Unit 2 Assignment: “Insurance Logistic Regression”

Alan Kessler, MSDS 411, Section 56

Intro

The goal of the project is to use historical data to predict both the probability that an auto insurance customer has a claim and in the case of a claim, to determine the loss amount. The data contains information about the customer including previous claim history, individual characteristics such as age, and financial information like income.

To predict the probability of a claim, multiple logistic regression models are fit to the data and evaluated. Linear models are fit to the loss data as well. Final models are selected based on fit statistics and interpretability considerations. The final models contain a majority of the data elements included, with missing values imputed and several transformations applied.

While the topic of the assignment relates to Auto Insurance, the purpose of the assignment is for Northwestern University’s School of Professional Studies MSDS 411. It does not constitute Actuarial Services as defined in the Actuarial Standards of Practice No. 41. The variables and data used do not contemplate other Actuarial Standards of Practice. All variables are assumed to be acceptable from an implementation stand point.

Bonus

For bonus points, I used decision trees to assess the potential for interactions and applied probit regression as one of the models for the binary target.

For optional bonus points, I incorporated impact/likelihood encoding to enhance the modeling process. Additionally, I used a gamma error distribution for the claim severity model.

Data Exploration

The first step in exploring the data is to analyze the distributions of both the binary and continuous targets. For the binary target, approximately a quarter of the 8,161 observations have a claim. While these classes are unbalanced, claim cases are common enough that sampling techniques on this small data set are unnecessary.

The continuous target has a heavy tail characteristic of a loss distribution as shown in Figure 1. Taking the log of the data does not normalize the sample, so it is important to consider other error distributions such as the Gamma distribution.

Prior to further exploration, currency variables require transformation to change their format to numeric. Additionally, binary categorical variables are transformed to be numeric indicators for the purpose of clearer interpretation.

Figure 1 – QQ-Plots of Continuous Target

Next, missing values are analyzed. Figure 2 shows the percentage of observations missing in the training data for variables with at least one missing value. Missing values are observed to be relatively rare. In the case of driver age, a missing value indicator is not necessary due to the rarity of that value being missing.

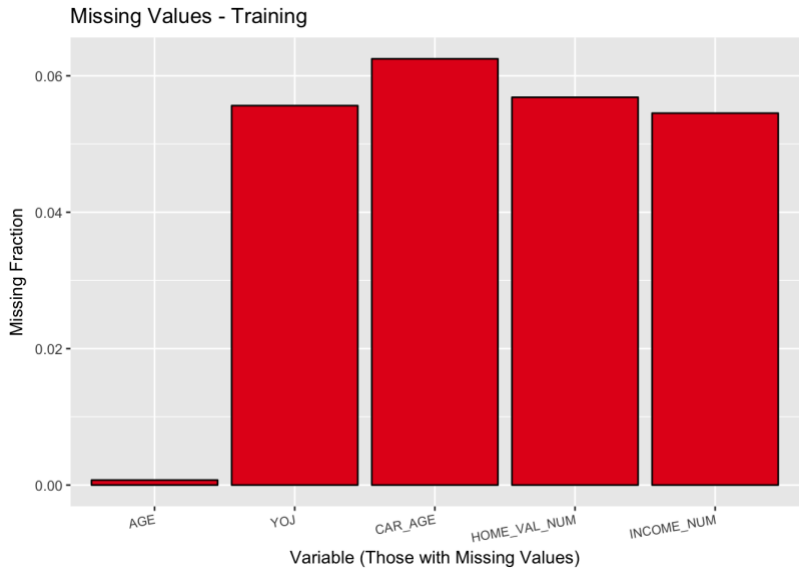

Figure 2 – Missing Value Frequency

In the analysis of predictor variables, continuous predictors are considered as a group. While many are discrete in nature, they can be treated as numeric for modeling purposes. Figure 3 shows the distributions of both training and test data, indicating that they are likely from the same population. Some of the variables appear skewed. This suggests keeping a close eye on the relationship with the target for a linear relationship and applying transformations as needed.

