Chelsea Clinger

411 Generalized Linear Methods 56

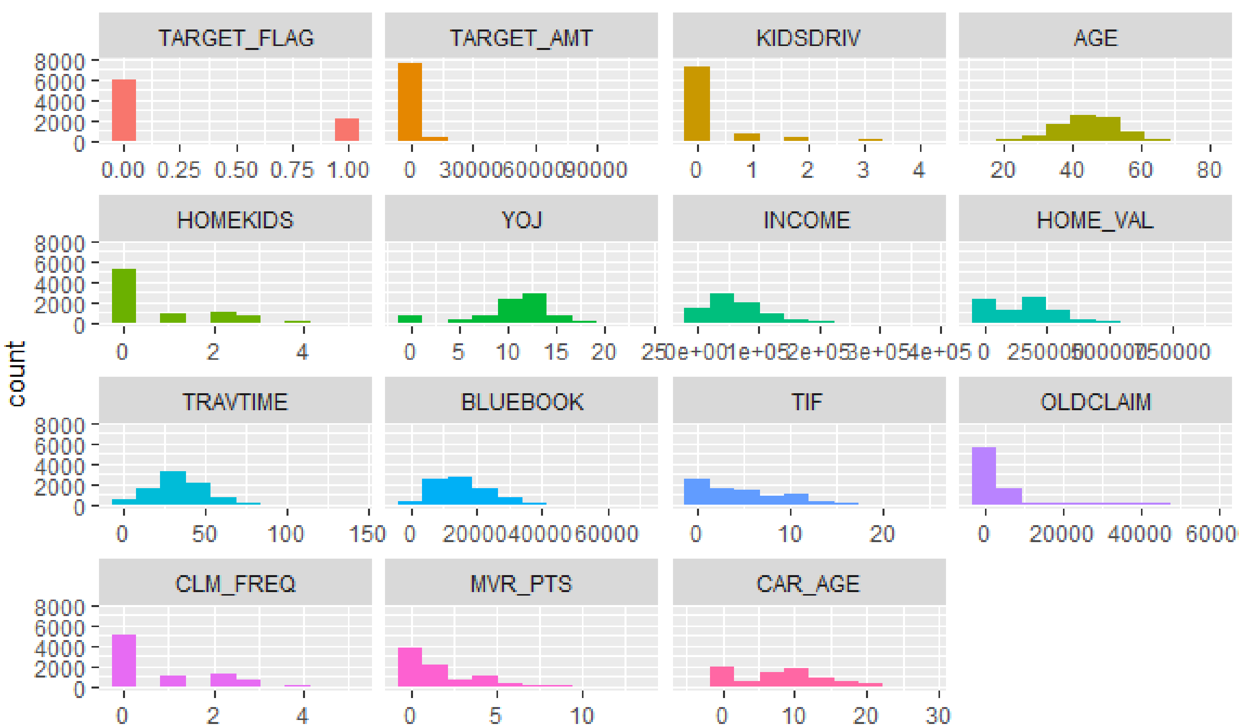
Unit 2

Introduction

Any insurance company will have a large amount of information on any particular client including information about the client themselves as well as the car that they are responsible for insuring. This information can be compiled and used in many different types of analysis. In this assignment two predictive models will be created using logistic regression to determine both the likelihood that a given person will file an insurance claim in the future and if so, how much that claim will have to pay out.

Section 1 – Data Exploration

Before beginning any analysis on the variables, an initial view into each of them will be conducted. Looking first at the two target variables, we can see that most individuals do not experience an auto accident (0). It is also easy to quickly identify that many of the predictor variables contain 0 values. It will need to be determined if these values are correct 0s or not.



VARIABLE	PROBLEM	ADJUSTMENT
CAR AGE	There are negative values for age	Adjusted values lower than the bottom quartile to become equal to the bottom quartile
REVOKED	Yes and No are categorical	Changed Yes to 1 and No to 0

MSTATUS	Yes and No are categorical	Changed Yes to 1 and No to 0
CAR_USE	Private and Commercial are categorical	Changed Private to 1 and Commercial to 0
RED_CAR	Yes and No are categorical	Changed Yes to 1 and No to 0
URBANICITY	Options are categorical	Highly Urban/Urban = 1, Highly Rural/Rural = 0
CAR_TYPE	Some of the options fit into smaller categories	Minivan and Van = Van Pickup and Pallet Truck = Truck

The below table demonstrates how many 0s and NAs there are for each variable. YOJ, INCOME and HOME_VAL have a very high count of Nas that need to be resolved.

