Homework 2 – PREDICT 411

Nate Bitting

Introduction and Data Exploration

For this assignment, the dataset we will be working with is the insurance dataset that consists of approximately 8000 customer records and contains 13 numeric variables, 10 categorical variables and two target variables. The two target variables include the TARGET_FLAG and TARGET_AMT. TARGET_FLAG is represented as a "1" if a customer was in a crash or a "0" if they were not. TARGET_AMT has a value of zero for anyone not in a crash and a value greater than zero for those who were in a crash. The objective of this assignment is to build a predictive model in order to calculate the probability that a customer will get into a crash and if so, estimate the value for damages. In Table 1 (below) is the list of all of the variables we will consider including in the model.

Table 1: Insurance Dataset Variable List

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Format	Label
5	AGE	Num	8	4.	Age
17	BLUEBOOK	Num	8	DOLLAR10.	Value of Vehicle
25	CAR_AGE	Num	8	4.	Vehicle Age
19	CAR_TYPE	Char	11		Type of Car
16	CAR_USE	Char	10		Vehicle Use
22	CLM_FREQ	Num	8		#Claims(Past 5 Years)
13	EDUCATION	Char	13		Max Education Level
6	HOMEKIDS	Num	8	4.	#Children @Home
10	HOME_VAL	Num	8	DOLLAR10.	Home Value
8	INCOME	Num	8	DOLLAR10.	Income
1	INDEX	Num	8		
14	JOB	Char	13		Job Category
4	KIDSDRIV	Num	8	4.	#Driving Children
11	MSTATUS	Char	5		Marital Status
24	MVR_PTS	Num	8	5.	Motor Vehicle Record Points
21	OLDCLAIM	Num	8	DOLLAR12.	Total Claims(Past 5 Years)
9	PARENT1	Char	3		Single Parent
20	RED_CAR	Char	3		A Red Car
23	REVOKED	Char	3		License Revoked (Past 7 Years)
12	SEX	Char	3		Gender
3	TARGET_AMT	Num	8		
2	TARGET_FLAG	Num	8		
18	TIF	Num	8		Time in Force
15	TRAVTIME	Num	8	4.	Distance to Work
26	URBANICITY	Char	21		Home/Work Area
7	YOJ	Num	8	4.	Years on Job

In Table 2 you can see a sample of 10 observations from the provided insurance dataset in its raw form. We will explore each of the variables in order to assess what we are working with and if any cleanup will be required before we can enter the model building phase.

Table 2: 10 Observations of Insurance Dataset (does not include all variables to save space)

Obs	INDEX	TARGET_FLAG	TARGET_AMT	KIDSDRIV	AGE	HOMEKIDS	YOJ	INCOME	PARENT1	HOME_VAL	MSTATUS	SEX	EDUCATION	JOB	TRAVTIME	CAR_USE
1	1	0	0	0	60	0	11	$67,349	No	$0	z_No	M	PhD	Professional	14	Private
2	2	0	0	0	43	0	11	$91,449	No	$257,252	z_No	M	z_High School	z_Blue Collar	22	Commercial
3	4	0	0	0	35	1	10	$16,039	No	$124,191	Yes	z_F	z_High School	Clerical	5	Private
4	5	0	0	0	51	0	14	.	No	$306,251	Yes	M	<High School	z_Blue Collar	32	Private
5	6	0	0	0	50	0	.	$114,986	No	$243,925	Yes	z_F	PhD	Doctor	36	Private
6	7	1	2946	0	34	1	12	$125,301	Yes	$0	z_No	z_F	Bachelors	z_Blue Collar	46	Commercial
7	8	0	0	0	54	0	.	$18,755	No	.	Yes	z_F	<High School	z_Blue Collar	33	Private
8	11	1	4021	1	37	2	.	$107,961	No	$333,680	Yes	M	Bachelors	z_Blue Collar	44	Commercial
9	12	1	2501	0	34	0	10	$62,978	No	$0	z_No	z_F	Bachelors	Clerical	34	Private
10	13	0	0	0	50	0	7	$106,952	No	$0	z_No	M	Bachelors	Professional	48	Commercial

As can be seen in Table 2 above, there are several fields that contain missing values that must be addressed before we can even begin to build a predictive model.

Missing Values

For this assignment we will be using logistic regression to estimate the probability of a customer getting into a crash and as a result, we must address the missing values in order for the model to perform well. In Table 3 we can see the output from the MEANS procedure for all of the numeric variables. What we are interesting in investigating is the identification of which input variables have missing values and review the mean, median, min, max, and quantiles for each variable.

Table 3: Summary from SAS Means Procedure

Variable	Label	N	N Miss	Mean	Median	Minimum	5th Pctl	50th Pctl	90th Pctl	95th Pctl	99th Pctl	Maximum
INDEX		8161	0	5152	5133	1	509	5133	9282	9791	10197	10302
TARGET_FLAG		8161	0	0	0	0	0	0	1	1	1	1
TARGET_AMT		8161	0	1504	0	0	0	0	4904	6452	19867	107586
KIDSDRIV	#Driving Children	8161	0	0	0	0	0	0	1	1	2	4
AGE	Age	8155	6	45	45	16	30	45	56	59	64	81
HOMEKIDS	#Children @Home	8161	0	1	0	0	0	0	3	3	4	5
YOJ	Years on Job	7707	454	10	11	0	0	11	15	15	17	23
INCOME	Income	7716	445	61898	54028	0	0	54028	123217	152283	215536	367030
HOME_VAL	Home Value	7697	464	154867	161160	0	0	161160	316587	374931	500309	885282
TRAVTIME	Distance to Work	8161	0	33	33	5	7	33	54	60	75	142
BLUEBOOK	Value of Vehicle	8161	0	15710	14440	1500	4900	14440	27460	31110	39090	69740
TIF	Time in Force	8161	0	5	4	1	1	4	11	13	17	25
OLDCLAIM	Total Claims(Past 5 Years)	8161	0	4037	0	0	0	0	9583	27090	42820	57037
CLM_FREQ	#Claims(Past 5 Years)	8161	0	1	0	0	0	0	3	3	4	5
MVR_PTS	Motor Vehicle Record Points	8161	0	2	1	0	0	1	5	6	8	13
CAR_AGE	Vehicle Age	7651	510	8	8	-3	1	8	16	18	21	28