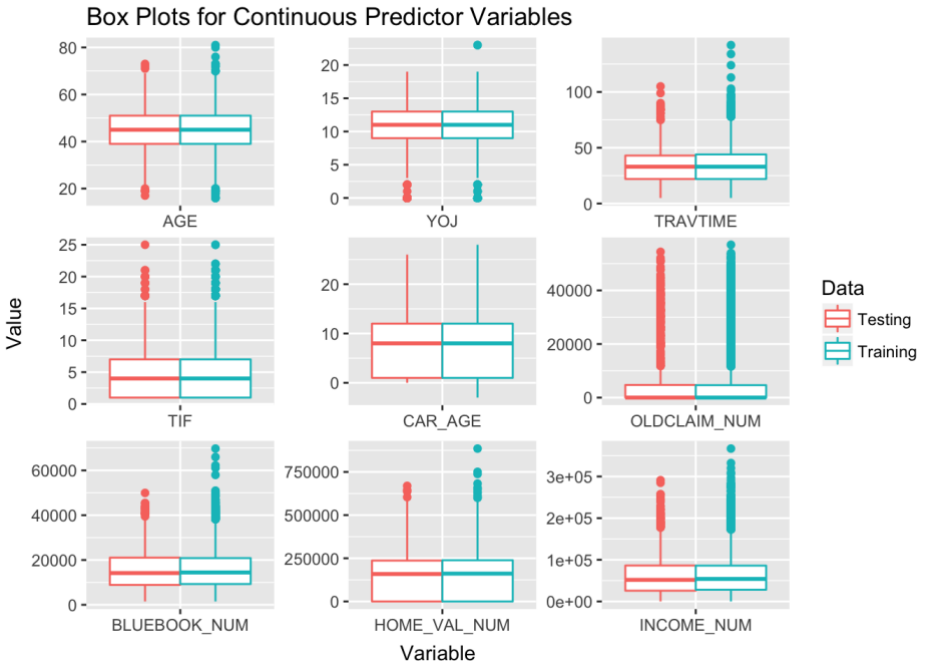

Figure 3 – Continuous Predictor Box Plots

Figure 4 shows the same data in the form of histograms. From observing the data in this format, there appears to be unrealistic negative values for car age that should be investigated. Additionally, several predictors have frequent values that suggest that an indicator variable for those values could prove predictive.

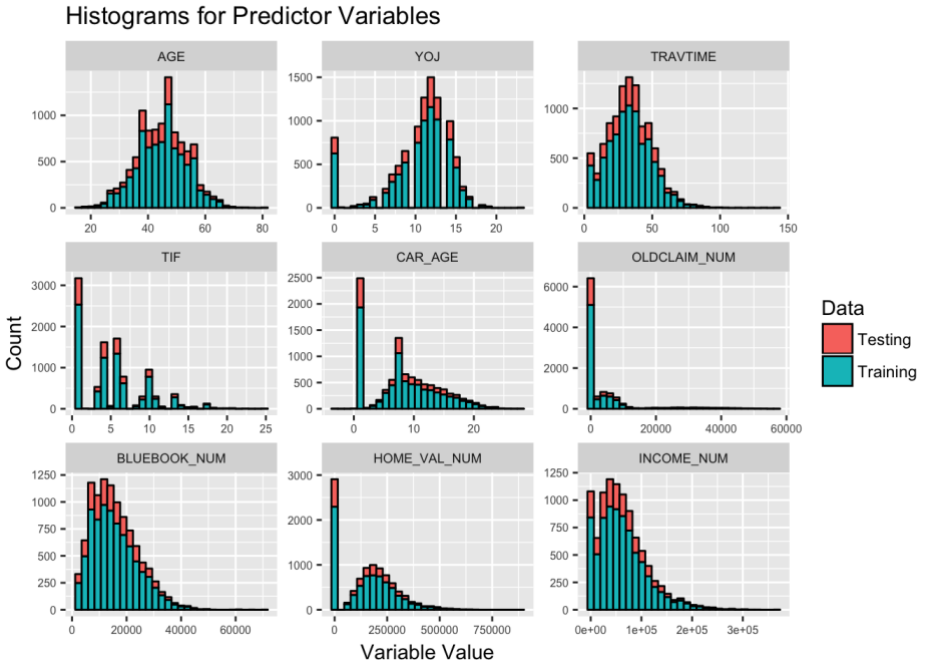

Figure 4 – Continuous Predictor Histograms

An analysis of continuous predictor correlation is shown in Figure 5. None of the variables show extremely high correlation with each other. The strongest correlation is between income and home value, which is intuitive.

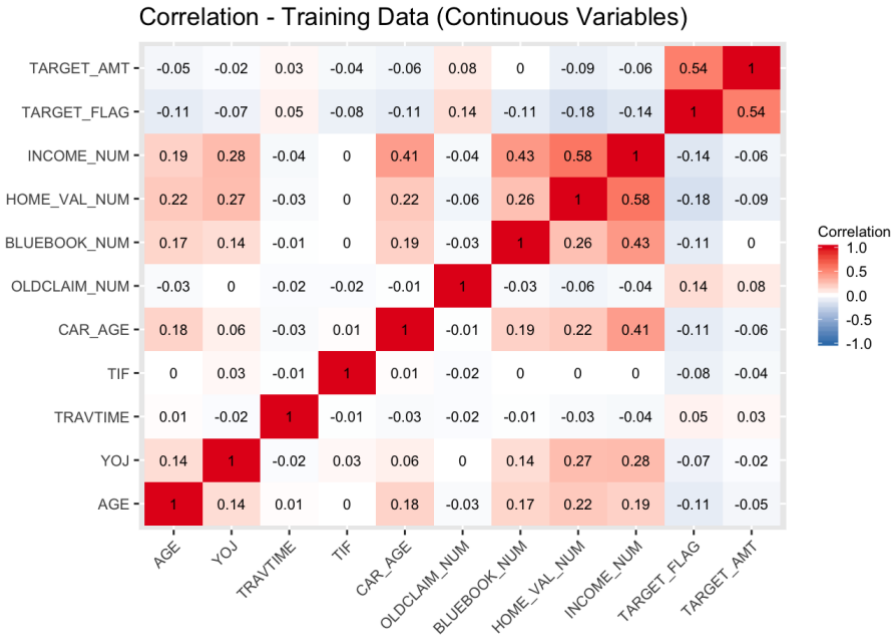

Figure 5 – Continuous Predictor Correlation

In terms of the continuous predictors and their relationships with the targets, the continuous target shows little promise of building a strong predictive model. Predicting the size of a claim is difficult based solely on driver characteristics. However, the univariate relationships with the binary targets show significant segmentation in the log-odds. Figure 6 shows the continuous predictors binned into deciles with associated training log-odds. The 95% confidence interval is displayed as a band around the point estimates.

