant at alpha level 0.5. Both the bluebook and imputed home value are log transformed. Out of all the models, this model performed best and this has very minimal multicollinearity problem. This model ROC curve area scored little above 0.81 and both the AIC & BIC values are decreased.

Figure 16: Model 3 ROC Curve

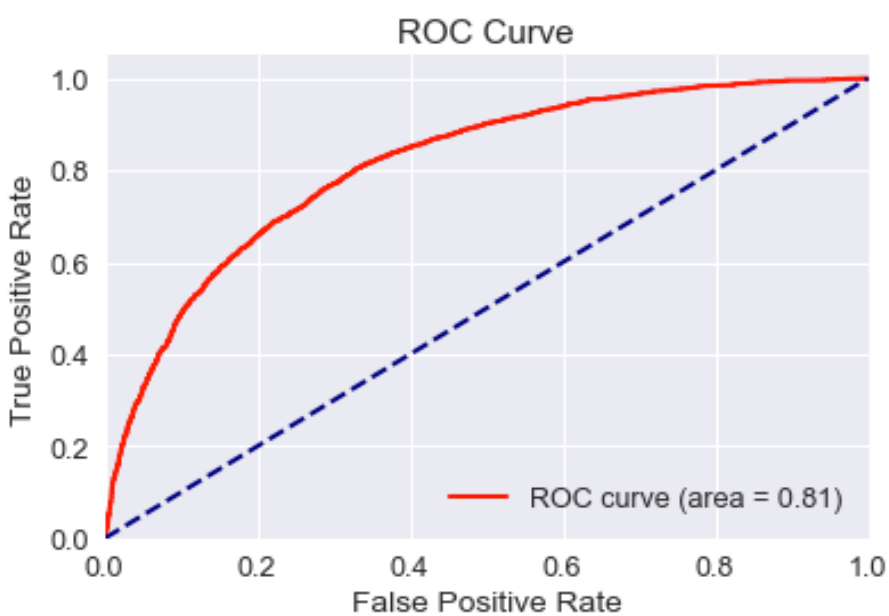


Table 6: Model 3 Logit Summary

Results: Logit

Model:	Logit	Pseudo R-squared:	0.223
Dependent Variable:	target_flag	AIC:	7358.1882
Date:	2018-02-17 01:16	BIC:	7505.3378
No. Observations:	8161	Log-Likelihood:	-3658.1
Df Model:	20	LL-Null:	-4709.0
Df Residuals:	8140	LLR p-value:	0.0000
Converged:	1.0000	Scale:	1.0000
No. Iterations:	7.0000		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Intercept	4.9605	1.2350	4.0164	0.0001	2.5398	7.3811
car_type[T.Panel Truck]	0.3888	0.1332	2.9183	0.0035	0.1277	0.6499
car_type[T.Pickup]	0.5104	0.0975	5.2351	0.0000	0.3193	0.7015
car_type[T.Sports Car]	0.9364	0.1065	8.7923	0.0000	0.7276	1.1451
car_type[T.Van]	0.5870	0.1194	4.9160	0.0000	0.3529	0.8210
car_type[T.z_SUV]	0.7101	0.0849	8.3610	0.0000	0.5437	0.8766
kidsdriv	0.4133	0.0547	7.5549	0.0000	0.3061	0.5206
mvr_pts	0.0981	0.0136	7.2272	0.0000	0.0715	0.1248
tif	-0.0541	0.0073	-7.3996	0.0000	-0.0684	-0.0398
travtime	0.0154	0.0019	8.0223	0.0000	0.0116	0.0191
IMP_urbanicity	2.2918	0.1124	20.3932	0.0000	2.0715	2.5121
IMP_revoked	0.9811	0.0845	11.6148	0.0000	0.8156	1.1467
IMP_mstatus	-0.4562	0.0780	-5.8497	0.0000	-0.6091	-0.3034
IMP_car_use	0.7106	0.0730	9.7347	0.0000	0.5675	0.8537
IMP_parent1	0.4627	0.0933	4.9572	0.0000	0.2797	0.6456
home_own	-0.3154	0.0712	-4.4291	0.0000	-0.4549	-0.1758
university_degree	-0.5015	0.0676	-7.4209	0.0000	-0.6339	-0.3690
white_collar	-0.5142	0.0845	-6.0867	0.0000	-0.6798	-0.3486
hadFewOldClaim	0.5720	0.0678	8.4332	0.0000	0.4391	0.7049
log_bluebook	-0.3534	0.0539	-6.5592	0.0000	-0.4589	-0.2478
log_IMP_home_val	-0.4518	0.0977	-4.6255	0.0000	-0.6433	-0.2604

