Insurance

Introduction

In this report we are analyzing an auto insurance company's customer data in order to predict which customers are most likely to crash their cars. We will be using the Insurance data set to build logistic regression models to predict car crashes. Creating an accurate predictive model provides the auto insurance company with valuable insight that can help the company identify and thus avoid insuring potentially risky new customers. 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 and decision trees. 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 Insurance data set contains 8161 observations with 13 numeric variables, 10 categorical variables, 2 target variables, and 1 index variable concerning an auto insurance company's customer data. For this report, we will be focusing on predicting the TARGET_FLAG variable, which signifies whether or not a customer crashed his car. In **Table 1**, we can see that approximately a fourth of all customers crash their cars, as denoted by the "1" in the TARGET_FLAG column.

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

Table 1: Customer Car Crash Summary

In order to predict which customers are most likely to crash their cars, we must first look at the data we have. In **Table 2**, we can see that most of the numeric variables contain all 8161 data points. Of the 13 numeric variables in the data set, five have missing data: AGE, YOJ, INCOME, HOME_VAL, and CAR_AGE. Because these variables with missing data are missing at most 6 percent of their values, we can fix them by imputation. We will discuss in detail our method of imputation in the next section. For now, most of the data for the variables seems reasonable, with the exception of the minimum for CAR_AGE. A car's age cannot be negative, so we know we will have to correct what is likely a typo during our data preparation stage.

Variable	Label	Minimum	Maximum	Mean	Median	N	N Miss
KIDSDRIV	# Driving Children	0	4.0000000	0.1710575	0	8161	0
AGE	Age	16.0000000	81.0000000	44.7903127	45.0000000	8155	6
HOMEKIDS	# Children @ Home	0	5.0000000	0.7212351	0	8161	0
YOJ	Years on Job	0	23.0000000	10.4992864	11.0000000	7707	454
INCOME	Income	0	367030.26	61898.10	54028.17	7716	445
HOME_VAL	Home Value	0	885282.34	154867.29	161159.53	7697	464
TRAVTIME	Distance to Work	5.0000000	142.1206304	33.4887972	32.8709696	8161	0
BLUEBOOK	Value of Vehicle	1500.00	69740.00	15709.90	14440.00	8161	0
TIF	Time in Force	1.0000000	25.0000000	5.3513050	4.0000000	8161	0
OLDCLAIM	Total Claims (Past 5 Years)	0	57037.00	4037.08	0	8161	0
CLM_FREQ	# Claims (Past 5 Years)	0	5.0000000	0.7985541	0	8161	0
MVR_PTS	Motor Vehicle Record Points	0	13.0000000	1.6955030	1.0000000	8161	0
CAR_AGE	Vehicle Age	-3.0000000	28.0000000	8.3283231	8.0000000	7651	510

Table 2: Overview of Numeric Variables in Insurance Data Set

After reviewing the numeric variables in **Table 2**, we turn our attention to the categorical variables. We discover that only one variable is missing values, JOB. **Table 3** shows us the frequency of jobs among customers and reveals that 526 values, or approximately 6 percent, of all values are missing. Given that JOB is, generally, inherently related to some of the numeric variables, such as INCOME and HOME_VAL, we will likely be able to impute some of the missing jobs based on some of these numeric variables.

Job Category				
JOB	Frequency	Percent	Cumulative Frequency	Cumulative Percent
	526	6.45	526	6.45
Clerical	1271	15.57	1797	22.02
Doctor	246	3.01	2043	25.03
Home Maker	641	7.85	2684	32.89
Lawyer	835	10.23	3519	43.12
Manager	988	12.11	4507	55.23
Professional	1117	13.69	5624	68.91
Student	712	8.72	6336	77.64
z_Blue Collar	1825	22.36	8161	100.00

Table 3: Frequency of Jobs

After looking at summary tables to identify the missing values in both the numeric and categorical variables in the Insurance data set, we continue our data exploration by looking at the distribution and probability plots for the numeric variables. We do this to understand the data better and get a sense of possible outliers.

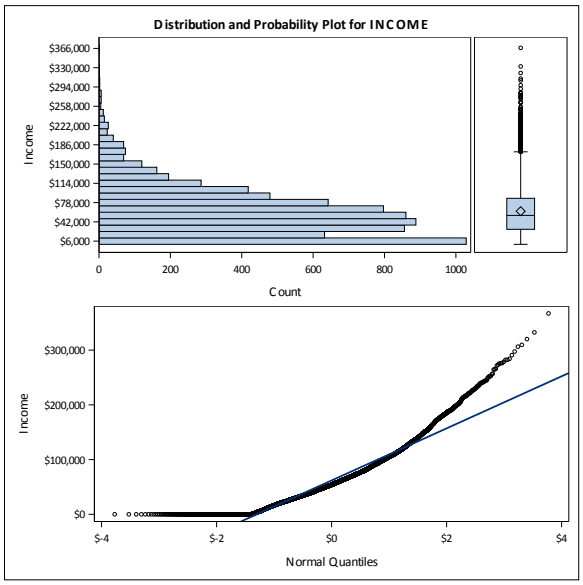

Figure 1: Distribution of INCOME

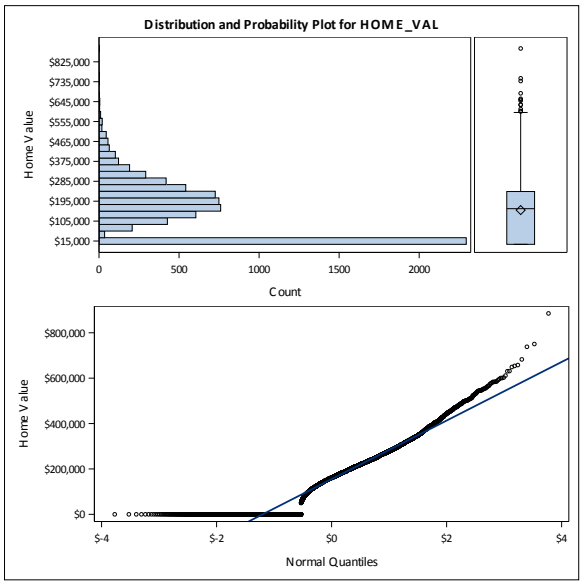

Figure 2: Distribution of HOME_VAL

In **Figures 1 and 2**, we can see the distribution of values for INCOME and HOME_VAL, prior to correcting the data for missing values. The plots do not reveal any abnormalities from our assumptions: we expect most of the data to fall in the mid- to lower-end of the distributions since most people do not earn an exorbitant salary or own a super expensive home.