Figure 6 indicates that flooring, capping, and other transformations are necessary in the Data Preparation section. Particularly for driver age, the relationship with the target is not linear as both very young and very old drivers are most likely to have claims.

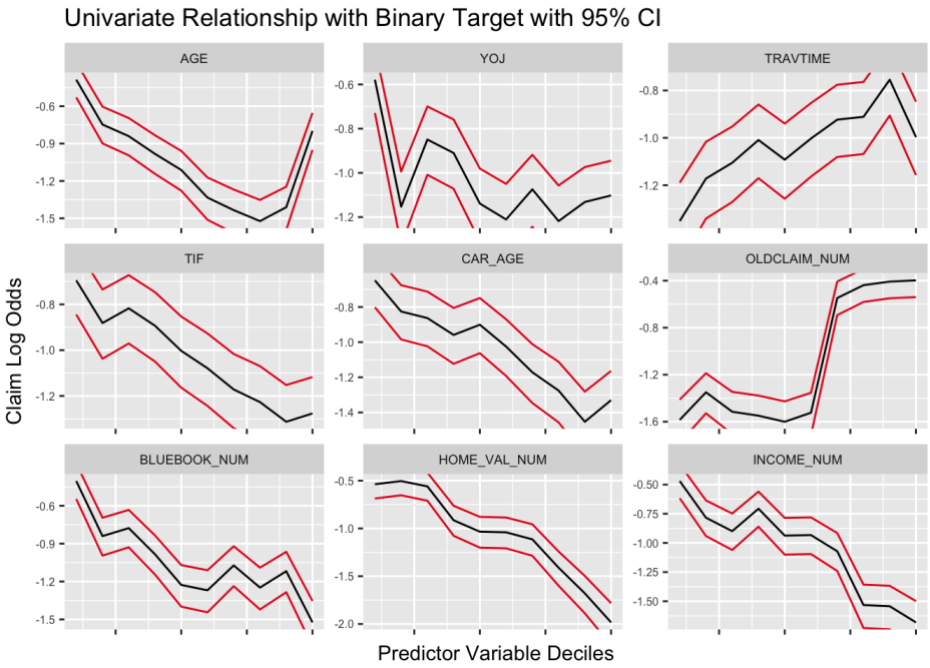

Figure 6 – Univariate Relationship with Binary Target

Other variables, including those that appear numeric but have very few distinct values, are treated as categorical. The first check on these variables is to determine if the various levels are populated sufficiently to build a model. Figures 7 and 8 display the number of training and testing observations in each category. Based on the results, grouping categorical levels appears necessary for only a subset of these variables.