N = number of observations;
N Miss = number of missing observations
Mean = mean value for each numeric variable

As shown in Table 3, there are five variables with missing values including Age, Years on the Job, Income, Home Value, and Vehicle Age. We will explore how we plan to handle these missing values in the data preparation phase. Although some of the variables in Table 3 are numeric in nature, we thought it would be good to explore them as if they were categorical variables including KIDSDRIV, HOMEKIDS, MVR_PTS, and CLM_FREQ.

Table 4: Frequency Table for KIDSDRIV, HOMEKIDS, MVR_PTS, and CLM_FREQ

#Driving Children				
KIDSDRIV	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	7180	87.98	7180	87.98
1	636	7.79	7816	95.77
2	279	3.42	8095	99.19
3	62	0.76	8157	99.95
4	4	0.05	8161	100.00

#Children @Home				
HOMEKIDS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	5289	64.81	5289	64.81
1	902	11.05	6191	75.86
2	1118	13.70	7309	89.56
3	674	8.26	7983	97.82
4	164	2.01	8147	99.83
5	14	0.17	8161	100.00

Motor Vehicle Record Points				
MVR_PTS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	3712	45.48	3712	45.48
1	1157	14.18	4869	59.66
2	948	11.62	5817	71.28
3	758	9.29	6575	80.57
4	599	7.34	7174	87.91
5	399	4.89	7573	92.80
6	266	3.26	7839	96.05
7	167	2.05	8006	98.10
8	84	1.03	8090	99.13
9	45	0.55	8135	99.68
10	13	0.16	8148	99.84
11	11	0.13	8159	99.98
13	2	0.02	8161	100.00

#Claims(Past 5 Years)				
CLM_FREQ	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	5009	61.38	5009	61.38
1	997	12.22	6006	73.59
2	1171	14.35	7177	87.94
3	776	9.51	7953	97.45
4	190	2.33	8143	99.78
5	18	0.22	8161	100.00

-

As you can see in Table 4, 88% of customers have no children who drive at home and roughly 65% do not have children. Another interesting takeaway is that 61% of customers have not had a claim in the past five years. In order to see if there is any correlation with these variables, we will examine a crosstab of each variable in Table 4 and the frequency by TARGET_FLAG.

Table 5: Crosstab of Categorical Variables vs TARGET_FLAG Frequency %

	TARGET_FLAG	
#Driving Children	**0**	**1**
0	75%	25%
1	63%	37%
2	60%	40%
3	50%	50%
4	50%	50%
Grand Total	**74%**	**26%**

	TARGET_FLAG	
Education	**0**	**1**
<High School	68%	32%
z_High School	66%	34%
Bachelors	77%	23%
Masters	80%	20%
PhD	83%	17%
Grand Total	**74%**	**26%**

	TARGET_FLAG	
#Children @Home	**0**	**1**
0	78%	22%
1	66%	34%
2	66%	34%
3	66%	34%
4	65%	35%
5	57%	43%
Grand Total	**74%**	**26%**

	TARGET_FLAG	
JOB	**0**	**1**
z_Blue Collar	67%	33%
Clerical	71%	29%
Student	63%	37%
Professional	78%	22%
Home Maker	72%	28%
Lawyer	82%	18%
Manager	86%	14%
Doctor	88%	12%
Grand Total	**74%**	**26%**

	TARGET_FLAG	
MVR_PTS	**0**	**1**
0	81%	19%
1	77%	23%
2	72%	28%
3	68%	32%
4	66%	34%
5	63%	37%
6	61%	39%
7	44%	56%
8	35%	65%
9	27%	73%
10	15%	85%
11	18%	82%
13	0%	100%
Grand Total	**74%**	**26%**

	TARGET_FLAG	
URBANICITY	**0**	**1**
Highly Urban/ Urban	69%	31%
z_Highly Rural/ Rural	93%	7%
Grand Total	**74%**	**26%**

	TARGET_FLAG	
CLM_FREQ	**0**	**1**
0	82%	18%
1	61%	39%
2	60%	40%
3	60%	40%
4	58%	42%
5	61%	39%
Grand Total	**74%**	**26%**

	TARGET_FLAG	
REVOKED	**0**	**1**
Yes	56%	44%
No	76%	24%
Grand Total	**74%**	**26%**

Table 5 provides us with keen insight as to the frequency of customers who were in a car crash based on different categories within each variable. We will leverage the data presented in Table 5 as the basis for how we will group our indicator variables. This will hopefully alleviate any multicollinearity that may exist among the predictor variables. Below are a list of takeaways from examining each of the crosstabs in Table 5.