Table 7: Model 3 Regression Summary

```
                           Logit Regression Results
==============================================================================
Dep. Variable:            target_flag   No. Observations:                 8161
Model:                          Logit   Df Residuals:                     8140
Method:                           MLE   Df Model:                           20
Date:                Sat, 17 Feb 2018   Pseudo R-squ.:                  0.2232
Time:                        01:16:30   Log-Likelihood:                -3658.1
converged:                       True   LL-Null:                       -4709.0
                                        LLR p-value:                     0.000
============================================================================================
                               coef    std err          z      P>|z|      [0.025      0.975]
--------------------------------------------------------------------------------------------
Intercept                    4.9605      1.235      4.016      0.000       2.540       7.381
car_type[T.Panel Truck]      0.3888      0.133      2.918      0.004       0.128       0.650
car_type[T.Pickup]           0.5104      0.098      5.235      0.000       0.319       0.702
car_type[T.Sports Car]       0.9364      0.106      8.792      0.000       0.728       1.145
car_type[T.Van]              0.5870      0.119      4.916      0.000       0.353       0.821
car_type[T.z_SUV]            0.7101      0.085      8.361      0.000       0.544       0.877
kidsdriv                     0.4133      0.055      7.555      0.000       0.306       0.521
mvr_pts                      0.0981      0.014      7.227      0.000       0.072       0.125
tif                         -0.0541      0.007     -7.400      0.000      -0.068      -0.040
travtime                     0.0154      0.002      8.022      0.000       0.012       0.019
IMP_urbanicity               2.2918      0.112     20.393      0.000       2.072       2.512
IMP_revoked                  0.9811      0.084     11.615      0.000       0.816       1.147
IMP_mstatus                 -0.4562      0.078     -5.850      0.000      -0.609      -0.303
IMP_car_use                  0.7106      0.073      9.735      0.000       0.568       0.854
IMP_parent1                  0.4627      0.093      4.957      0.000       0.280       0.646
home_own                    -0.3154      0.071     -4.429      0.000      -0.455      -0.176
university_degree           -0.5015      0.068     -7.421      0.000      -0.634      -0.369
white_collar                -0.5142      0.084     -6.087      0.000      -0.680      -0.349
hadFewOldClaim               0.5720      0.068      8.433      0.000       0.439       0.705
log_bluebook                -0.3534      0.054     -6.559      0.000      -0.459      -0.248
log_IMP_home_val            -0.4518      0.098     -4.626      0.000      -0.643      -0.260
--------------------------------------------------------------------------------------------
```

Model Selection

The following table shows the comparison for all the models that are described above. Based on all the three models summary, I chose the model 3 as the best performing one based on its AIC and BIC values and also based on the ROC curve area value.

Table 8: Model Comparison

	Model 1	Model 2	Model 3
ROC	0.67	0.73	0.81
AIC	8898	8277	7358
BIC	8919	8319	7505
Pseudo R-Squared	0.056	0.122	0.223

```
                              Results: Logit
========================================================================
Model:                  Logit              Pseudo R-squared:   0.223
Dependent Variable:     target_flag        AIC:                7358.1882
Date:                   2018-02-17 01:16   BIC:                7505.3378
No. Observations:       8161               Log-Likelihood:     -3658.1
Df Model:               20                 LL-Null:            -4709.0
Df Residuals:           8140               LLR p-value:        0.0000
Converged:              1.0000             Scale:              1.0000
No. Iterations:         7.0000
------------------------------------------------------------------------
                         Coef.  Std.Err.    z     P>|z|   [0.025  0.975]
------------------------------------------------------------------------
Intercept                4.9605   1.2350  4.0164 0.0001  2.5398  7.3811
car_type[T.Panel Truck]  0.3888   0.1332  2.9183 0.0035  0.1277  0.6499
car_type[T.Pickup]       0.5104   0.0975  5.2351 0.0000  0.3193  0.7015
car_type[T.Sports Car]   0.9364   0.1065  8.7923 0.0000  0.7276  1.1451
car_type[T.Van]          0.5870   0.1194  4.9160 0.0000  0.3529  0.8210
car_type[T.z_SUV]        0.7101   0.0849  8.3610 0.0000  0.5437  0.8766
kidsdriv                 0.4133   0.0547  7.5549 0.0000  0.3061  0.5206
mvr_pts                  0.0981   0.0136  7.2272 0.0000  0.0715  0.1248
tif                     -0.0541   0.0073 -7.3996 0.0000 -0.0684 -0.0398
travtime                 0.0154   0.0019  8.0223 0.0000  0.0116  0.0191
IMP_urbanicity           2.2918   0.1124 20.3932 0.0000  2.0715  2.5121
IMP_revoked              0.9811   0.0845 11.6148 0.0000  0.8156  1.1467
IMP_mstatus             -0.4562   0.0780 -5.8497 0.0000 -0.6091 -0.3034
IMP_car_use              0.7106   0.0730  9.7347 0.0000  0.5675  0.8537
IMP_parent1              0.4627   0.0933  4.9572 0.0000  0.2797  0.6456
home_own                -0.3154   0.0712 -4.4291 0.0000 -0.4549 -0.1758
university_degree       -0.5015   0.0676 -7.4209 0.0000 -0.6339 -0.3690
white_collar            -0.5142   0.0845 -6.0867 0.0000 -0.6798 -0.3486
hadFewOldClaim           0.5720   0.0678  8.4332 0.0000  0.4391  0.7049
log_bluebook            -0.3534   0.0539 -6.5592 0.0000 -0.4589 -0.2478
log_IMP_home_val        -0.4518   0.0977 -4.6255 0.0000 -0.6433 -0.2604
------------------------------------------------------------------------
```