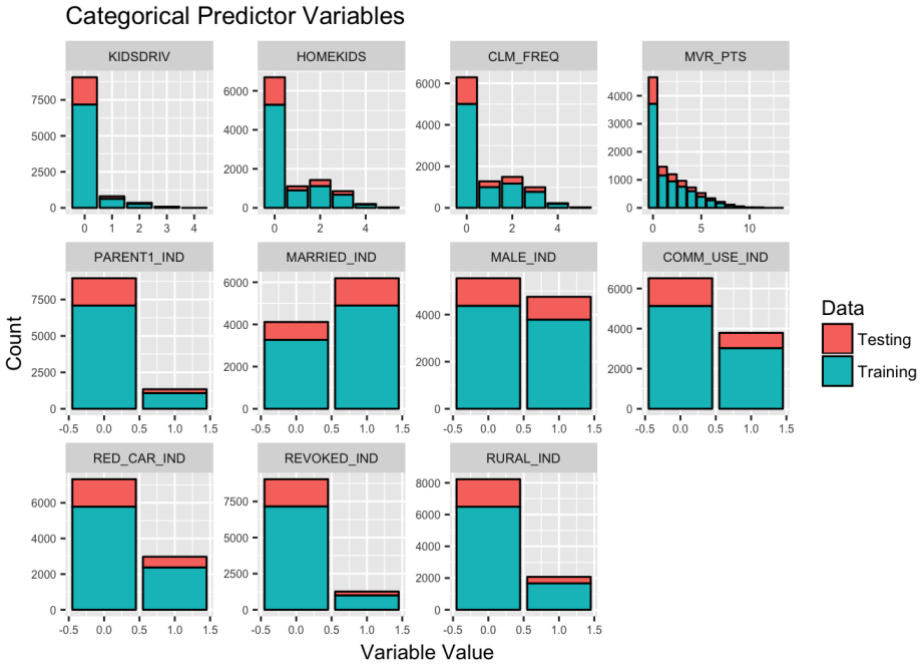

Figure 7 – Categorical Predictor Variables Population

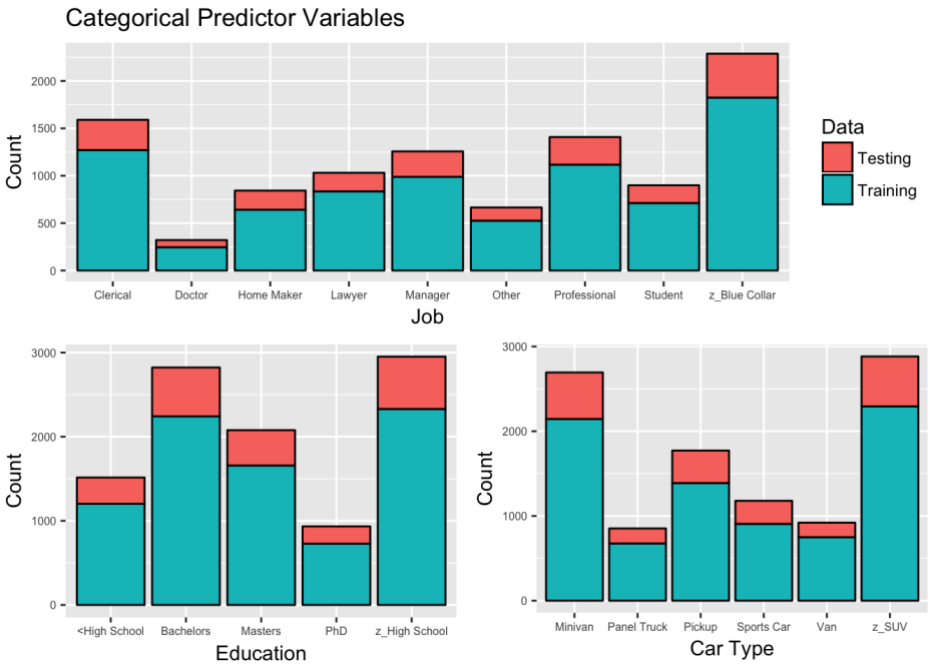

Figure 8 – Categorical Predictor Variables Population

The univariate relationships with the binary target shown in Figures 9-10 further substantiate the need to group some of the levels in the Data Preparation section. Some levels show little segmentation. However, the segmentation found in the categorical variables as a whole is very promising for building predictive models.