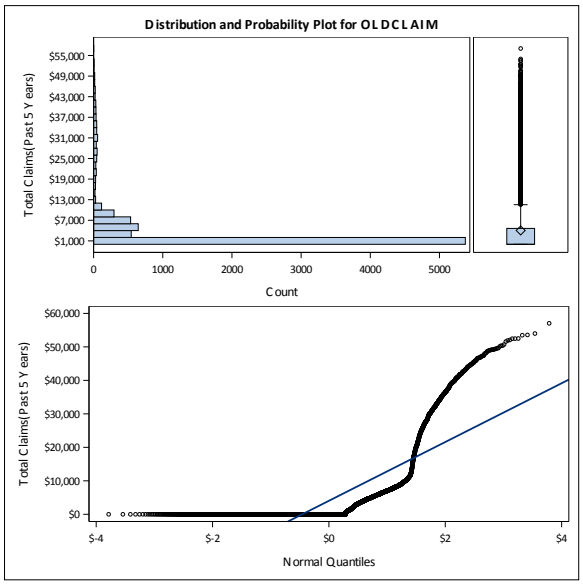

Figure 3: Distribution of OLDCLAIM

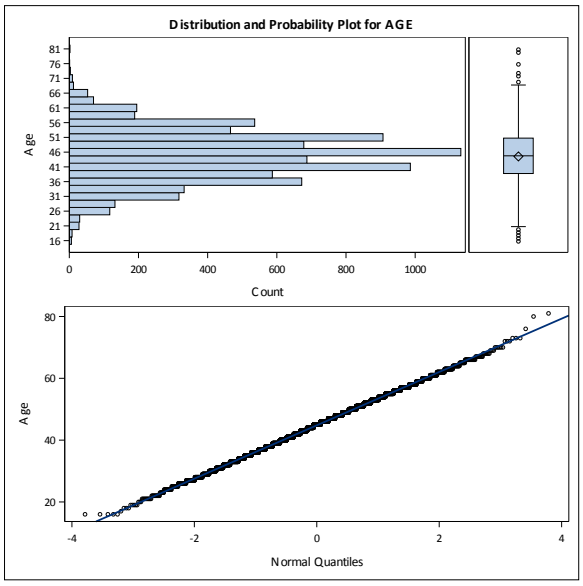

Figure 4: Distribution of AGE

As we examine the distribution plots of the other numeric variables, we try to determine if the data seems logical and whether or not we might want to modify a variable's data during the data preparation stage. We are not concerned with making sure the variables exhibit a normal distribution—each variable's distribution just must seem reasonable. For instance, the plot for OLDCLAIM, shown in **Figure 3**, is not even slightly normal. This is okay since we expect most people to have no claims in the past 5 years, some people to have small claim amounts, and only a few people to have high previous claims. This data, while abnormally distributed, is reasonable. We might want to bin the data during our data preparation stage to yield better predictions, but we

are not concerned with the distribution of the data on a whole. As for AGE, shown in **Figure 4**, this variable does follow a normal distribution. This is something we expect. If the plot showed mostly 20-year-olds and 80-year-olds, then we would have to question the quality of our data.

We continue our data exploration of numeric variables by plotting out the distributions of values classified by TARGET_FLAG. Most of the histogram plots show a similar distribution of values between customers who crashed their cars (TARGET_FLAG = 1) and customers who did not crash their cars (TARGET_FLAG = 0). However, a few variables exhibited some noticeable differences. In **Figure 5**, we see that the percentage of customers who rent their homes (as represented by the highest vertical bar in each histogram) is much higher among customers who crash their cars than customers who do not crash their cars. If you recall from **Figure 1**, only about 25 percent of customers crash their cars. Of the customers who crash their cars, about 40 percent are people who rent. On the other hand, renters make up only about 25 percent of the customers who do not crash their cars.

In **Figure 6**, we see another pattern. Approximately 70 percent of customers who do not crash their cars have a claim frequency of 0, but only 40 percent of customers who do crash their cars have a claim frequency of 0. The vertical bars representing the percentage of claims greater than 0 are overall higher for customers who crash their cars than for customers who do not crash their cars. For instance, the percentage of people who filed 2 claims in the past 5 years is about twice as high for customers who crashed their cars as for customers who did not crash their cars. This suggests that customers with a higher number of claims generally have a higher risk of crashing.

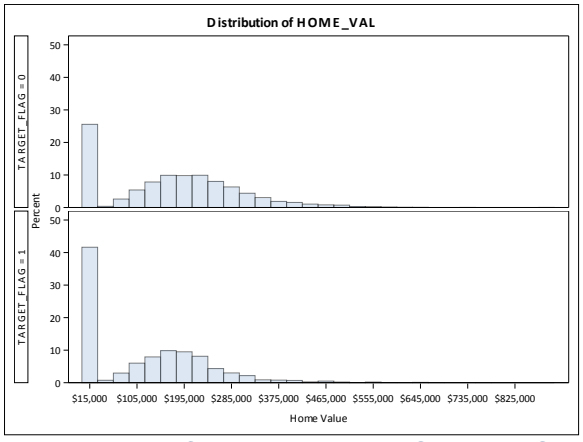

Figure 5: HOME_VAL by TARGET_FLAG

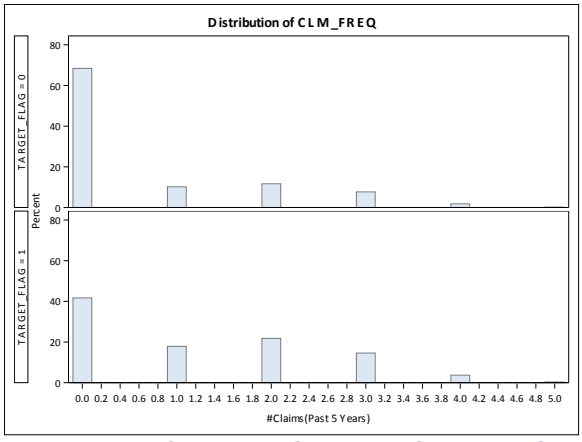

Figure 6: CLM_FREQ by TARGET_FLAG

After examining the plots for the numeric variables, we created frequency tables and histograms of the categorical variables. Seven of the 10 variables contain only two categories, so while it was interesting to see the distributions, the results were not surprising. If anything, plotting histograms confirmed our assumptions about the data. For instance, in **Figure 7**, we see that three times as many customers live or work in a highly urban or urban area as opposed to a highly rural or rural area. Given that the population density of cities is much greater than smaller towns, it makes sense that the majority of the customers lives or works in urban areas. We checked histograms of each categorical variable to make sure the data distributions seemed reasonable. They did.