Key Takeaways

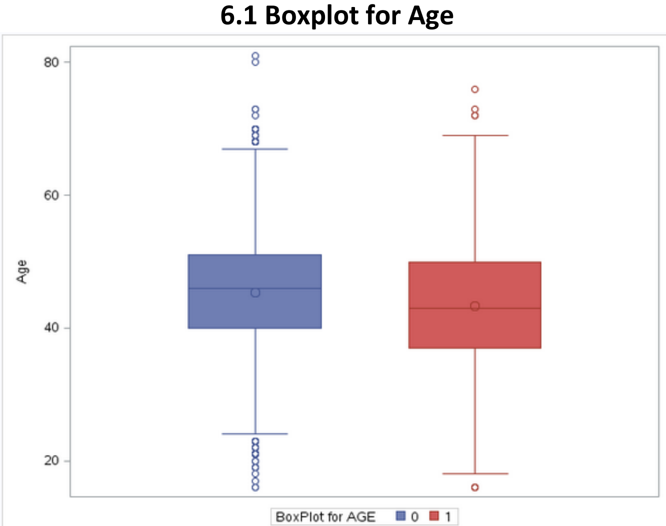
- You will notice for KIDSDRIV, it appears that the more children you have, the more likely you are to get into an accident.
- For HOMEKIDS, there are far less customers getting into a crash who do not have children vs those who do. However, it appears that there is not much of a difference for customers with one to four children, but increases for those who have five children at home.
- As one would expect, you will also notice that those with more points on their driving record tend to have a higher chance of getting into an accident.
- Those who have not had a claim in the past five years have a lower chance of getting into an accident, but the chance of getting into an accident is roughly the same regardless of how many claims submitted over the past five years.
- Students have the highest chance of getting into a crash, with blue collar workers coming in 2nd for the most likely to get in an accident.
- Those living in a rural area have a significantly smaller chance of getting into an accident.
- Those who have had their license revoked in the past 7 years have a higher probability of getting into a crash.
- Lastly, the more education one receives, the lower the probability of getting into an accident.

Assessment of Outliers

As with any data exploration exercise, we must review each predictor variable to know if any extreme observations are present in the data. To accomplish this task, we will review boxplots of the numeric variables from the insurance dataset.

Table 6: Boxplots to explore outliers

6.1 Boxplot for Age



As you can see in figure 6.1, for the age values for those not in an accident, some customers our outside of the 95th percentile, therefore they must be addressed. There are fewer outliers for those who were in a crash than those who were not.

6.2 Boxplot for Income

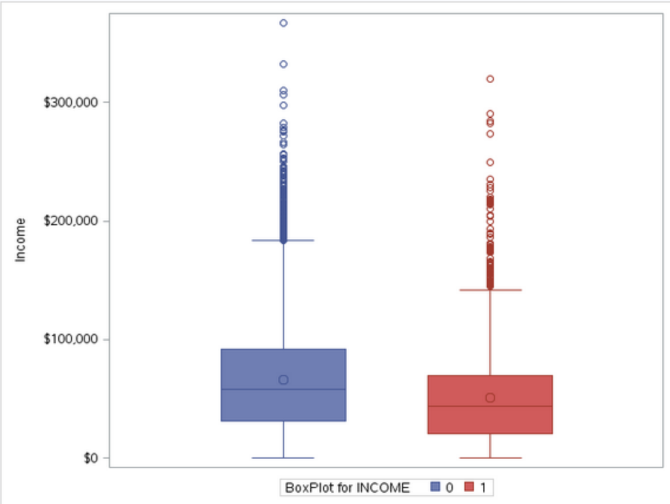
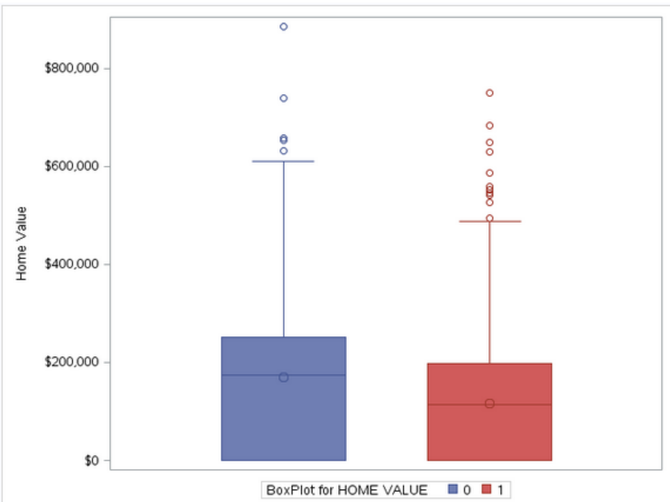


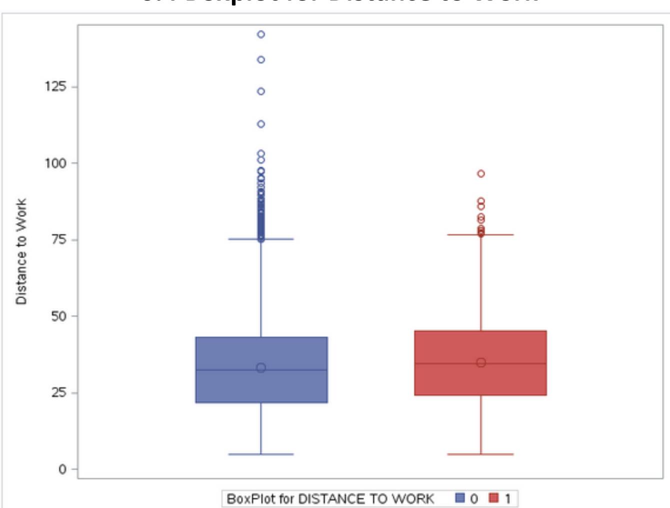
Figure 6.2 clearly shows we have outliers to deal with the Income variable. We will explore possible solutions such as normalization, trimming, standardization, or log transformations to address this issue. One interesting thing to note is that those with lower income seem to get in accidents more than those with higher income levels.