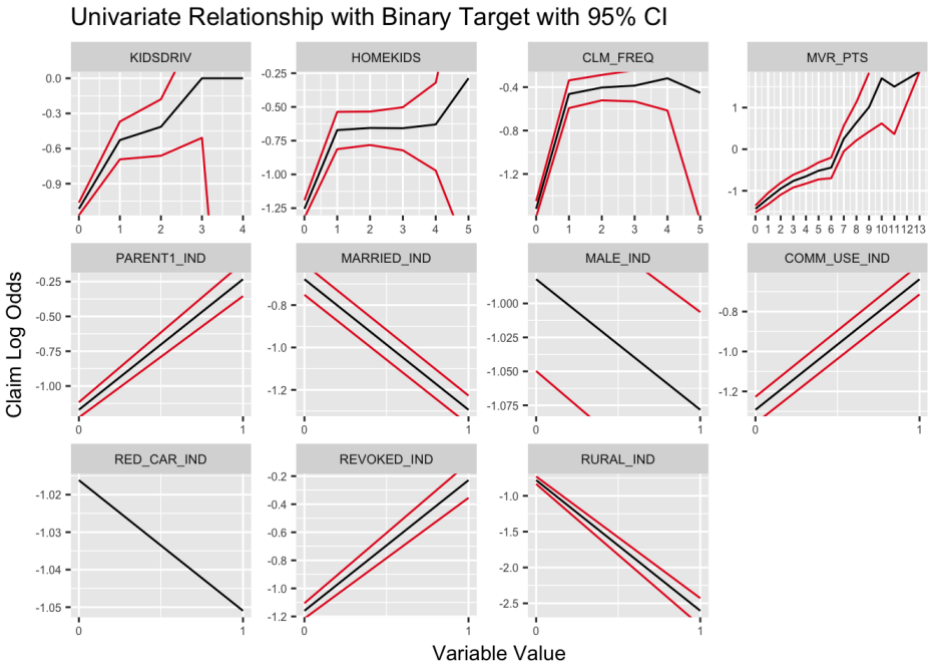

Figure 9 – Univariate Relationships with Binary Target

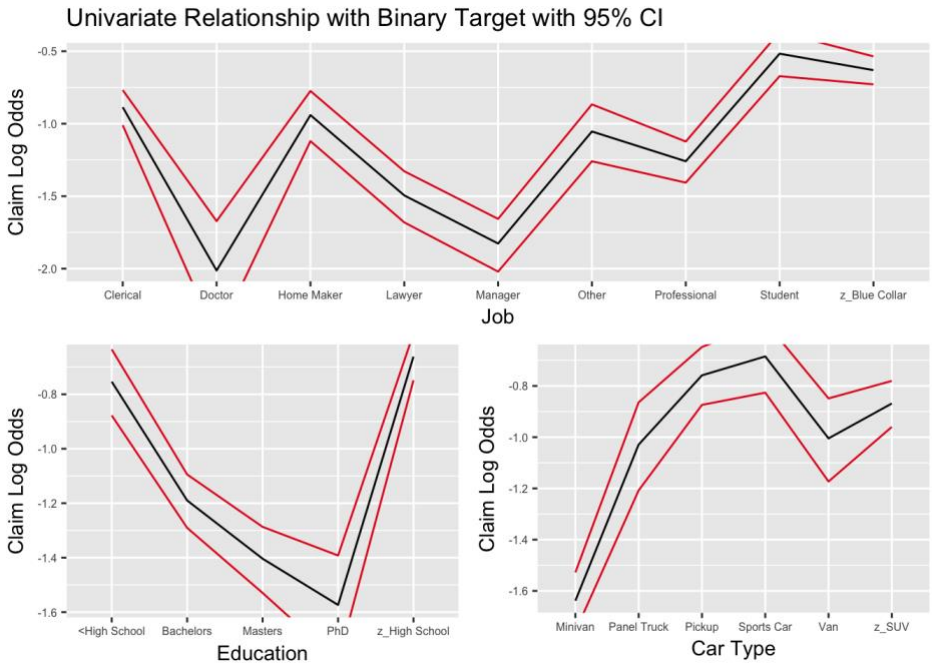

Figure 10 – Univariate Relationships with Binary Target

In the case of the claim severity data, like the continuous predictors, the univariate relationships are weaker than for the binary target, so the expectations for that model should be lower than the logistic regression.

Data Preparation

The first step in preparing the data is to drop the indicator for the presence of a red car. The variable does not appear to have a strong univariate relationship and there is a strong lack of a causal link between the color of a car and its claim experience.

Each of the variables with missing values are imputed using the means of the data. Missing value indicators are created with the exception of driver age due the small number of observations actually missing. Car age is floored to have a minimum value of one to avoid unintuitive results.

Variables serving as indicators for common values seen in continuous variables such as travel time of five and income of zero are created. These variables should account for any unique behavior related to those values.

Continuous variables are then capped. For years on job, the value is capped at 15 to reflect the lack of lift after that year. Similar analysis is used to cap travel time at 65 and MVR points at 7. Other continuous variables are capped at the 98th percentile of the training data. The large and skewed values for the Blue Book estimate and previous claim costs result in the need for a log-transformation of those variables.

Next, categorical variable levels are grouped. For the number of children drivers, number of children at home, and previous claim frequency, the values greater than one are grouped together to create indicator variables. For car type, panel trucks and vans are grouped together due to their similar experience and physical attributes.

To account for the relatively higher cardinality in the education, job, and car type variables, impact or likelihood encoding is used. This entails replacing the category with the empirical training probability of a claim. This results in a new numeric variable. The benefit is that it reduces the number of dummy variables created and preserves the relationships between variable levels. An issue with this method is that it results in less intuitive interpretations of the coefficients in a model and introduces some information leakage by incorporating the target itself. However, because these variables have relatively low cardinality and the target is not rare, the impact of this leakage on the final model is minimal.

In the context of estimating risk, preserving the unique univariate relationship between age and the binary target is important. Over much of the age range, there is a monotonically decreasing relationship between age and claim probability. There are spikes for very old and very young drivers. This relationship can be incorporated with likelihood encoding as well. First the ages are grouped into deciles and the deciles are replaced with the claim probability in that decile.

With the exception of the likelihood encoded variables, the final variables and their relation to the binary target are shown in Figures 11-13. The transformations create a strong set of predictors to use in model building. The same transformations, with likelihood encoded variables excluded, are considered for used in the claim severity model as well.