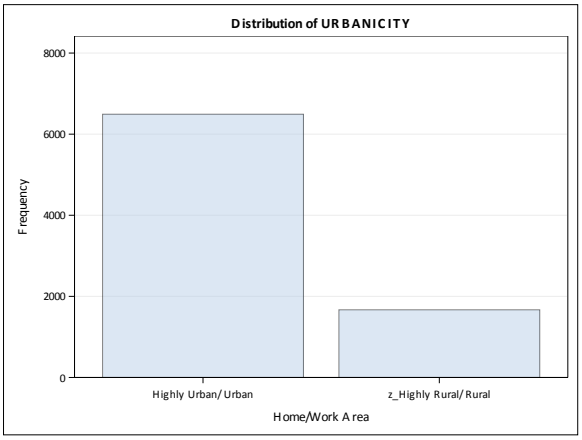

Figure 7: Distribution of URBANICITY

After confirming that the data seemed logical from the histograms, we created frequency tables to show each categorical variable classified by TARGET_FLAG. We did this to detect possible patterns and discover which categories might be predictive of car crashes.

A common assumption people have is that men are riskier drivers than women. By looking at **Table 4**, can see that this assumption is not necessarily true. When we look at the row where TARGET_FLAG = 1, we see that 14.61 percent of the total percentage of crashes from the data set, 26.38, is attributed to women. Men are responsible for only 11.78 percent of the total crashes. Although the table indicates that women are more likely to have car accidents, we must keep in mind that women make up a larger portion of the customer data set. The row percentages for TARGET_FLAG = 1 are similar to the total percentages of how the data set is divided between men and women. Should the data set have an equal number of data for men and women, the crash frequencies might differ.

Table of TARGET_FLAG by SEX			
TARGET_FLAG	**SEX (Gender)**		
Frequency Percent Row Pct Col Pct	**M**	**z_F**	**Total**
0	2825 34.62 47.02 74.62	3183 39.00 52.98 72.75	6008 73.62
1	961 11.78 44.64 25.38	1192 14.61 55.36 27.25	2153 26.38
Total	3786 46.39	4375 53.61	8161 100.00

Table 4: TARGET_FLAG by SEX

In **Table 5**, we see a similar story to the table above. Customers who have earned a PhD make up the smallest percentage of crashes (only 1.53 percent of 26.38 percent), but they also represent the smallest portion of the entire data set (728 of 8161 records). However, if we compare customers with a Bachelors degree with customers with a High School diploma, we find

something more interesting. Both of these levels of education represent a similar portion of the overall data set (27.47 percent for Bachelors, 28.55 percent for z_High School), yet customers with a High School diploma are about 1.5 times more likely to have a car crash than customers with a Bachelors degree (9.72 to 6.41 percent). The row percentage for z_High School is 36.83 percent, which is also considerably higher than the total percentage of crashes, 26.38 percent. This indicates that customers with a High School diploma are more likely to crash their cars than the average person. These observations suggest that EDUCATION is a predictive variable, which is something to keep in mind when we prepare and build our models.

Table of TARGET_FLAG by EDUCATION						
TARGET_FLAG	**EDUCATION (Max Education Level)**					
Frequency Percent Row Pct Col Pct	**<High School**	**Bachelors**	**Masters**	**PhD**	**z_High School**	**Total**
0	818 10.02 13.62 68.00	1719 21.06 28.61 76.67	1331 16.31 22.15 80.28	603 7.39 10.04 82.83	1537 18.83 25.58 65.97	6008 73.62
1	385 4.72 17.88 32.00	523 6.41 24.29 23.33	327 4.01 15.19 19.72	125 1.53 5.81 17.17	793 9.72 36.83 34.03	2153 26.38
Total	1203 14.74	2242 27.47	1658 20.32	728 8.92	2330 28.55	8161 100.00

Table 5: TARGET_FLAG by EDUCATION

Looking at the frequencies and distributions of the variables, especially in relation to TARGET_FLAG, has given us some ideas how we might approach the data preparation, model creation, and model selection stages in this project. In the next section we will discuss in detail how we used the observations we made during our data exploration in order to prepare the data for modeling.

Data Preparation

The first step we took in preparing the data was to fix the missing values. **Table 6** shows just the numeric variables with missing values. The first variable we fixed was AGE. Because only six values were missing and because the mean and median for AGE were both approximately 45, we decided to impute this value for the missing values. The next variable we fixed was CAR_AGE. The mean and median for this variable were both approximately 8, so we imputed this value for all the missing values. We also noticed that CAR_AGE had one negative value. Because it is impossible for a car to be -3 years old, we interpreted this as a typo and imputed 3 for this value.

Variable	Label	Minimum	Maximum	Mean	Median	N	N Miss
AGE	Age	16.0000000	81.0000000	44.7903127	45.0000000	8155	6
YOJ	Years on Job	0	23.0000000	10.4992864	11.0000000	7707	454
INCOME	Income	0	367030.26	61898.10	54028.17	7716	445
HOME_VAL	Home Value	0	885282.34	154867.29	161159.53	7697	464
CAR_AGE	Vehicle Age	-3.0000000	28.0000000	8.3283231	8.0000000	7651	510

Table 6: Overview of Numeric Variables with Missing Values

With both AGE and CAR_AGE, we thought it best to impute the median because the values for these variables were discrete, so imputing the mean, though it was nearly the same value, would be inconsistent with how the rest of the values for these variables were populated. Also, we determined that it might be difficult to impute the missing values for AGE and CAR_AGE using some of the other variables as references since these two variables do not have any obvious inherent relationships to other variables. Variables such as JOB, INCOME, and HOME_VAL, for instance, all have an inherent relationship we will discuss shortly.

The next variable we decided to fix was YOJ. Although the mean and median for this variable are both nearly the same, we decided against imputing 11 for all the missing values. We figured imputing 11 years on the job for a 20-year-old or a similarly young person would be unrealistic—most people do not start working when they are 9 years old. In order to impute YOJ, we first created bins for the age variable that contained the imputed ages, IMP_AGE. We specified the age bins as follows:

Bin 0: IMP_AGE ≤ 25
Bin 1: $25 <$ IMP_AGE ≤ 35
Bin 2: $35 <$ IMP_AGE ≤ 45
Bin 3: $45 <$ IMP_AGE ≤ 55
Bin 4: $55 <$ IMP_AGE ≤ 65
Bin 5: $65 <$ IMP_AGE

We then created a table of YOJ classified by the age bins. In **Table 7**, we can see the distribution of YOJ per bin and how many YOJ values are missing from each bin. We decided to impute the missing YOJ values by using the mean rounded to the nearest whole number for each bin. For instance, we imputed the missing values as 8 for AGE_BIN 0, as 10 for AGE_BIN 1, etc.