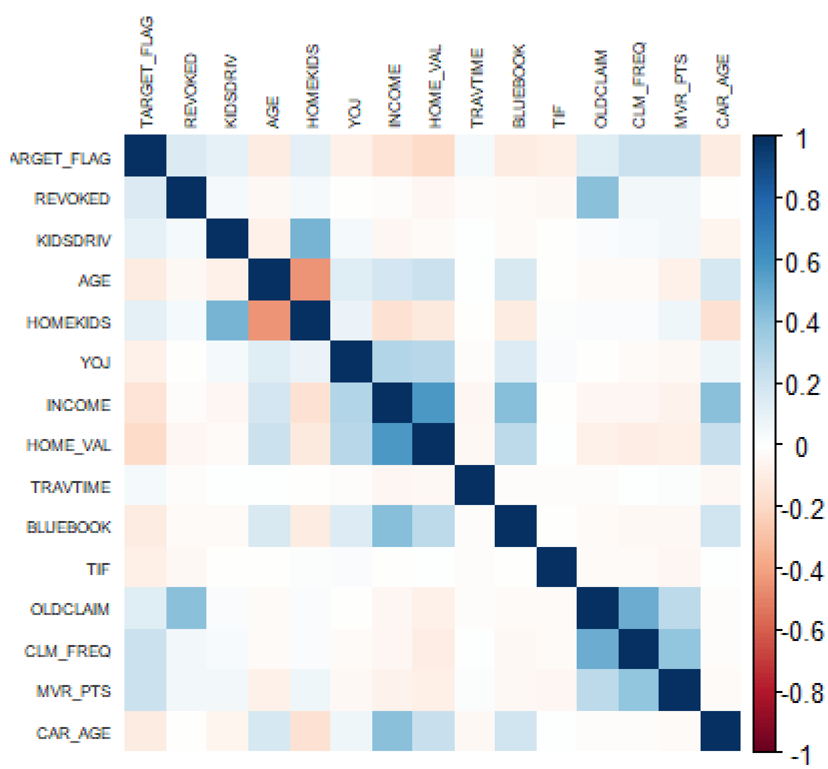
	variable	q_zeros	p_zeros	q_na	p_na	q_inf	p_inf	type	unique
1	INDEX	0	0.00	0	0.00	0	0	integer	8161
2	TARGET_FLAG	6008	73.62	0	0.00	0	0	integer	2
3	TARGET_AMT	6008	73.62	0	0.00	0	0	numeric	1949
4	KIDSDRIV	7180	87.98	0	0.00	0	0	integer	5
5	AGE	0	0.00	6	0.07	0	0	integer	60
6	HOMEKIDS	5289	64.81	0	0.00	0	0	integer	6
7	YOJ	625	7.66	454	5.56	0	0	integer	21
8	INCOME	615	7.54	445	5.45	0	0	numeric	6612
9	PARENT1	0	0.00	0	0.00	0	0	factor	2
10	HOME_VAL	2294	28.11	464	5.69	0	0	numeric	5106
11	MSTATUS	0	0.00	0	0.00	0	0	factor	2
12	SEX	0	0.00	0	0.00	0	0	factor	2
13	EDUCATION	0	0.00	0	0.00	0	0	factor	5
14	JOB	0	0.00	0	0.00	0	0	factor	9
15	TRAVTIME	0	0.00	0	0.00	0	0	integer	97
16	CAR_USE	0	0.00	0	0.00	0	0	factor	2
17	BLUEBOOK	0	0.00	0	0.00	0	0	numeric	2789
18	TIF	0	0.00	0	0.00	0	0	integer	23
19	CAR_TYPE	0	0.00	0	0.00	0	0	factor	6
20	RED_CAR	0	0.00	0	0.00	0	0	factor	2
21	OLDCLAIM	5009	61.38	0	0.00	0	0	numeric	2857
22	CLM_FREQ	5009	61.38	0	0.00	0	0	integer	6
23	REVOKED	0	0.00	0	0.00	0	0	factor	2
24	MVR_PTS	3712	45.48	0	0.00	0	0	integer	13
25	CAR_AGE	3	0.04	510	6.25	0	0	integer	30
26	URBANICITY	0	0.00	0	0.00	0	0	factor	2

After the missing values have been removed, we can calculate the correlation between variables.