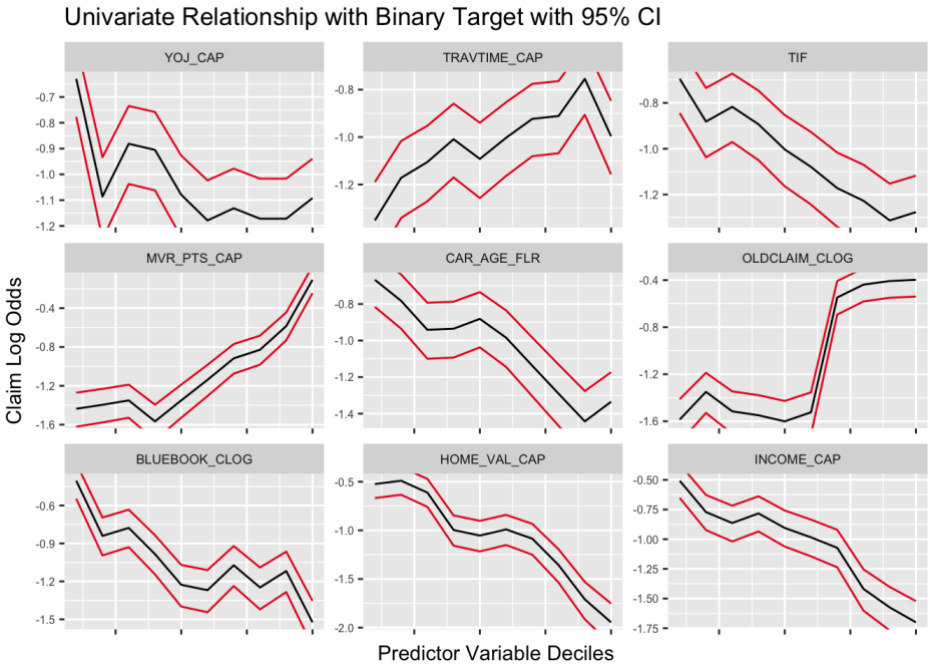

Figure 11 – Univariate Relationships with Binary Target for Final Continuous Predictors

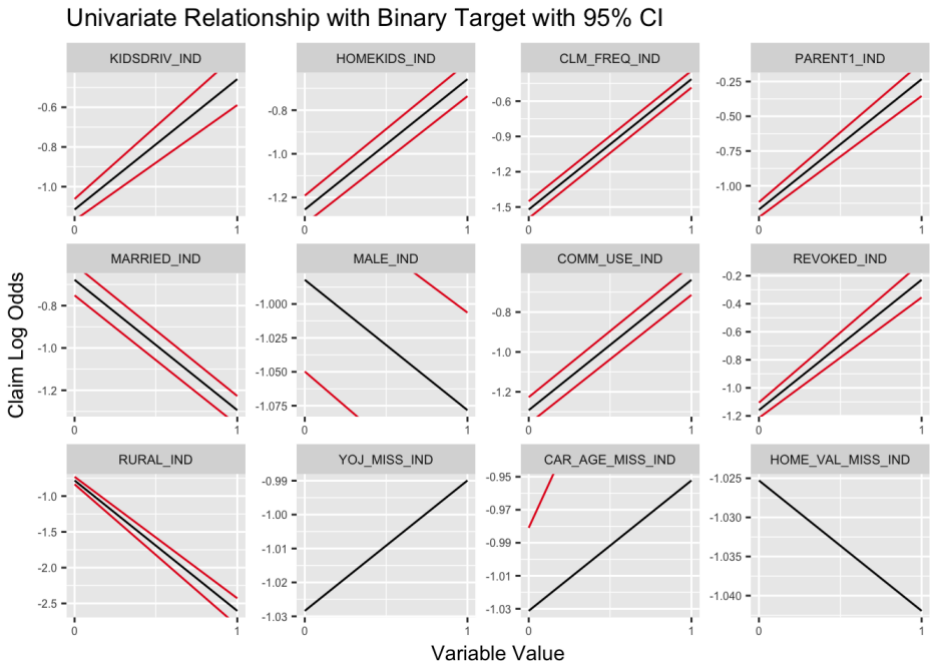

Figure 12 – Univariate Relationships with Binary Target for Final Categorical Predictors

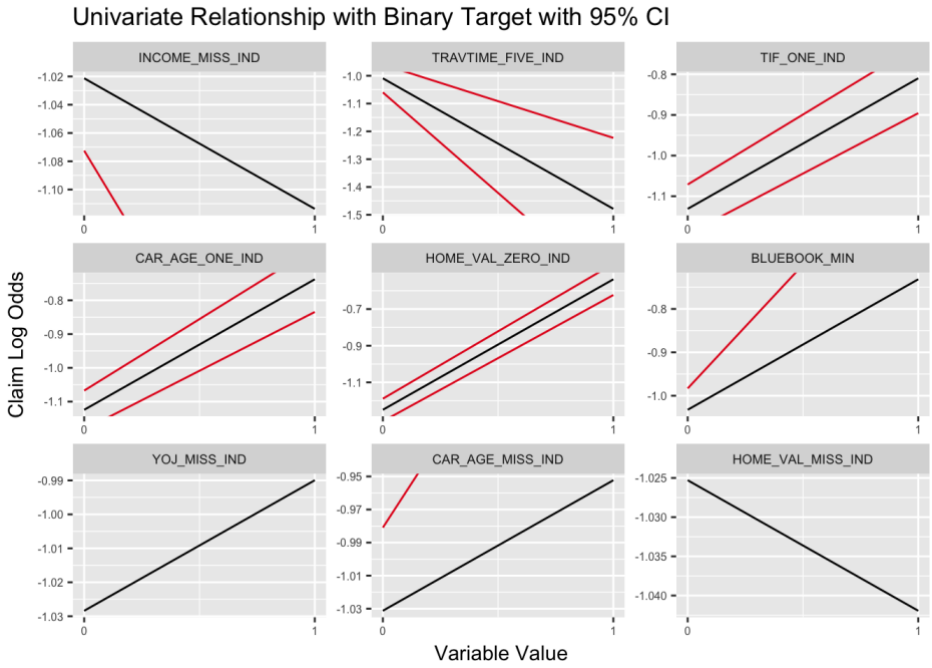

Figure 13 – Univariate Relationships with Binary Target for Final Categorical Predictors

Build Models

For binary classification, the models are assessed through five-fold cross validation. The area under the curve (AUC) and Kolmogorov Smirnov (KS) statistics are based on the out-of-fold predictions to minimize overfitting.