Analysis Variable : YOJ Years on Job						
AGE_BIN	N Obs	N	N Miss	Minimum	Maximum	Mean
0	96	93	3	0	15.0000000	8.2365591
1	1064	1004	60	0	18.0000000	9.8486056
2	3124	2953	171	0	19.0000000	10.4564849
3	3019	2857	162	0	18.0000000	10.4441722
4	795	740	55	0	19.0000000	11.8162162
5	63	60	3	0	23.0000000	13.3833333

Table 7: YOJ per AGE_BIN

After imputing the missing values for YOJ, we started imputing the missing values for INCOME. Because income is, in general, strongly tied to a person's job, we decided to impute the missing incomes using the mean income for each job. For instance, if a customer's record showed she was a Doctor, we imputed the missing salary as $128,679.71. **Table 8** shows the mean values we imputed for each missing value of INCOME. If a record was missing the person's JOB and INCOME, we imputed the overall mean INCOME, $61,898.10, from **Table 6** as the income.

Analysis Variable : INCOME Income						
Job Category	**N Obs**	**N**	**N Miss**	**Minimum**	**Maximum**	**Mean**
	526	502	24	5821.51	332338.99	118852.94
Clerical	1271	1198	73	2954.14	113845.14	33861.19
Doctor	246	238	8	0	367030.26	128679.71
Home Maker	641	598	43	0	185747.34	12073.33
Lawyer	835	799	36	0	282292.36	88304.81
Manager	988	938	50	0	309628.41	87461.56
Professional	1117	1057	60	0	231003.56	76593.11
Student	712	659	53	0	93066.98	6309.65
z_Blue Collar	1825	1727	98	5014.93	198319.70	58957.01

Table 8: INCOME per JOB

Once we fixed all the missing incomes, we began fixing the missing jobs using the imputed incomes. Because there is a range of salaries for each job and because salaries can overlap among professions, we decided to use a ballpark value of the imputed income range as the cutoff for imputing jobs. We chose our ballpark values as somewhere between the 50th and 75th percentiles for each of the highest paying professions. Our cutoffs are as follows:

Doctor: IMP_INCOME > $125,000
Lawyer: $95,000 < IMP_INCOME ≤ $125,000
Manager: $75,000 < IMP_INCOME ≤ $95,000
Professional: $60,000 < IMP_INCOME ≤ $75,000

For any values that fell below the lowest cutoff, we imputed that job as z_Blue Collar because, as **Figure 8** reveals, blue-collar jobs are the most common in the data set. While imputing some of the missing jobs as Clerical, Home Maker, or Student would have likely created a more accurate data set, determining cutoff values for these jobs would have been resource-intensive because there is so much income overlap among these jobs. Furthermore, because we had already imputed missing incomes for records without a job listed as $61,898.10, we thought that imputing z_Blue Collar for the remaining jobs was most prudent given that the mean salary for blue collar workers, $58,957.01 per **Table 8**, was closest to the mean income we had imputed. The mean incomes for Clerical, Home Maker, and Student, are far below $61,898.10, so imputing these jobs into the records where INCOME was imputed as $61,898.10 would have reduced the accuracy of the data as a whole.

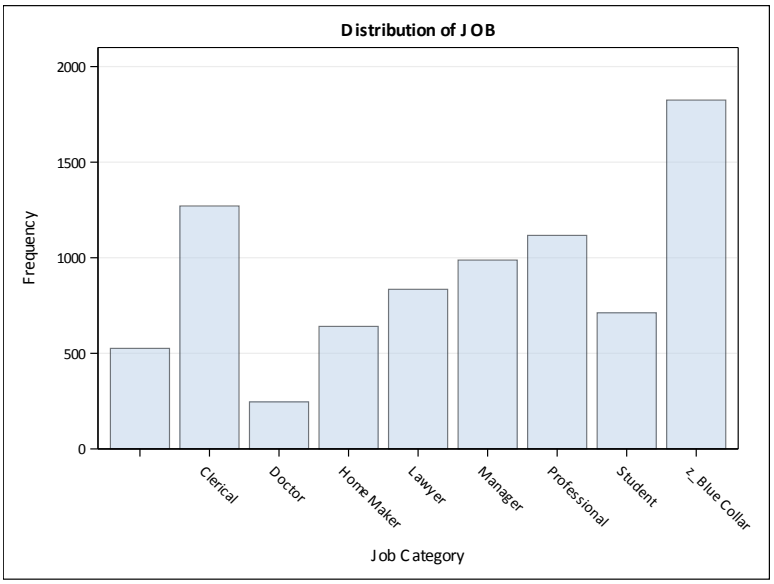

Figure 8: Distribution of JOB

After imputing all the missing jobs, we used the mean home value for each imputed job to impute the missing home values. For instance, we assigned a home value of $117,635.99 to any customer classified as having a Clerical job; we imputed a home value of $269,556.66 to any customer classified as a Doctor; etc. **Table 9** shows the home values for each imputed job.

Analysis Variable : HOME_VAL Home Value						
IMP_JOB	**N Obs**	**N**	**N Miss**	**Minimum**	**Maximum**	**Mean**
Clerical	1271	1204	67	0	308611.18	117635.99
Doctor	455	421	34	0	885282.34	269556.66
Home Maker	641	603	38	0	449553.34	86861.73
Lawyer	925	881	44	0	600591.12	201922.45
Manager	1060	1003	57	0	682633.59	199290.20
Professional	1191	1118	73	0	579056.49	192245.39
Student	712	667	45	0	236049.57	15385.90
z_Blue Collar	1906	1800	106	0	502062.72	156413.52

Table 9: HOME_VAL per IMP_JOB

The imputations we have described thus far remained constant for the three models we will discuss later in this report. The only additional data preparation we performed involved addressing outliers by capping the high values for some variables. As **Figure 9** shows, IMP_YOJ contained a couple outliers. The two highest values for IMP_YOJ were 23, so we decided to cap the data at the next highest value, 19, to take care of these outliers. We decided that we would also cap a few more variables with obvious outliers.