6.3 Boxplot for Home Value



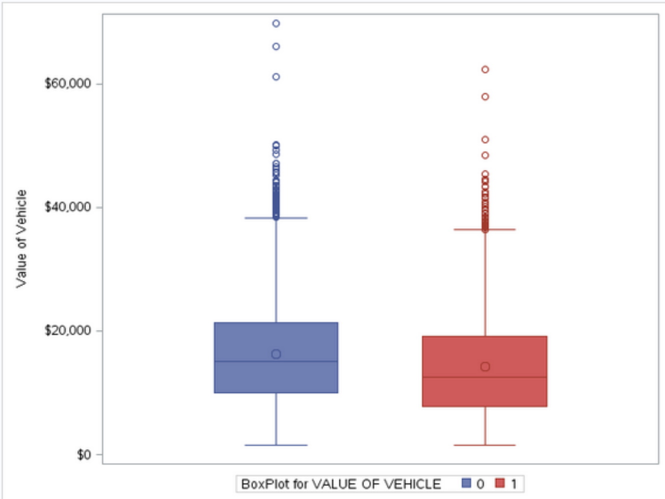
There does not seem to be much of a difference for the home values when presented in the way we have it in figure 6.3 to the left. To make things easier, we will consider creating a dummy variable to determine if a customer is a home owner or not called "home_owner." This will possibly mitigate any need for dealing with outliers in this variable.

6.4 Boxplot for Distance to Work



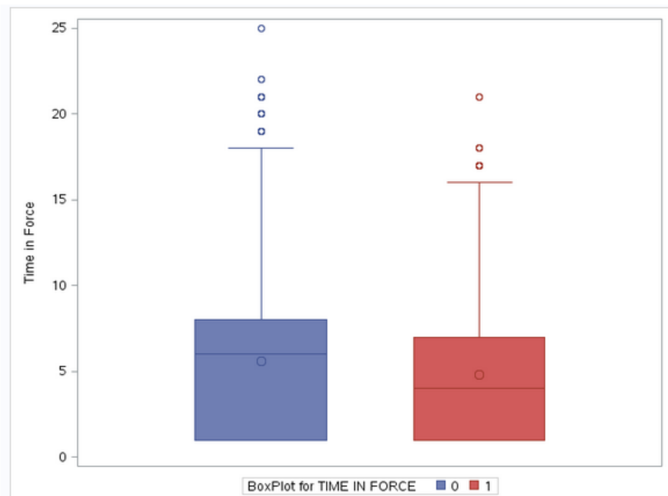
Based on the theoretical notion that the further the commute would result in higher risk of getting into an accident. However, figure 6.4 shows the boxplot for the distance to work for each customer by those who were in an accident vs those who were not. There appears to be no effect on the probability of getting into an accident based on travel distance to work. We do notice there are some outliers in this variable that must be addressed as well.

6.5 Boxplot for Value of Vehicle



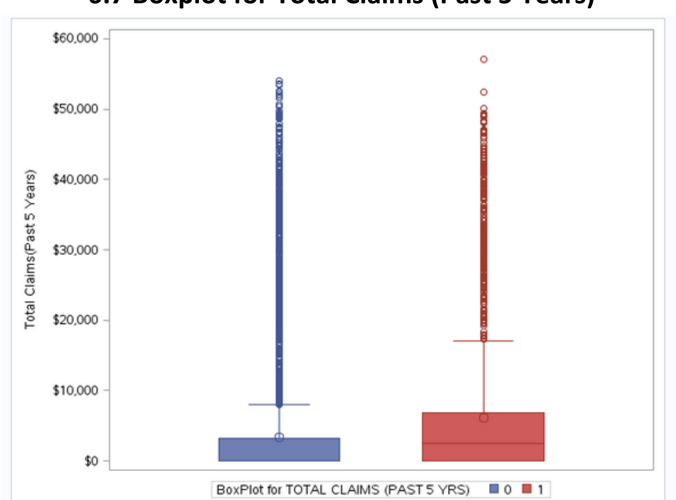
For both customers who were in an accident and those who were not have many values exceeding outside the boundaries of the 95th percentile. We will explore various techniques for dealing with these outliers.

6.6 Boxplot for Time in Force



Based on what we see in figure 6.6, there is not too much of a concern with regards to outliers for this particular variable. We can explore leveraging something as simple as trimming at the 95th percentile to address this variable.

6.7 Boxplot for Total Claims (Past 5 Years)



In figure 6.7, we can see there exists many extreme observations outside of the 95th percentile with the old claims variable. This variable must be addressed as it will most likely have a negative impact on model performance if left in its raw form.

Data Preparation

Address Missing Values

As mentioned in the previous section, the variables with missing values include Age, Years on the Job, Income, Home Value, and Vehicle Age.