The first model considered is a baseline containing only an indicator whether the driver had a prior claim. This variable is a strong predictor as the presence of a prior claim is an indicator of future claims. Figure 14 shows the ROC curve. Based on the shape of the curve, there is opportunity for a more complex model to better fit the data.

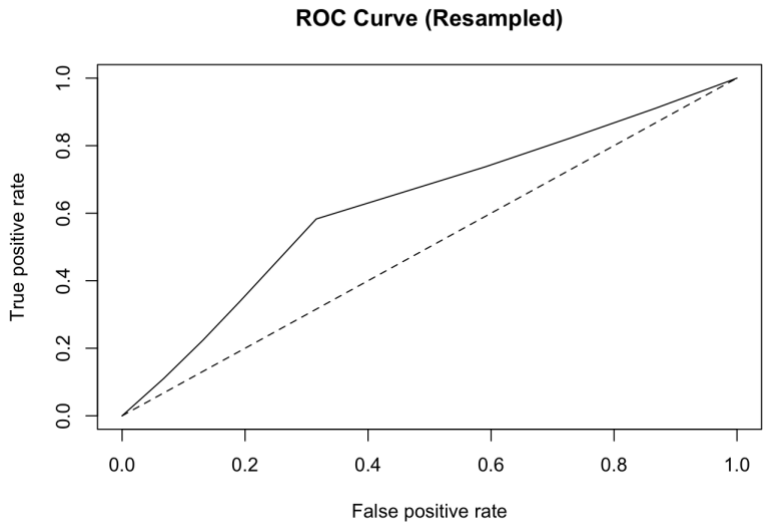

Figure 14 – Model 1 ROC Curve

Figure 15 shows actual and expected log-odds for the data split into 50 equally sized groups. This basic model splits the data into only two groups and does not rank the observations well.

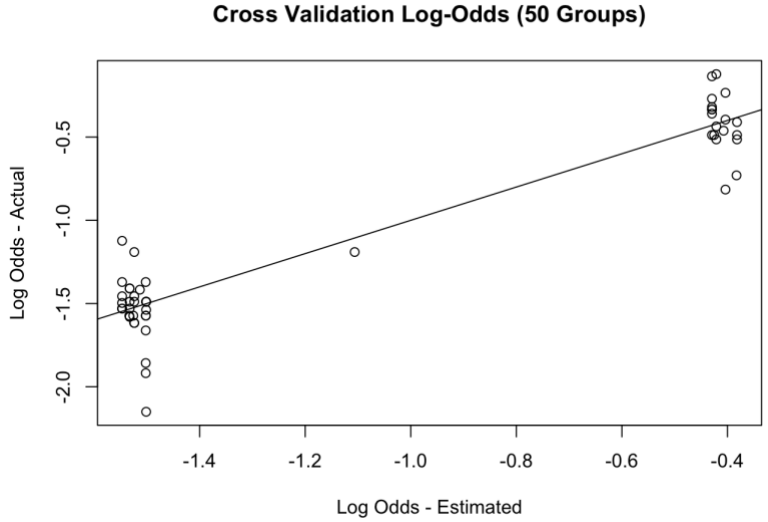

Figure 15 – Model 1 Log-Odds

Figure 16 shows a modified lift chart for Model 1. Deciles of model predictions are shown on the x-axis and the actual probability is shown on each of the bars. A successful model shows monotonically increasing bar heights and large separation between the lowest and highest bars. The ratio of their heights can also serve as a metric of how well the model ranks risks. Model 1

does not have strong separation outside of the two large groups. The high-risk group is approximately twice as likely to have a claim as the lower risk group.