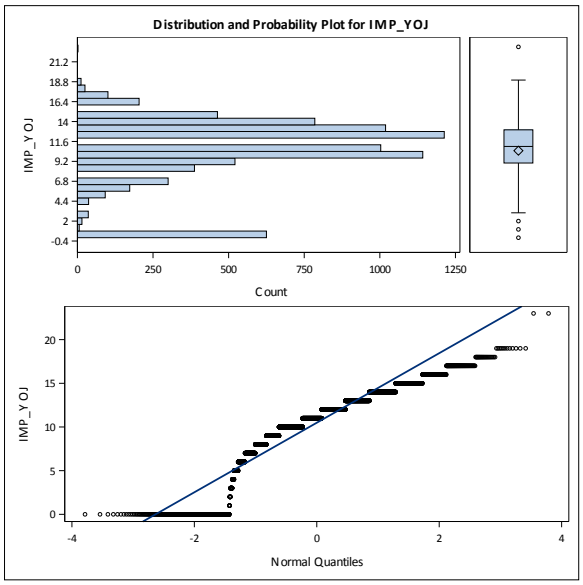

Figure 9: Distribution of IMP_YOJ

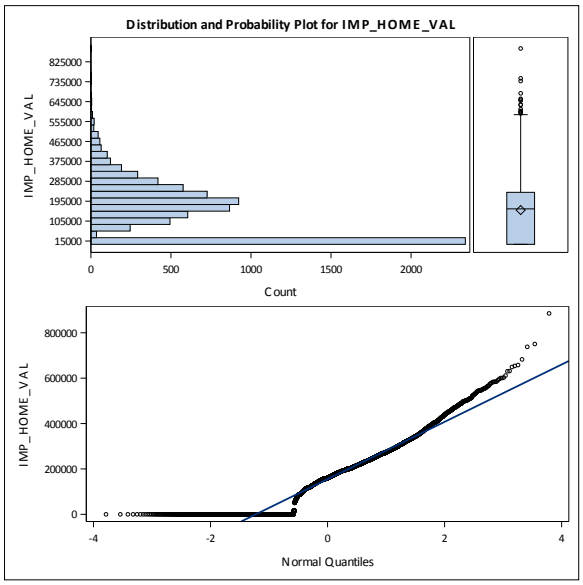

Figure 10: Distribution of IMP_HOME_VAL

In order to address the outliers present in **Figure 10**, we capped IMP_HOME_VAL at the 99th percentile. We decided to also cap TRAVTIME and BLUEBOOK, shown in **Figures 11 and 12**, at the 99th percentile in order to try to improve the predictive accuracy of the models.

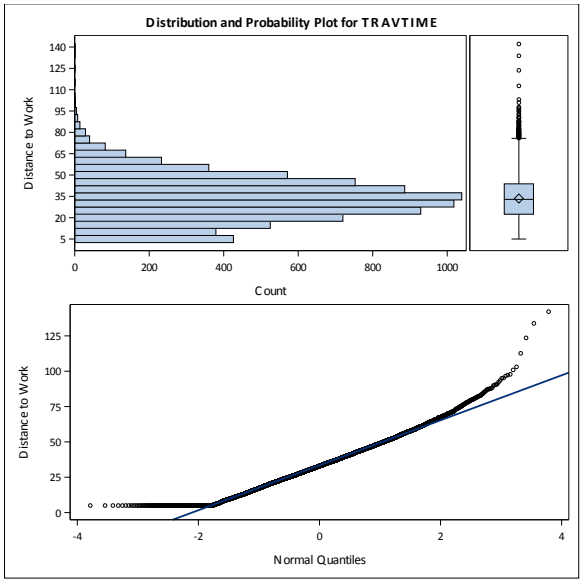

Figure 11: Distribution of TRAVTIME

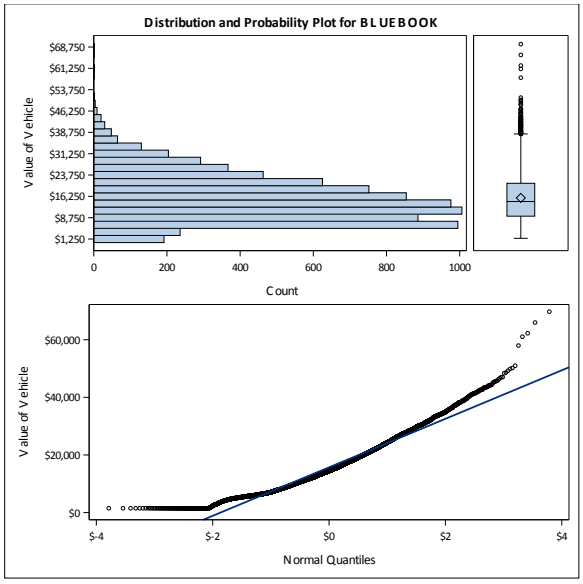

Figure 12: Distribution of BLUEBOOK

We examined the distribution and probability plots for all variables after we finished imputing missing values, but we determined that, aside from the variables mentioned above, the rest of the variables did not contain outliers that required fixing. We decided to not modify any more of the data set because doing so would be resource-intensive for possibly minimal gains, or it might even result in our model becoming over-fit to the training data set. The data preparation we described throughout this section is what we believe leads to the best predictive model.

Model Creation

After exploring the data and preparing them so that they would be ready for modeling, we began selecting variables to include in the models. In total we built approximately six distinct models, but we will focus on just three in this report as several of the models used similar variable selection techniques with only slight differences in data preparation or variable inclusion, which thus led to very similar levels of predictive accuracy.

For the first model, Model A, we used only the data set where we had imputed missing values; we did not use the data set that capped outliers. We used the stepwise automated variable selection technique in order to decide which variables to include in the model. The formula for Model A for predicting car crashes is:

YHAT = - 1.3968
- 0.7140*(CAR_TYPE in ("Minivan"))
- 0.1013*(CAR_TYPE in ("Panel Truck"))
- 0.1779*(CAR_TYPE in ("Pickup"))
+ 0.2551*(CAR_TYPE in ("Sports Car"))
- 0.0781*(CAR_TYPE in ("Van"))
+ 0.8061*(CAR_USE in ("Commercial"))
+ 0.00687*(EDUCATION in ("<High School"))
- 0.4219*(EDUCATION in ("Bachelors"))
- 0.3762*(EDUCATION in ("Masters"))
- 0.2939*(EDUCATION in ("PhD"))
+ 0.1463*(IMP_JOB in ("Clerical"))
- 0.4179*(IMP_JOB in ("Doctor"))
+ 0.0324*(IMP_JOB in ("Home Maker"))
- 0.0900*(IMP_JOB in ("Lawyer"))
- 0.7489*(IMP_JOB in ("Manager"))
- 0.0766*(IMP_JOB in ("Professional"))
- 0.0161*(IMP_JOB in ("Student"))
- 0.4710*(MSTATUS in ("Yes"))
- 0.4616*(PARENT1 in ("No"))
- 0.8952*(REVOKED in ("No"))
+ 2.3863*(URBANICITY in ("Highly Urban/ Urban"))
- 0.00002*BLUEBOOK
+ 0.1973*CLM_FREQ
- 1.34E-06*IMP_HOME_VAL
- 3.25E-06*IMP_INCOME
+ 0.4172*KIDSDRIV
+ 0.1145*MVR_PTS
- 0.00001*OLDCLAIM
- 0.0556*TIF
+ 0.0145*TRAVTIME