<u>Results of assessing predictors with missing values</u>

- AGE: Median is 45 and 6 values are missing
- YOJ: Median is 11 and 454 values are missing
- INCOME: Median is $54,028 and 445 values are missing
- HOME_VAL: Mean is $154,867 and 464 values are missing
- CAR_AGE: Median 8 and 510 values are missing
- JOB: z_Blue Collar is the most common and 526 values are missing

For the values listed above we will create new imputed variables leveraging the median value for each variable with the exception of HOME_VAL where we will use the mean. We chose to use the median for most of the variables to ensure any extreme observations did not influence the value chosen for imputing the values. Lastly, for all values we had to impute, we created new variables with the "IMP" prefix (i.e. IMP_AGE). For any observation we had to impute, we also created a flag to indicate which observations had an imputed value. We did this by using the "M" prefix for this indicator variable (i.e. M_AGE). For the Job variable, we chose to impute the missing values with "z_Blue Collar" given it was the most common job role amongst the existing customers.

Variable Transformations

As covered in the previous section, we have several predictor variables containing extreme observations that must be addressed before we move into the model building phase. The variables covered in the previous section with outliers include AGE, INCOME, HOME_VAL, TRAVTIME, BLUEBOOK, TIF, and OLDCLAIM. We also had some categorical variables that appeared to have multicollinearity among the various categories including CAR_TYPE, EDUCATION, JOB, HOMEKIDS, KIDSDRIV, CLM_FREQ, and MVR_PTS. In Table 7 below, you can see what we chose to leverage to mitigate the outliers and any multicollinearity that exist in the data.

<u>Table 7: Variable Transformations for Variables with Outliers</u>

<table>
<tr><th>Variable</th><th>Transformation</th></tr>
<tr><td>AGE</td><td>For the age variable, we chose to leverage a binning technique based upon the following bins:
16-18 – age_1
19-20 – age_2
20-30 – age_3
30-40 – age_4
40-50 – age_5
50-60 – age_6
60 and up – no indicator variable necessary</td></tr>
<tr><td>INCOME</td><td>We chose to use a trimming transformation at the 95th percentile. The new variable contains the "T95" prefix – T95_INCOME.</td></tr>
</table>

HOME_VAL	For Home Value, we noticed in the data exploration section that there didn't seem to be much of a difference between customers who were in an accident vs those who were not. Given this fact, we chose to create a dummy variable called "home_owner" to indicate if the HOME_VAL was greater than $0 or not.
TRAVTIME	Given the outliers we saw in Table 7 for Distance to Work, we chose to use a trimming method at the 95th percentile. The new variable contains the "T95" prefix – T95_TRAVTIME.
BLUEBOOK	As with Income and Distance to work, we also chose to apply the trimming method at the 95 percentile to the BLUEBOOK variable as well. This variable also contains the "T95" prefix – T95_BLUEBOOK.
TIF	There didn't seem to be too many outliers in the TIF variable, but we chose to create bins for this variable to see if there was any difference based upon how long one was in the work force. The bins for the TIF variable are as follows: • 1 year – tif_1 • 2-4 years – tif_24 • 5-10 years – tif_510 • 11 years or higher – tif_11up
OLDCLAIM	Due to the significant number of extreme observations outside the 95th percentile, we chose to create three indicator variables for low, medium and high. We first isolated the observations to only those who had a claim greater than $0 in the past five years before we calculated the quintiles used for binning. The basis for each bin are as follows: • $0 - $3,662 (25th Percentile) – oldclaim_low • $3,663 – $9,867 (25th to 75th percentile) – oldclaim_med • $9,867 and higher (> 75th percentile) – oldclaim_high
CAR_TYPE	We created indicator variables for each category in CAR_TYPE. The only caveat is that we chose to combine Panel Truck and Van given their probabilities of getting into a car crash were almost the same. This new variable was called OTHER_CAR.
EDUCATION	For education, we created three indicator variables: bachelors, high_edu (Masters or PhD), and highschool.
JOB	Given there were many job levels with similar probabilities of getting into an accident, we chose to combine multiple categories into a single category. As a result, we chose to create three new job variables as follows: • BLUE_COLLAR – "Clerical" or "z_Blue Collar" • WHITE_COLLAR – "Doctor" or "Professional" or "Lawyer" or "Manager" • JOB_OTHER – "Student" or "Home Maker"

HOMEKIDS	Home kids is a count variable, so we chose to treat it by creating dummy variables for each number as follows: 1 Kid – homekids_1, 2 Kids – homekids_2, etc.
KIDSDRIV	For Kids Driving, we followed a similar approach as HOMEKIDS. We created four dummy variables for each number of kids driving at home: kidsdriv_1, kidsdriv_2, etc.
CLM_FREQ	AS with the previous two variables, CLM_FREQ is also a count variable and we chose to create indicator variables for each value in the data: clm_freq_1, clm_freq_2, clm_freq_3up
MVR_PTS	For MVR_PTS, there was a wider range of possible values, so we chose to bin the values into three buckets: mvr_pts_1 (1 pt), mvr_pts_25 (2-5 pts), mvr_pts_6up (6 pts or higher).

Model Building

For this section, we will review three separate models and evaluate based upon various model fit statistics and other tools for selection the optimal model.

Model A: Simple model using all input variables and the imputed variables we created to handle missing values

The first model we will review is the same model as provided in the shell code which includes all of the input variables as well as those containing imputed variables for those with missing values. The SAS output for Model A is shown in Table 8 below.