The plot on the right illustrates a heat map of correlation between variables.

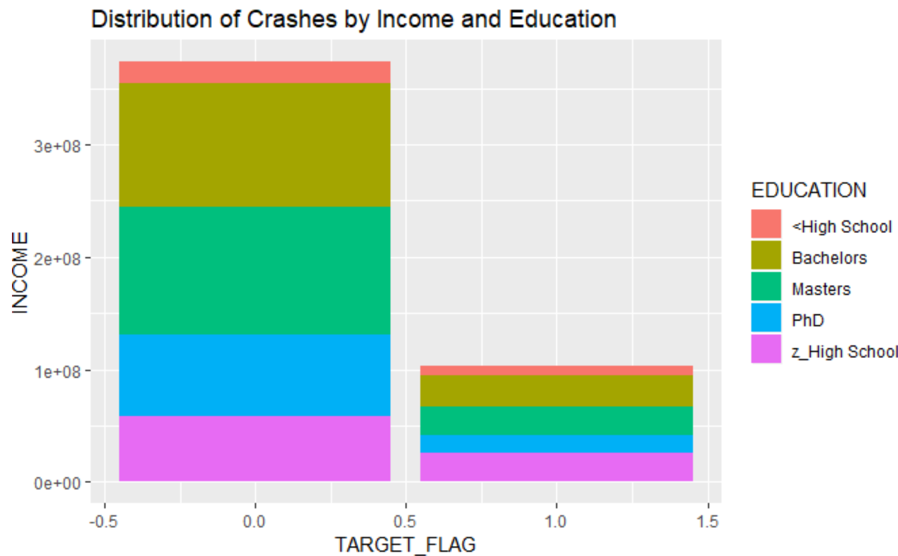
Initially looking to determine which variables are closely correlated with the target value, it is notable that claim frequency and previous traffic tickets have the highest positive influence while home value and income have the highest negative influence.

Some other correlations exist between age and number of kids (the younger you are the less likely you are to have children) and home value and income (the higher your income, the higher your home value).



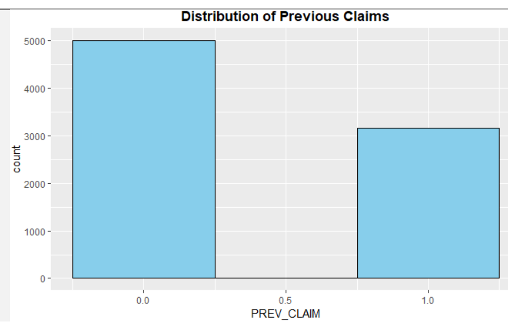
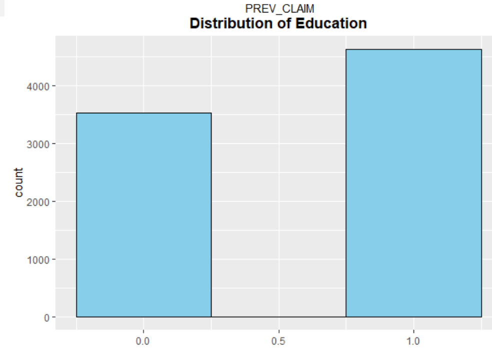
Since income and the target value had a noticeable negative relationship, it may be worthwhile to explore the relationship between the target value, income and education level.