We capped the YHAT values at -999 and 999 to optimize SAS's performance, and then we calculated the predicted car crashes, P_TARGET_FLAG, with the following equation:

P_TARGET_FLAG = exp(YHAT)/(1+exp(YHAT))

For the most part this formula makes sense. Positive coefficients indicate a customer is a more dangerous driver, while negative coefficients indicate a customer is less risky. For instance, it seems logical that people who drive a Sports Car are considered more dangerous than people who drive other types of cars. The other car types are more utilitarian, while a sports car is more for leisure. Drivers are more likely to speed and take driving risks in a sports car than they are if they are driving a minivan with children in the backseat or hauling material in a pickup truck. It also makes sense that customers who drive for commercial purposes have a greater risk of crashing since they drive more often than customers who do not drive commercially.

The coefficients also make sense for EDUCATION. The only risky coefficient is <High School, while all the customers with higher education are indicative of less risk. This supports the theory that more educated people drive safer. The coefficients for JOB, however, are not quite as intuitive as the coefficients for some of the other categorical variables. In theory, people with white-collar jobs are safer drivers. The coefficients for Doctor, Lawyer, Manager, and Professional all support this. Clerical jobs are also white-collar, but the coefficient is positive and indicative of the customer being a greater risk. Perhaps the fact that people with clerical jobs generally earn far less than people with the other white-collar jobs listed is a reason why customers with clerical jobs are considered more dangerous drivers. The fact that a Student is considered a safe driver is also counterintuitive since students are typically younger people and thus are assumed to be riskier drivers, but perhaps there is more information to uncover.

When looking at the formula, the coefficients for MSTATUS and PARENT1 seem reasonable. The negative coefficients support the theory that married people are safer drivers, and the fact that single parents have more at stake if they get into a car crash supports why they might be more cautious drivers.

The formula also logically suggests that customers who have not had their license revoked in the past 7 years are considered safer drivers, while customers who live or work in Highly Urban or Urban areas are at greater risk for a car crash. Given that cities have a higher population density, it makes sense that there is a greater risk of crashing while driving in urban areas.

With the exception of OLDCLAIM, the rest of the coefficients are intuitive based on commonly held beliefs. It is more logical for a driver's risk to increase than decrease if they have more claims from the past 5 years. Further investigation into the counterintuitive coefficients is necessary before deploying this model.

While the results for Model A are overall logical and easy to interpret, we decided to try creating a better model with even more intuitive coefficients. For Model B, we prepared the data with the same imputations we used in Model A, but we also capped a few of the variables with outliers. We capped IMP_YOJ to exclude the two highest values, and we capped IMP_HOME_VAL, TRAVTIME, and BLUEBOOK at the 99th percentile to try to mitigate the effects the outliers we detected in our data exploration stage might have on the model. Just as in Model A, we used stepwise variable selection in Model B. The formula for Model B is:

```
YHAT =      - 1.3892
            - 0.7099*(CAR_TYPE in ("Minivan"))
            - 0.0827*(CAR_TYPE in ("Panel Truck"))
            - 0.1766*(CAR_TYPE in ("Pickup"))
            + 0.2555*(CAR_TYPE in ("Sports Car"))
            - 0.0646*(CAR_TYPE in ("Van"))
            + 0.8075*(CAR_USE in ("Commercial"))
```

+ 0.00557*(EDUCATION in ("<High School"))
- 0.4198*(EDUCATION in ("Bachelors"))
- 0.3760*(EDUCATION in ("Masters"))
- 0.2978*(EDUCATION in ("PhD"))
+ 0.1437*(IMP_JOB in ("Clerical"))
- 0.4180*(IMP_JOB in ("Doctor"))
+ 0.0287*(IMP_JOB in ("Home Maker"))
- 0.0855*(IMP_JOB in ("Lawyer"))
- 0.7469*(IMP_JOB in ("Manager"))
- 0.0752*(IMP_JOB in ("Professional"))
- 0.0278*(IMP_JOB in ("Student"))
- 0.4603*(MSTATUS in ("Yes"))
- 0.4630*(PARENT1 in ("No"))
- 0.8943*(REVOKED in ("No"))
+ 2.3843*(URBANICITY in ("Highly Urban/ Urban"))
- 0.00002*IMP_BLUEBOOK
+ 0.1968*CLM_FREQ
- 1.43E-06*IMP_HOME_VAL
- 3.18E-06*IMP_INCOME
+ 0.4182*KIDSDRIV
+ 0.1143*MVR_PTS
- 0.00001*OLDCLAIM
- 0.0556*TIF
+ 0.0151*IMP_TRAVTIME

As in Model A, we capped the YHAT values at -999 and 999 to optimize SAS's performance, and then we calculated the predicted car crashes, P_TARGET_FLAG, with the following equation:

P_TARGET_FLAG = exp(YHAT)/(1+exp(YHAT))

From looking at the formula above, we can see that the signs of the coefficients for Model B are the same as in Model A. Only the values of the coefficients changed slightly. The analysis for Model A holds true for Model B, so we would still like to try to create a model where there are no counterintuitive coefficients. In order to try to create a better, more intuitive model, we decided to keep the same data set we created for Model B that contains imputations and capped outliers, and we used a decision tree to select the variables for Model C. **Figure 13** shows the output from the decision tree we created using R. All the variables are the same as in the data set for Model B, but we did have to change some variables from being stored as currency values to being stored as numeric values. Variables with a "_n," such as OLDCLAIM_n, indicate that we converted them from currency to numeric.