Table 8: Model A (AIC: 7458.409 KS: 0.2073 AUC: 0.808 Somers' D: 0.616)

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	9419.962	7458.409
SC	9426.969	7661.615
-2 Log L	9417.962	7400.409

Association of Predicted Probabilities and Observed Responses

Percent Concordant	80.8	Somers' D	0.616
Percent Discordant	19.2	Gamma	0.616
Percent Tied	0.0	Tau-a	0.239
Pairs	12935224	c	0.808

Kolmogorov-Smirnov Test for Variable phat Classified by Variable TARGET_FLAG

TARGET_FLAG	N	EDF at Maximum	Deviation from Mean at Maximum
0	6008	0.691578	9.620928
1	2153	0.221087	-16.071630
Total	8161	0.567455	

Maximum Deviation Occurred at Observation 1046

Value of phat at Maximum = 0.253976

Kolmogorov-Smirnov Two-Sample Test (Asymptotic)

KS	0.207346	D	0.470491
KSa	18.731245	Pr > KSa	<.0001

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.1351	0.1910	125.0215	<.0001
CAR_TYPE	Minivan	1	-0.7294	0.0853	73.0654	<.0001
CAR_TYPE	Panel Truck	1	-0.1830	0.1485	1.5179	0.2179
CAR_TYPE	Pickup	1	-0.1904	0.0923	4.2527	0.0392
CAR_TYPE	Sports Car	1	0.2434	0.0972	6.2723	0.0123
CAR_TYPE	Van	1	-0.1046	0.1188	0.7750	0.3787
CAR_USE	Commercial	1	0.7864	0.0903	75.8719	<.0001
EDUCATION	<High School	1	-0.0222	0.0938	0.0559	0.8131
EDUCATION	Bachelors	1	-0.4162	0.0827	25.3311	<.0001
EDUCATION	Masters	1	-0.4601	0.1144	16.1814	<.0001
EDUCATION	PhD	1	-0.3568	0.1563	5.2098	0.0225
IMP_JOB	Clerical	1	0.1462	0.1042	1.9697	0.1605
IMP_JOB	Doctor	1	-0.5606	0.2584	4.7081	0.0300
IMP_JOB	Home Maker	1	0.00648	0.1385	0.0022	0.9627
IMP_JOB	Lawyer	1	0.00181	0.1498	0.0001	0.9904
IMP_JOB	Manager	1	-0.7719	0.1220	40.0024	<.0001
IMP_JOB	Professional	1	-0.0590	0.1097	0.2894	0.5906
IMP_JOB	Student	1	-0.00371	0.1210	0.0009	0.9756
MSTATUS	Yes	1	-0.4755	0.0789	36.2697	<.0001
PARENT1	No	1	-0.4686	0.0933	25.1987	<.0001
URBANICITY	Highly Urban/ Urban	1	2.4164	0.1123	463.3175	<.0001
BLUEBOOK		1	-0.00002	4.683E-6	24.2674	<.0001
CLM_FREQ		1	0.1484	0.0253	34.3553	<.0001
IMP_HOME_VAL		1	-1.41E-6	3.381E-7	17.4620	<.0001
IMP_INCOME		1	-3.58E-6	1.064E-6	11.3565	0.0008
KIDSDRIV		1	0.4414	0.0546	65.4376	<.0001
MVR_PTS		1	0.1099	0.0134	66.9886	<.0001
TIF		1	-0.0568	0.00729	60.7253	<.0001
TRAVTIME		1	0.0144	0.00187	59.6336	<.0001

As you can see in Table 8, the area under the curve (AUC, aka c) is reasonable at 0.808, however when assessing the p-values for each of the predictor variables, we notice there are several variables that have a p-value exceeding the 0.05 threshold. We also can assume that without any variable transformations, the model will most likely not perform well I production given the noise that exists in the data. Let's explore our second model by applying some basic transformations on some of the variables we discussed in the data preparation section.

Model B: Addition of new calculated variables and binning techniques

Due to the significant issues with the p-values in Model A, we will explore performing variable transformations on some of the variables to see if the model improves. For Model B, we chose to perform some trimming on some of the variables containing outliers such as AGE, BLUEBOOK, TRAVTIME, INCOME, etc. As shown in Table 9, the AUC improved slightly from Model A, going from 0.808 up to 0.810.

Table 9: Model B (AIC: 6969.213 KS: 0.2090 AUC: 0.810 Somers' D: 0.619)

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	8818.622	6969.213
SC	8825.562	7156.607
-2 Log L	8816.622	6915.213

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	81.0	Somers' D	0.619
Percent Discordant	19.0	Gamma	0.619
Percent Tied	0.0	Tau-a	0.241
Pairs	11331506	c	0.810

Kolmogorov-Smirnov Test for Variable phat Classified by Variable TARGET_FLAG			
TARGET_FLAG	N	EDF at Maximum	Deviation from Mean at Maximum
0	5618	0.713599	9.388352
1	2017	0.239465	-15.668495
Total	7635	0.588343	
Maximum Deviation Occurred at Observation 4351			
Value of phat at Maximum = 0.269520			