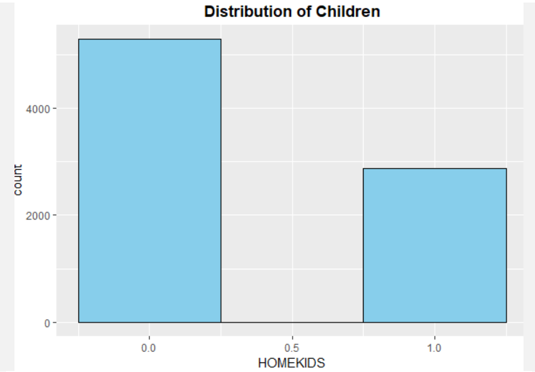
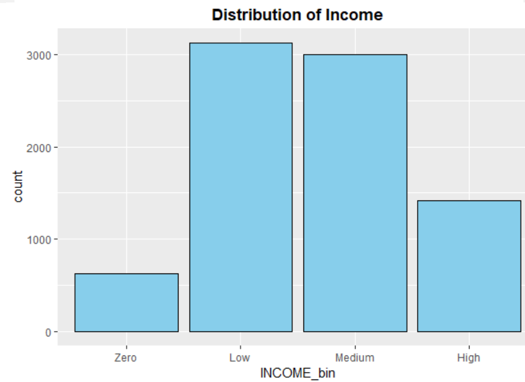
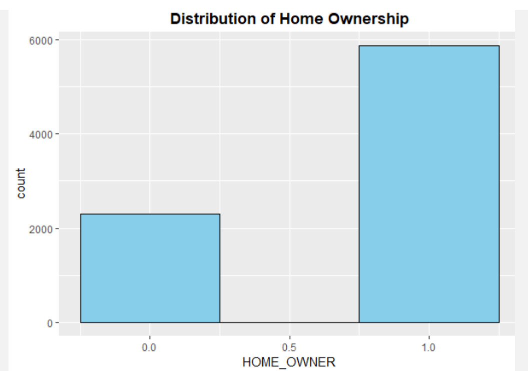
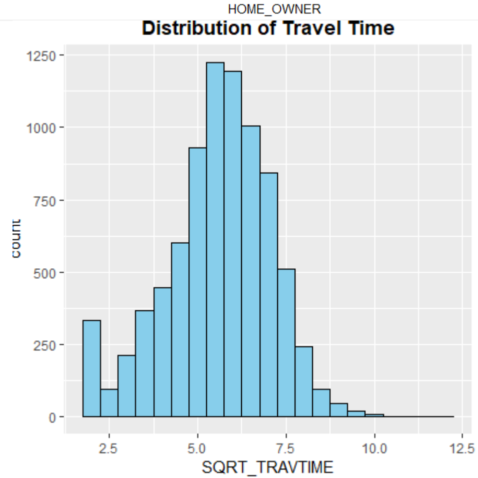
The graph on the right illustrates that while the majority of individuals who have had an accident have a higher education, a high percentage of those either still in high school or with just a high school diploma have caused accidents.

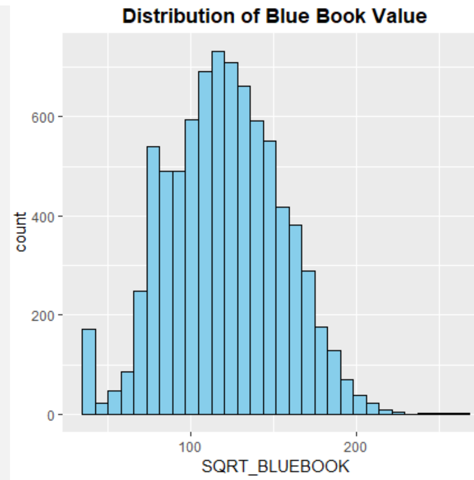
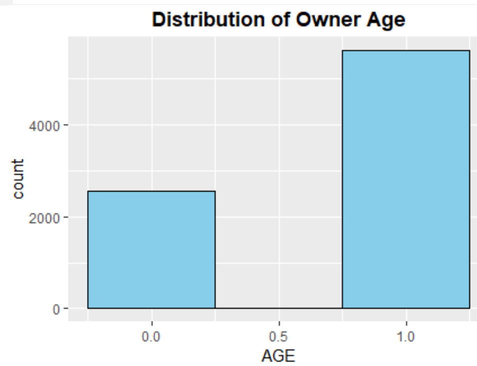


Section 2 – Data Preparation

Additional transformations of variables into binary variables will assist in creating a more significant correlation.