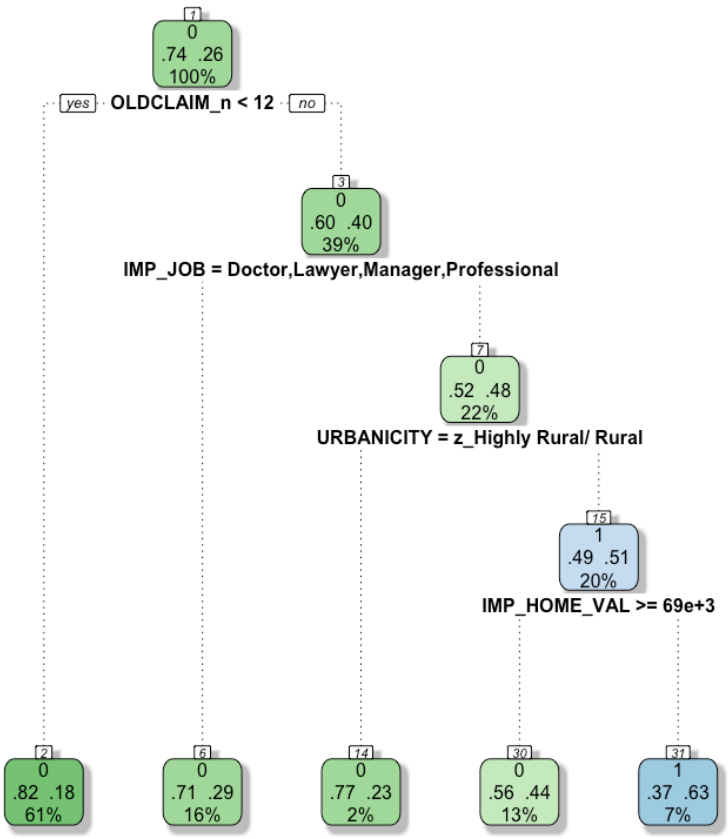

Figure 13: Decision Tree for Model C

Each leaf on the tree in **Figure 13** has three rows of data. The top row indicates the TARGET_FLAG class that makes up the majority of the data in the leaf, the middle row indicates the probability of each class of TARGET_FLAG of that leaf, and the bottom row indicates the percentage of observations from the entire data set present in each leaf. The variables beneath the leaves determine how the tree branches, or splits, from each leaf. For instance, the top leaf, or root, shows that 100 percent of the data is being considered in the tree, and that 74 percent of the data is TARGET_FLAG = 0, while 26 percent of the data is TARGET_FLAG = 1. This is consistent with the frequencies we saw in **Table 1** at the beginning of the report. Because most of the customers in the data set do not crash their cars, the leaf has a “0” in the top row. When we consider the variable OLDCLAIM_n, we see that the majority of observations in the data set have fewer than 12 claims. Customers with fewer than 12 claims branch to the left, while customers with 12 claims or more branch to the right. In the next leaf labeled with IMP_JOB = Doctor, Lawyer, Manger, Professional, we see that 39 percent of the data set is being considered. TARGET_FLAG is 0 for 60 percent of the data being considered, while TARGET_FLAG is 1 for remaining 40 percent of the data. If the customer is not a Doctor, Lawyer, Manger, or Professional, the tree branches to the right and then considers the URBANCITY variable and then the IMP_HOME_VAL variable. The right branches of the tree all support observations and assumptions we have already made about the data. Customers who have a high number of claims, who do not have high-paying white-collar jobs, who live or work in urban areas, and who have a low home value are more likely to get into car accidents. The bottom leaf farthest to the right indicates that while only 7 percent of the entire data set is represented in that leaf, 63 percent of the data being considered has TARGET_FLAG = 1. Given that the tree in **Figure 13** makes logical sense, we decided to create a model using just the variables listed in the tree. The formula for Model C is:

YHAT = - 2.0702
- 0.2079*(IMP_JOB in ("Clerical"))
- 1.1322*(IMP_JOB in ("Doctor"))
- 0.3468*(IMP_JOB in ("Home Maker"))
- 0.9459*(IMP_JOB in ("Lawyer"))
- 1.3787*(IMP_JOB in ("Manager"))
- 0.6326*(IMP_JOB in ("Professional"))
- 0.1073*(IMP_JOB in ("Student"))
+ 2.1996*(URBANICITY in ("Highly Urban/ Urban"))
+ 0.00002*OLDCLAIM
- 3.34E-06*IMP_HOME_VAL

As in Models A and B, we capped the YHAT values at -999 and 999 to optimize SAS's performance, and then we calculated the predicted car crashes, P_TARGET_FLAG, with the following equation:

P_TARGET_FLAG = exp(YHAT)/(1+exp(YHAT))

Several of the signs for the coefficients are different in Model C than in the previous two models. The coefficient signs for Clerical and Home Maker jobs, and the sign for OLDCLAIM, all switched. Only the sign change for OLDCLAIM makes sense since more claims indicate a riskier driver. While some of the signs are intuitive, others are not. The interpretability of Model C is not an improvement on Models A and B. Given that some of the signs switched, we would want to investigate the data further before we would consider deploying this model.

In the next section we will discuss which of the three models explored above we consider the best model for predicting car crashes.

Model Selection

While examining all the models we created, we chose Model B as the best model. The criteria we used for selecting the best model are AIC, ROC Curve AUC, and KS statistic. We compared the three models using all these metrics and decided that the model that scored the best in all three or at least two of the three metrics would be a sound model to deploy. **Table 10** shows the AIC values for each model. Model B has the lowest AIC value, but Model A is only slightly higher. Given that Model C included a different combination of variables and had some coefficients with the opposite signs as coefficients in Models A and B, we are not surprised that the AIC value for Model C is much worse.

Model Fit Statistics			
	Model A	Model B	Model C
Criterion	Intercept and Covariates	Intercept and Covariates	Intercept and Covariates
AIC	7368.186	7360.775	8208.54
SC	7585.406	7577.996	8285.618
-2 Log L	7306.186	7298.775	8186.54

Table 10: AIC Comparison for Models A, B, and C

Because AIC alone is not a completely reliable indicator of a good model, we decided to examine the ROC Curves AUC and KS statistics for each model as well.