Model Explanation (P_TARGET_FLAG)

In this section we will talk about the above selected model for predicting P_TARGET_FLAG. Most of the variables chosen for this model based on their statistical significance at alpha 0.05. As explained during the EDA, each car type had some sort of effect on the car crash and the claim frequency. Hence each type has been considered separately in the model building. Also five variables have been created by imputing the value of Single Parent, Car use, marital status, and the revoked status. For all the missing data, median has been used and for the missing job information, new category “Missing Job Info” has been created. And then rest of job information has been split into different categories. IMP_urbancity represents whether the person lives in the urban or not. IMP_revoked indicates whether the person's license has been revoked or not. Also for the old claims, I have taken only the one that had few claims.

P_TARGET_FLAG = 4.96 + 0.39 * car_type[T.Panel Truck] + 0.51 * car_type[T.Pickup] + 0.94 * car_type[T.Sports Car] + 0.59 * car_type[T.Van] + 0.71 * car_type[T.z_SUV] + 0.41 * kidsdriv + 0.09 * mvr_pts – 0.05 * tif + 0.01 * travtime + 2.29 * IMP_urbanicity + 0.98 * IMP_revoked – 0.45 * IMP_mstatus + 0.71 * IMP_car_use + 0.46 * IMP_parent1 – 0.31 * home_own – 0.5 * university_degree – 0.51 * white_collar + 0. 57 * hadFewOldClaim – 0.35 * log_bluebook – 0.45 * log_IMP_home_val

Model Explanation (P_TARGET_AMT)

In this model, we predict the amount it costs if the person gets into the crash. The target amount 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. I have created a simple OLS regression model, instead of using the mean to predict this amount. Based on the correlation between the features and the target variable, I chose Blue book, marital status and the motor vehicle record points. Based on this, the following table shows the results of the OLS model.