VARIABLE NAME	DESCRIPTION	CORRELATION BEFORE/AFTER	DISTRIBUTION
PREVIOUS CLAIM?	Indicates if an individual has had a previous claim in the past or not.	.21 -> .24	
EDUCATION	Indicates if education level is higher than high school or not **Highschool or lower:** 0 **Higher:** 1	N/A	

HOME_KIDS	Indicates rather if there are kids at home than how many	.11 -> .13	
INCOME	Categorize into Zero, Low, Medium and High. **Zero**: 0 **Low**: 1 – 50,000 **Medium**: 50,000 – 100,000 **High**: 100,000 – Inf	N/A	
HOME OWNERSHIP	It is already known home value has a large influence on the target value. **Home Val = 0**: 0 **Home Val>0:** 1	-0.18 -> -.15	
TRAVEL TIME	Change to square root	.04 -> .05	

BLUE BOOK	Change to square root	-.10 -> -.11	
AGE	Separate younger and older car owners <=40: 0 >40:1	-0.103 -> -0.117	

Section 3 – Model Creation

Model 1

The first model will be created with manually selected variables. The variables in the first model will include:

VARIABLE NAME	DESCRIPTION	ESTIMATE
REVOKED	If the license has been revoked in the past 7 years	.8434
AGE	Indicates if the owner is older or younger than 40	-0.2027
INCOME_BINLOW	Income between $1 and $50,000	-0.2929
INCOME_BINMEDIUM	Income between $50,000 and $100,000	-.2884
INCOME_BINHIGH	Income above $100,000	-.5581
SQRT_TRAVTIME	Distance to work	0.0973

SQRT_BLUEBOOK	Square root of the blue book value	-0.003
HOME_OWNER	If the owner also owns a home	-0.5524
HOMEKIDS	If the owner has children	.3584
PREV_CLAIM	If a previous claim has been filed	1.0476
EDUCATION	Education level if above high school or not	-0.3866

Out of the coefficients mentioned above there are a few that stand out as interesting. The various income bins indicate that having a "low" income lowers your odds of an auto accident (compared to not having any income) by .29 while increasing your income to "medium" lowers your odds by an additional .29 and .56 more when increasing to "high". This indicates to me that income, and especially how high your income is, plays a large role in your odds of having to file a claim. It is also worth noting how significant the previous claim value is. The odds of an individual having to file a claim are 1.05 higher if a claim has been filed previously.

Model 2

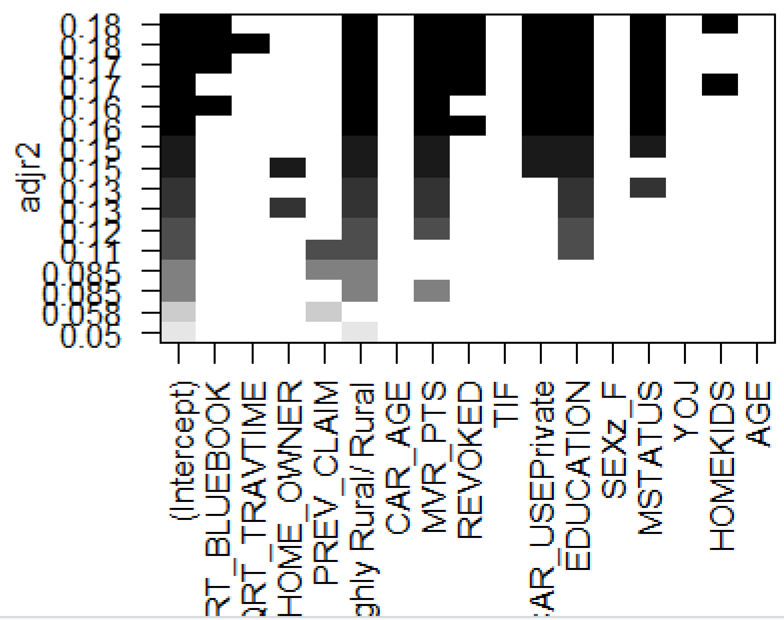
The next model will have the variables selected using the stepwise method. This method removed the red car indicator, owner gender, age of the car and years at job as variables that did not lower the AIC.