Kolmogorov-Smirnov Two-Sample Test (Asymptotic)			
KS	0.209043	D	0.474135
KSa	18.265894	Pr > KSa	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.4071	0.1959	151.0451	<.0001
CAR_TYPE	Minivan	1	-0.7155	0.0858	69.5322	<.0001
CAR_TYPE	Panel Truck	1	-0.0786	0.1658	0.2247	0.6355
CAR_TYPE	Pickup	1	-0.1508	0.0941	2.5683	0.1090
CAR_TYPE	Sports Car	1	0.2551	0.0973	6.8808	0.0087
CAR_TYPE	Van	1	-0.1080	0.1255	0.7409	0.3894
CAR_USE	Commercial	1	0.7737	0.0925	69.8894	<.0001
EDUCATION	<High School	1	-0.0312	0.0941	0.1103	0.7398
EDUCATION	Bachelors	1	-0.3780	0.0833	20.5968	<.0001
EDUCATION	Masters	1	-0.3563	0.1468	5.8859	0.0153
EDUCATION	PhD	1	-0.00757	0.1987	0.0014	0.9696
JOB	Clerical	1	0.0725	0.1064	0.4638	0.4958
JOB	Doctor	1	-0.8360	0.2964	7.9541	0.0048
JOB	Home Maker	1	-0.1068	0.1470	0.5281	0.4674
JOB	Lawyer	1	-0.1047	0.1872	0.3130	0.5758
JOB	Manager	1	-0.8567	0.1384	38.3002	<.0001
JOB	Professional	1	-0.1223	0.1186	1.0630	0.3025
JOB	Student	1	0.00756	0.1225	0.0038	0.9508
MSTATUS	Yes	1	-0.6221	0.0712	76.3330	<.0001
PARENT1	No	1	-0.4389	0.0963	20.7832	<.0001
URBANICITY	Highly Urban/ Urban	1	2.3673	0.1126	442.2263	<.0001
T95_BLUEBOOK		1	-0.00003	5.252E-6	25.8395	<.0001
T95_INCOME		1	-6.62E-6	1.233E-6	28.8062	<.0001
KIDSDRIV		1	0.4161	0.0555	56.2436	<.0001
MVR_PTS		1	0.1053	0.0144	53.6044	<.0001
LN_OLDCLAIM		1	0.0469	0.00753	38.8770	<.0001
T95_TRAVTIME		1	0.0166	0.00209	63.3978	<.0001

While the AIC, AUC, Somers' D, and KS improved from Model A, we still see some issues with the p-values for many of the variables. Based on the data exploration we covered earlier, we reasonably assume multicollinearity exists among the variables given they contain equal probability for getting into a crash for many category values. To address this, we will leverage the use of the binned variables we covered in the data preparation section earlier.

Model C: A combination of multiple variable transformation techniques

To address the issues we saw in both models A and B, we will explore the use of multiple variable transformations in Model C including binning and trimming. Please refer to Table 10 to see the output of Model C.

Table 10: Model C (AIC: 7323.139 KS: 0.2105 AUC: 0.815 Somers' D: 0.630)

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	9419.962	7323.139
SC	9426.969	7519.338
-2 Log L	9417.962	7267.139

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	81.5	Somers' D	0.630
Percent Discordant	18.5	Gamma	0.630
Percent Tied	0.0	Tau-a	0.245
Pairs	12935224	c	0.815

Kolmogorov-Smirnov Test for Variable phat Classified by Variable TARGET_FLAG			
TARGET_FLAG	N	EDF at Maximum	Deviation from Mean at Maximum
0	6008	0.724368	9.769056
1	2153	0.246633	-16.319075
Total	8161	0.598334	
Maximum Deviation Occurred at Observation 6336			
Value of phat at Maximum = 0.268431			

Kolmogorov-Smirnov Two-Sample Test (Asymptotic)			
KS	0.210538	D	0.477735
KSa	19.019639	Pr > KSa	<.0001