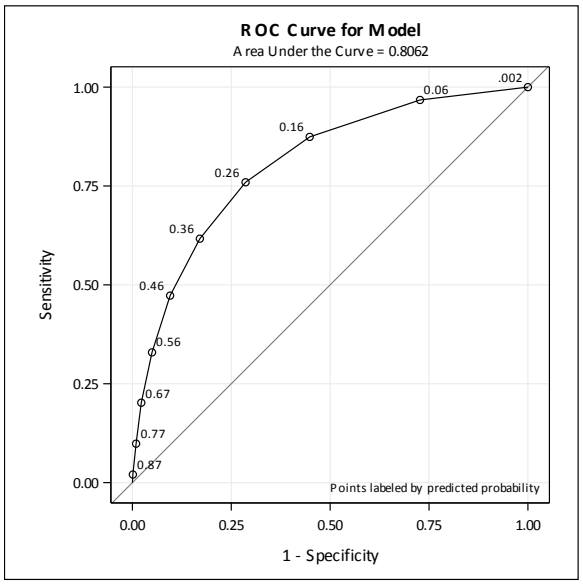

Figure 14: ROC Curve for Model A

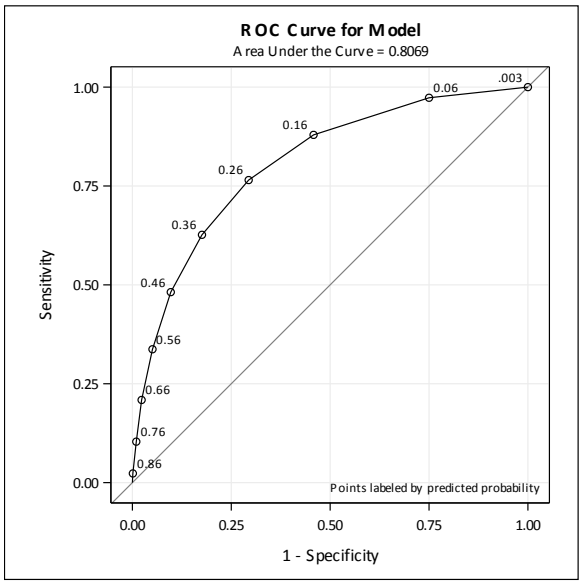

Figure 15: ROC Curve for Model B

In **Figures 14 and 15**, we can compare the ROC Curves for Models A and B. Like their AIC values, their ROC Curves are also very similar. Both exhibit a big bow that has all the lift in the beginning, which are characteristics of good models. If we make the cutoff value at the 0.26 node (i.e., customers with a probability score of 0.26 or greater), we find about 75 percent of the people who do crash their cars, but we also have a false positive rate of about 28 percent. The ROC Curves for Models A and B are nearly identical and complicate deciding which is the better model by visuals alone. We see that the AUC value for Model B is slightly higher than for Model A: 0.8069 v. 0.8062, respectively. This indicates that Model B is a slightly stronger model than Model A. When we calculate the KS statistic for Models A and B, we end up with 47 percent for both. **Table 11** shows the ROC Curve data for Models A and B. The cutoff values for the models differ slightly, but they result in the same Cumulative TP Rate, Cumulative FP Rate, and KS Difference.

GROUP	CUTOFF Model A	CUTOFF Model B	Cumulative TP Rate	Cumulative FP Rate	KS Difference
0	1.00000	1.00000	0%	0%	0%
1	0.62341	0.62894	28%	4%	24%
2	0.47615	0.48166	48%	10%	39%
3	0.36949	0.37590	63%	18%	45%
4	0.28698	0.29277	75%	28%	47%
5	0.21324	0.21766	84%	38%	46%
6	0.15057	0.15436	90%	49%	41%
7	0.10342	0.10532	94%	61%	33%
8	0.06495	0.06682	97%	74%	23%
9	0.03583	0.03674	99%	87%	13%
10	0.00285	0.00298	100%	100%	0%

Table 11: ROC Curve Data and KS Statistics for Models A and B

When we examine the ROC Curve for Model C in **Figure 16,** we see that it differs from the ROC Curves for Models A and B. The bow is not as full as in **Figures 14 and 15**, and the lift happens more gradually. These are indications that the model is not as good as Models A and B.

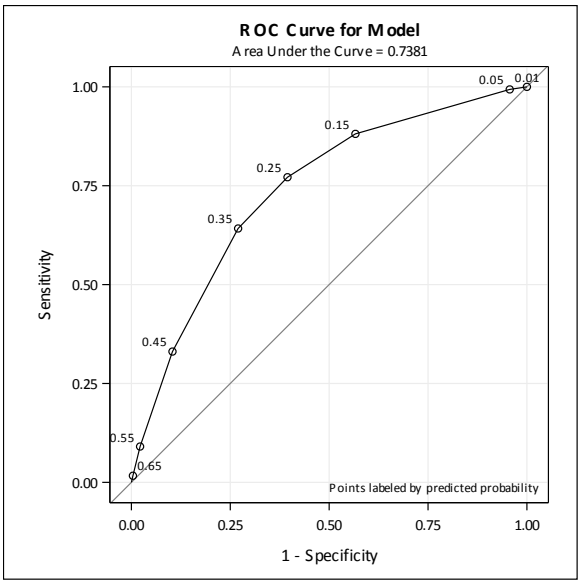

Figure 16: ROC Curve for Model C

If we make the cutoff value at the 0.25 node, we find about 75 percent of the people who do crash their cars, but we also have a false positive rate of about 38 percent. These are just ballpark estimates from glancing at the plot, but the false positive rate for Model C is far worse than for Models A and B given that we find about the same percentage of people who crash their cars. The AUC value is also much lower for Model C at 0.7381. These are indications that the model does not perform as well as the other two models.

GROUP	CUTOFF Model C	Cumulative TP Rate	Cumulative FP Rate	KS Difference
0	1.00000	0%	0%	0%
1	0.50629	20%	6%	15%
2	0.41324	39%	13%	26%
3	0.37679	55%	21%	34%
4	0.32361	68%	30%	38%
5	0.24268	78%	40%	38%
6	0.18149	85%	51%	34%
7	0.12574	91%	62%	29%
8	0.09307	96%	74%	21%
9	0.06258	98%	87%	11%
10	0.00921	100%	100%	0%

Table 12: ROC Curve Data and KS Statistic for Model C

In **Table 12**, we can see that the Cutoff values, Cumulative TP Rate, Cumulative FP Rate, and KS Difference are quite different from Models A and B. The KS statistic for Model C, 38 percent, is 9

percent lower than the KS statistic of the other models. Again, these observations confirm that Model C is far from the best model.

Although Models A and B had the same KS statistic and very similar AIC values and ROC Curve AUC values, Model B scored slightly better in the criteria we used for determining the best model. Its coefficients are also mostly intuitive, but because the model does contain a couple of counterintuitive coefficients, we would like to consult an auto insurance expert or investigate the data more deeply prior to deploying this model.

Conclusion

Our goal with this project was to create a model that best predicted which customers are most likely to crash their cars. We created several models based on the Insurance data set provided. We prepared the data using various methods, including imputation, binning, and capping. We then built the models using a couple of techniques, such as stepwise automated variable selection and decision trees. Our best model surpassed the other models by having the lowest AIC, highest AUC, and a higher or equally high KS statistic. Although a couple coefficients in our best model are counterintuitive, the model performed well. We would like to further investigate the counterintuitive coefficients provided that we have sufficient time to do so before the model must be deployed. In the meantime, we are confident that the model we created is suitable for predicting car crashes.