VARIABLE NAME	DESCRIPTION	ESTIMATE
SQRT_BLUEBOOK	Square root of the blue book value	-0.00592
SQRT_TRAVTIME	Distance to work	0.170913
HOME_OWNER	If the owner also owns a home	-0.33077
INCOME_BINLOW	Income between $1 and $50,000	-0.67351
INCOME_BINMEDIUM	Income between $50,000 and $100,000	-0.86092
INCOME_BINHIGH	Income above $100,000	-1.17181
PREV_CLAIM	If a previous claim has been filed	0.418946
URBANICITYZ_HIGHLY RURAL/RURAL	Indicates that a car owner lives in a rural area opposed to urban	-2.30509
MVR_PTS	Motor vehicle record points	0.098428
REVOKED	If the license has been revoked in the past 7 years	0.736419
CAR_TYPEPANEL TRUCK	Car being insured is a panel truck	0.604668

CAR_TYPEPICKUP	Car being insured is a pickup truck	0.560168
CAR_TYPESPORTS	Car being insured is a sports car	0.93131
CAR_TYPEVAN	Car being insured is a van	0.636462
CAR_TYPEZ_SUV	Car being insured is a SUV	0.708321
TIF	Time with insurance company	-0.05472
CAR_USEPRIVATE	Indicates a car is used for private purposes	-0.77187
JOBDOCTOR	Indicates the car owner’s job is a doctor	-0.38171
JOBHOME MAKER	Indicates the car owner’s job is a home-maker	-0.25845
JOBLAWYER	Indicates the car owner’s job is a lawyer	0.080747
JOBMANAGER	Indicates the car owner’s job is a manager	-0.64794
JOBPROFESSIONAL	Indicates the car owner’s job is a professional	-0.06862
JOBSTUDENT	Indicates the car owner’s job is a student	-0.39077
JOBZ_BLUE COLLAR	Indicates the car owner’s job is blue collar	-0.46137
EDUCATION	Education level if above high school or not	-0.59745
MSTATUS	Indicates if married or single	0.483711
HOMEKIDS	If the owner has children	-0.14291
AGE	Indicates if the owner is older or younger than 40	-0.46137

Model 3

For the final model all subsets regression will be conducted on the previously utilized predictor variables.



Utilizing the best subsets, the variables SQRT_Bluebook, Urbanicity Highly Rurual/Rural, MVR_PTS, Revoked, Car_Use Private, Education, Mstatus and Homekids are selected.

VARIABLE NAME	DESCRIPTION	ESTIMATE
SQRT_BLUEBOOK	Square root of the bluebook value	-0.0091
URBANICITY HIGHLY RURAL/RURAL	The location the insured car is primarily kept	-2.1222
MVR_PTS	Motor vehicle record points	.1605
CAR USE PRIVATE	The primary usage of the car is by the owner for personal needs	.7658
EDUCATION	The schooling level of the client	-0.8228
MSTATUS	The martial status of the client	-0.7127
HOME KIDS	If the client has children or not	.5538

Model 4

At this point it appears that Model 2 utilizing the stepwise method may be the most accurate model. Therefore an attempt to adjust some of the variables to enhance this model will be tried. Before continuing, both the income variable and the home value will be adjusted to be the square roots of their original values.

VARIABLE NAME	DESCRIPTION	ESTIMATE
SQRT_BLUEBOOK	Square root of the blue book value	-0.00544
SQRT_TRAVTIME	Distance to work	0.1693