Table 8: Target Amount Model

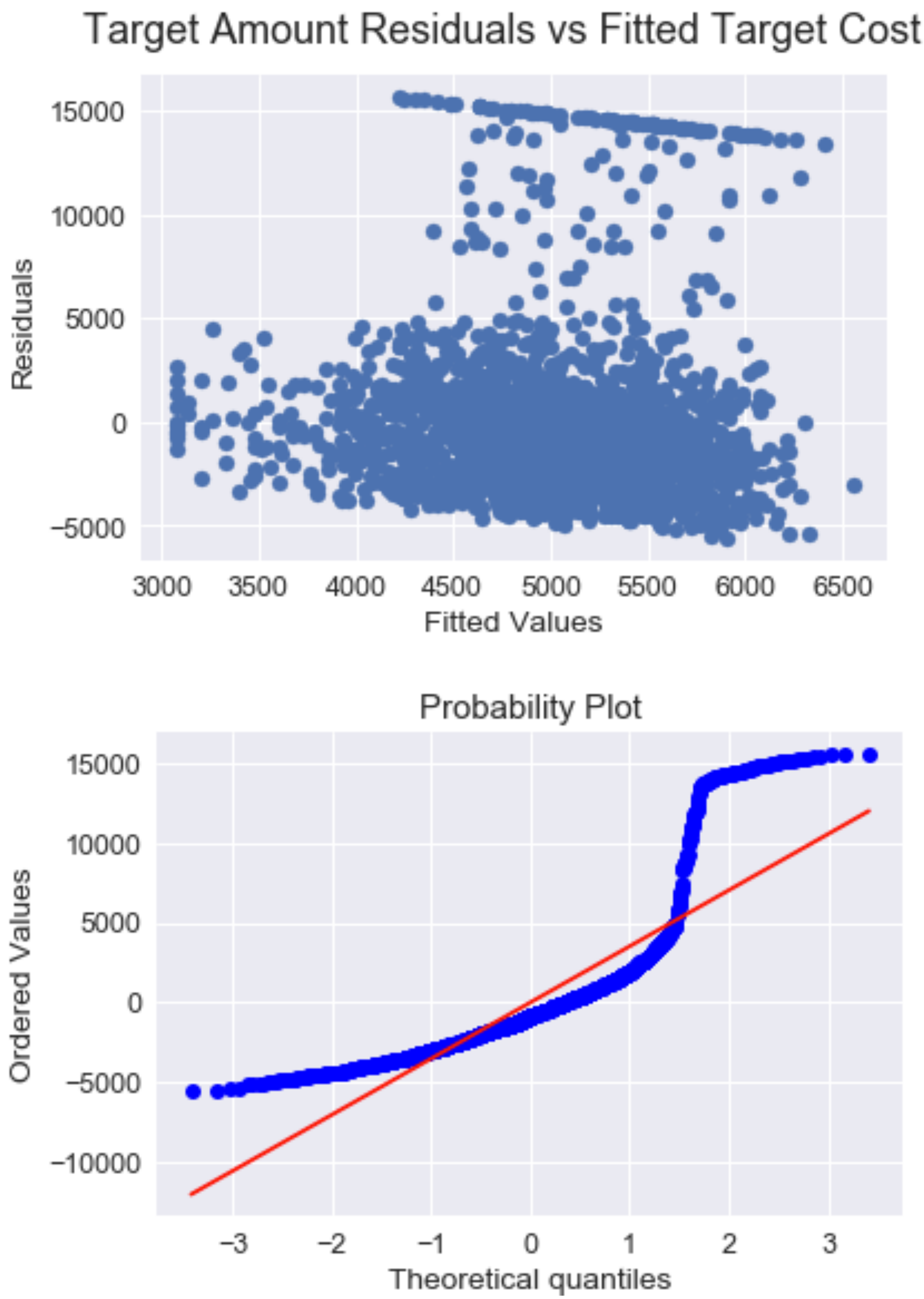
OLS Regression Results

Dep. Variable:	target_amt	R-squared:	0.019
Model:	OLS	Adj. R-squared:	0.018
Method:	Least Squares	F-statistic:	14.07
Date:	Sat, 17 Feb 2018	Prob (F-statistic):	4.41e-09
Time:	03:02:47	Log-Likelihood:	-20940.
No. Observations:	2153	AIC:	4.189e+04
Df Residuals:	2149	BIC:	4.191e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-2113.3893	1258.219	-1.680	0.093	-4580.844	354.065
log_bluebook	764.3108	132.754	5.757	0.000	503.971	1024.650
IMP_mstatus	-402.4122	174.962	-2.300	0.022	-745.524	-59.300
mvr_pts	63.9457	33.920	1.885	0.060	-2.573	130.465

Omnibus:	979.439	Durbin-Watson:	2.011
Prob(Omnibus):	0.000	Jarque-Bera (JB):	4316.894
Skew:	2.230	Prob(JB):	0.00
Kurtosis:	8.313	Cond. No.	142.

Figure 17: Residual plots



log_bluebook is the natural logarithm of a person's vehicular bluebook value. mvr_points are the number of motor vehicle record points a person has. Finally, IMP_mstatus should be 1 if a person is married, else 0. Although this model's AIC and BIC are quite high, it did perform better compared to the mean score. The QQ plot show in Figure 17 that they are not following the normal distribution.

P_TARGET_AMT = -2113.39 + 764.31 * log_bluebook – 402.41 * IMP_mstatus + 63.94 * mvr_pts

Conclusion

The objective of this assignment was to create a model that predicts best which customers are most likely to crash their cars. And if the person did get into crash, predict the amount that will cost. I have created several logistic regression models with the data provided. For the logistic regression it requires significant amount of data preparation. During the EDA and the data preparation phase, I have noticed that almost all the variables seem to play some of sort of the role in predicting the response variable. This means all the variables can be taken forward into the model. However, that produced a quite complex model and suffered from the poor prediction. I think some iterations of transformation can be carried out to refine the feature list for building the better model. From what I have seen out of all the models built, third model seems to be the best of all.