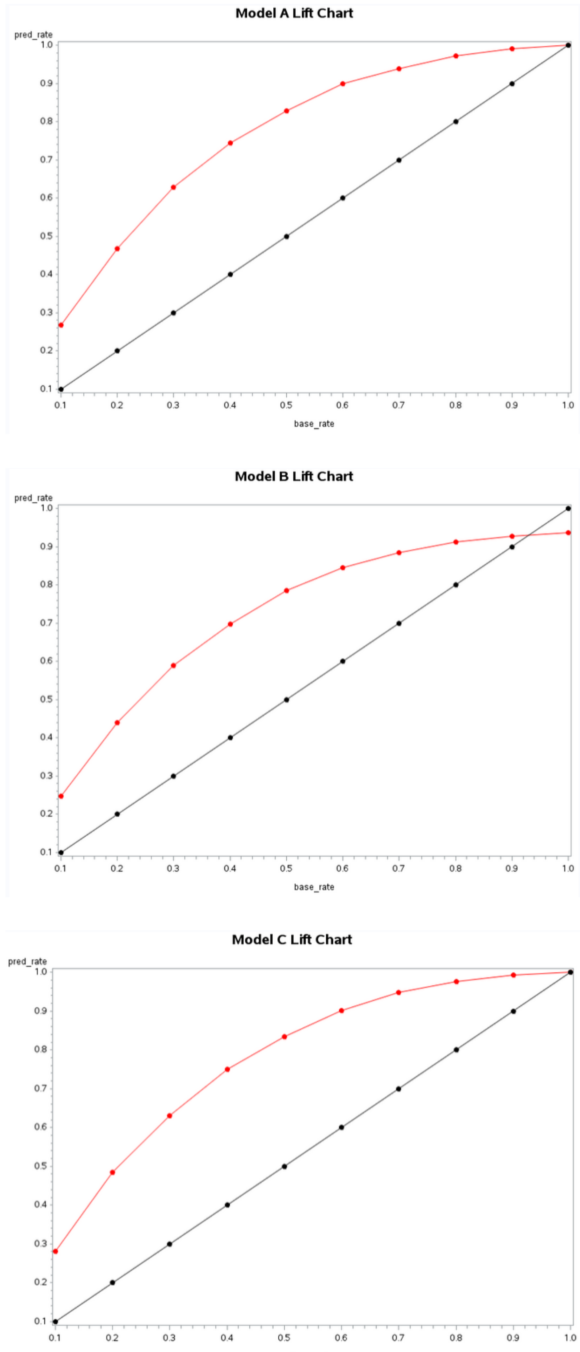
Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.3459	0.1930	48.6166	<.0001
CAR_USE	Commercial	1	0.6903	0.0694	98.9177	<.0001
MSTATUS	Yes	1	-0.5486	0.0812	45.6882	<.0001
PARENT1	No	1	-0.2805	0.0989	8.0496	0.0046
URBANICITY	Highly Urban/ Urban	1	2.3386	0.1125	431.9628	<.0001
REVOKED	No	1	-0.9737	0.0844	132.9967	<.0001
minivan		1	-0.6126	0.0764	64.3431	<.0001
sportscar		1	0.2563	0.0933	7.5457	0.0060
bachelor		1	-0.4155	0.0791	27.6257	<.0001
high_edu		1	-0.4166	0.0918	20.6034	<.0001
white_collar		1	-0.3415	0.0791	18.6221	<.0001
kidsdriv_1		1	0.7100	0.1065	44.4564	<.0001
kidsdriv_2		1	0.8666	0.1515	32.7172	<.0001
kidsdriv_3		1	1.0373	0.2996	11.9846	0.0005
tif_1		1	0.5087	0.0665	58.4513	<.0001
tif_24		1	0.3350	0.0768	19.0120	<.0001
mvr_pts_25		1	0.2767	0.0641	18.6483	<.0001
mvr_pts_6up		1	0.7335	0.1064	47.5198	<.0001
age_4		1	-0.4225	0.1100	14.7468	0.0001
age_5		1	-0.7164	0.1080	44.0328	<.0001
age_6		1	-0.4363	0.1167	13.9843	0.0002
oldclaim_low		1	0.5150	0.0941	29.9637	<.0001
oldclaim_med		1	0.5548	0.0750	54.6529	<.0001
home_owner		1	-0.2668	0.0764	12.2057	0.0005
T95_BLUEBOOK		1	-0.00003	4.353E-6	42.0722	<.0001
T95_INCOME		1	-5.49E-6	9.882E-7	30.9124	<.0001
T95_TRAVTIME		1	0.0173	0.00204	71.9078	<.0001
M_AGE		1	2.4979	1.2373	4.0761	0.0435

As we can see in Model C, the AIC was higher than Model B, but still better than Model A. However, the AUC, Somers's D, and KS statistic were highest in Model C. Lastly, our other objective was to develop a parsimonious model that only included the necessary variables to retain model predictive power and contain only variables with a p-values less than 0.05. In the next section, we will cover how we came to arrive at our final model to put into production.

Model Selection

After reviewing each model in depth in the prior section, we will review the model fit statistics in order to finalize on a model to select for deployment. Before we look at the criteria for selecting a final model, let us first review the lift charts for each model in Table 11 on the following page.

Table 11: Lift Charts for Models A, B, and C



The lift chart for Model A appears to be reasonable as it does not intersect with the black line. However, the lift chart for Model B does in fact intersect with the black line around 0.9 for the base rate, indicating the model could have some issues in production. Lastly, the lift chart for Model C appears to have slightly better performance than Model A based on the higher prediction rates towards the right side of the curve. We will not cover the criteria we used for the final model selection to move into production on the following page.

The primary measures we will use to assess the models include the following:

1. AIC
2. KS-Statistic
3. Somers' D
4. Area under the curve (AUC)
5. All predictors with p-values less than 0.05
6. Favorable Lift Chart

Table 12: Model Selection Criteria

	Model A	Model B	Model C
AIC: Intercept & Covariates	7458.409	6969.213	7323.139
KS-Statistic	0.207346	0.209043	0.210538
Somers' D	0.616	0.619	0.630
AUC	0.8082	0.8097	0.8149
P-values < .05	No	No	Yes
Favorable Lift Chart	Yes	No	Yes

As you can see in Table 12, Model B ended up with the lowest AIC, but did not have the highest AUC and the p-values for many of the variables exceeded 0.05. Model C ended up with the second lowest AIC, the highest KS-Statistic, highest Somers' D and AUC, all p-values were below 0.05, and the lift chart looked good. Therefore, we have selected Model C as our final model we will put into production.

Conclusion

The last step we will have to conduct, now that we have arrived at a final model, is to create the final data step needed to process new, out-of-sample, observations by running it through our model. This process involves the handling of missing values, creation of bins, and all trimming transformations.

As we discussed at the beginning of this report, the objective of this assignment was not only to predict the probability of a customer getting into an accident, but also to estimate the TARGET_AMT for potential damages that would be incurred if they got into an accident. Given the fact that the primary learning objective for this assignment was around logistic regression, we chose to use the mean value for TARGET_AMT from the training dataset after excluding all customers who did not get into an accident. The mean value we used to estimate TARGET_AMT on the training data was $5,702.18.

The data exploration and subsequent model building provided us with useful insight about our customer base and what factors have a significant impact on the likely chance of customers getting into an accident or not. We can provide this insight to the rest of the insurance organization so they can leverage it for useful information to provide to our customers.