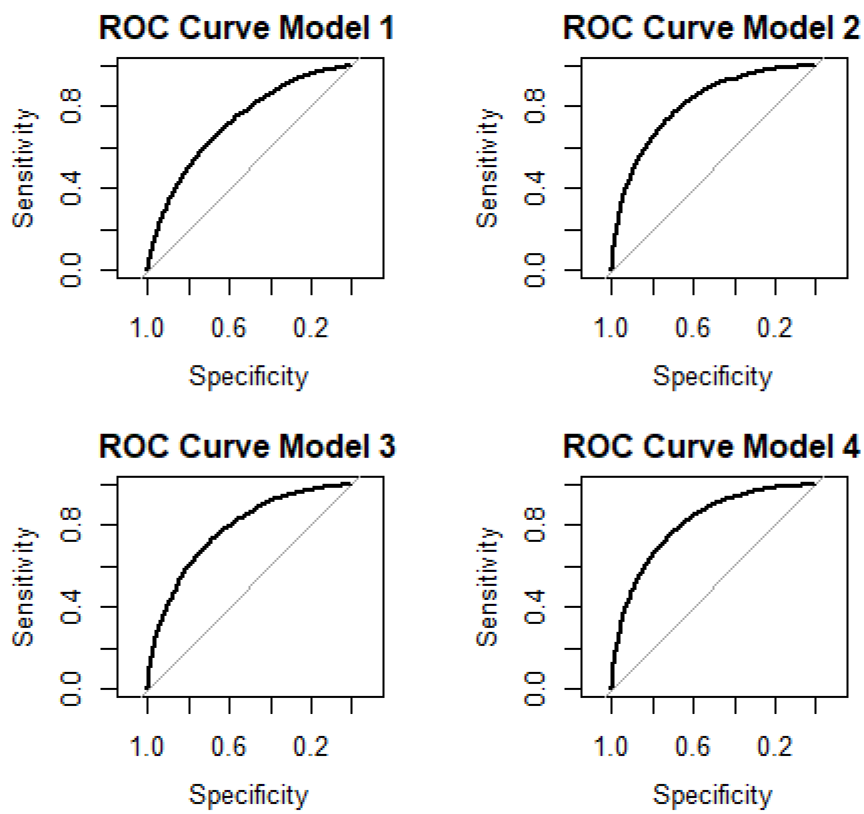
SQRT_HOMEVAL	Square root of the value of the home	-0.0008
SQRT_INCOME	Square root of the client's income	-0.0029
PREV_CLAIM	If a previous claim has been filed	0.4116
URBANICITYZ_HIGHLY RURAL/RURAL	Indicates that a car owner lives in a rural area opposed to urban	-2.3010
MVR_PTS	Motor vehicle record points	0.09983
REVOKED	If the license has been revoked in the past 7 years	0.7354
CAR_TYPEPANEL TRUCK	Car being insured is a panel truck	0.6178
CAR_TYPEPICKUP	Car being insured is a pickup truck	0.5530
CAR_TYPESPORTS	Car being insured is a sports car	0.9354
CAR_TYPEVAN	Car being insured is a van	0.6422
CAR_TYPEZ_SUV	Car being insured is a SUV	0.7066
TIF	Time with insurance company	-0.0550
CAR_USEPRIVATE	Indicates a car is used for private purposes	-0.7969
JOBDOCTOR	Indicates the car owner's job is a doctor	-0.2746
JOBHOME MAKER	Indicates the car owner's job is a home-maker	-0.2669
JOBLAWYER	Indicates the car owner's job is a lawyer	0.14551
JOBMANAGER	Indicates the car owner's job is a manager	-0.5826
JOBPROFESSIONAL	Indicates the car owner's job is a professional	-0.0067
JOBSTUDENT	Indicates the car owner's job is a student	-0.4403
JOBZ_BLUE COLLAR	Indicates the car owner's job is blue collar	0.06677
EDUCATION	Education level if above high school or not	-0.3895
MSTATUS	Indicates if married or single	-.5938
HOMEKIDS	If the owner has children	.4809
AGE	Indicates if the owner is older or younger than 40	-0.1186

When observing the changes in estimates from Model 2 to Model 4, the primary adjustments can be observed in the job variables. For example, the odds of an accident for a lawyer increased from .080 to

.146 and blue color workers increased from -.4614 to .06677. Another notable shift is the indicator if the client has children which increased significantly from -.1429 to .4809.

There are no coefficients in this model that don't align to the target value. For example, having prior tickets makes the insured significantly more risky while other factors such as having a higher paying job is less risky.

Section 4 – Model Selection



MODEL	AIC	DEVIANCE RATIO	R-SQUARED	ROC	KS STAT
MODEL 1	8412.12	.8907	.1093	.7245	.3249
MODEL 2	7380.86	.7775	.2224	.8112	.4637
MODEL 3	7766.8	.8228	.1772	.7802	.4145
MODEL 4	7369.4	.7767	.2233	.8118	.4668

Illustrated above are the various ROC curves for each models. The ROC curve for the first model is the most significantly different than the others with the fourth model indicating the better curve. Models 2 and 4 indicate a good accuracy of distinguishing between potential claims while 1 and 3 are definitively less accurate. Furthermore the additional statistics; KS Stat, AIC, Deviance and R-Squared, all indicate that the fourth model is the most reliable.