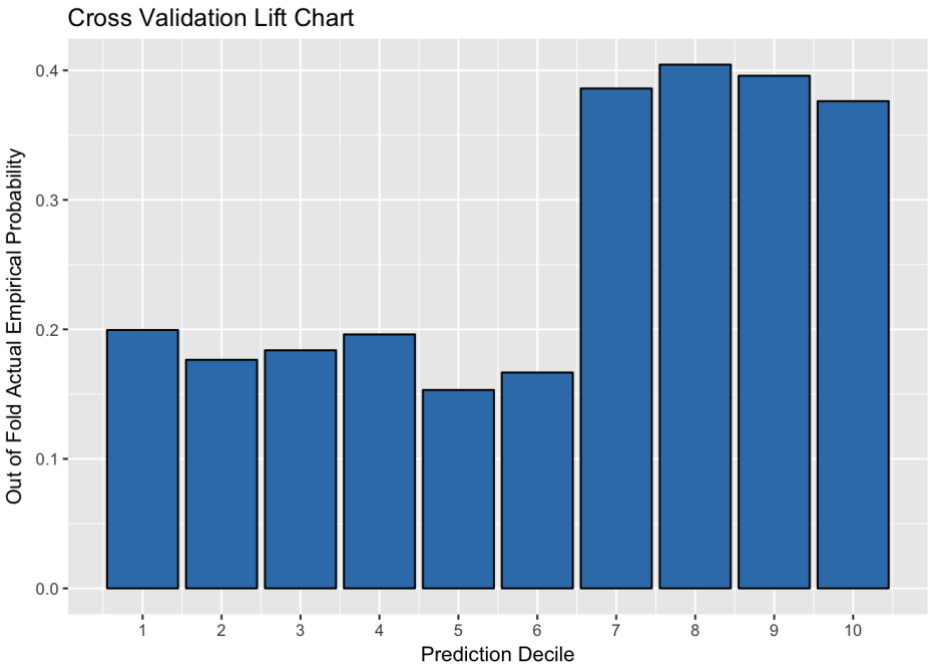

Figure 16 – Model 1 Lift

Model 2 uses forwards stepwise selection based on AIC to select variables. In the cross-validation step, separate forwards selection occurs for each fold. Only main effects are included in this iteration of the model build. Figure 17 shows the ROC curve for Model 2. The shape of the curve indicates that the model fits the data well.

This model includes: an indicator for rural areas, the job likelihood encoding, an indicator for the presence of a prior claim, an indicator for being a single parent, the car type likelihood encoding, an indicator for a revoked license, home value, travel time, time in force, MVR points, an indicator for kids driving, the age likelihood encoding, an indicator for commercial use, the bluebook value, an indicator for being married, the education likelihood encoding, an indicator for zero income, previous claim loss amount, income, and years on the job.

All of the model coefficients match their univariate and intuitive relationships where applicable with the exception of prior claim costs transformed. Upon further investigation, that variable has a variance inflation factor of over 40 due to its correlation with claim frequency. As a result, it should not be included in a final model.

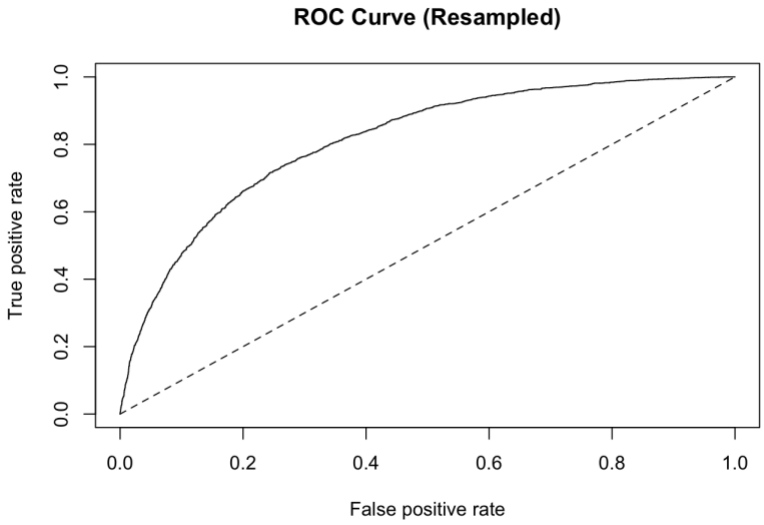

Figure 17 – Model 2 ROC Curve

The log-odds shown in Figure 18 indicate that the model ranks the probability well.

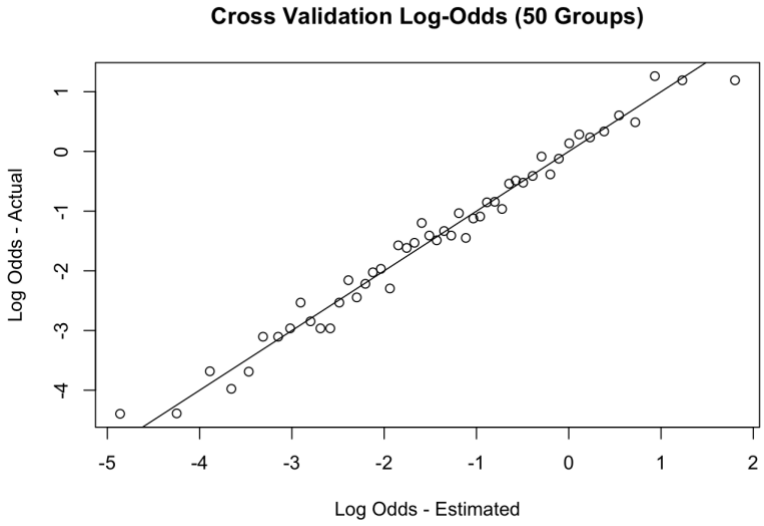

Figure 18 – Model 2 Log-Odds

Unlike Model 1, the modified lift shown in Figure 19 shows clear separation in risk between deciles and consistent monotonically increasing probabilities. This figure shows that the model fits the data well. The highest risk decile is almost 40 times more likely than the bottom decile to have a claim.

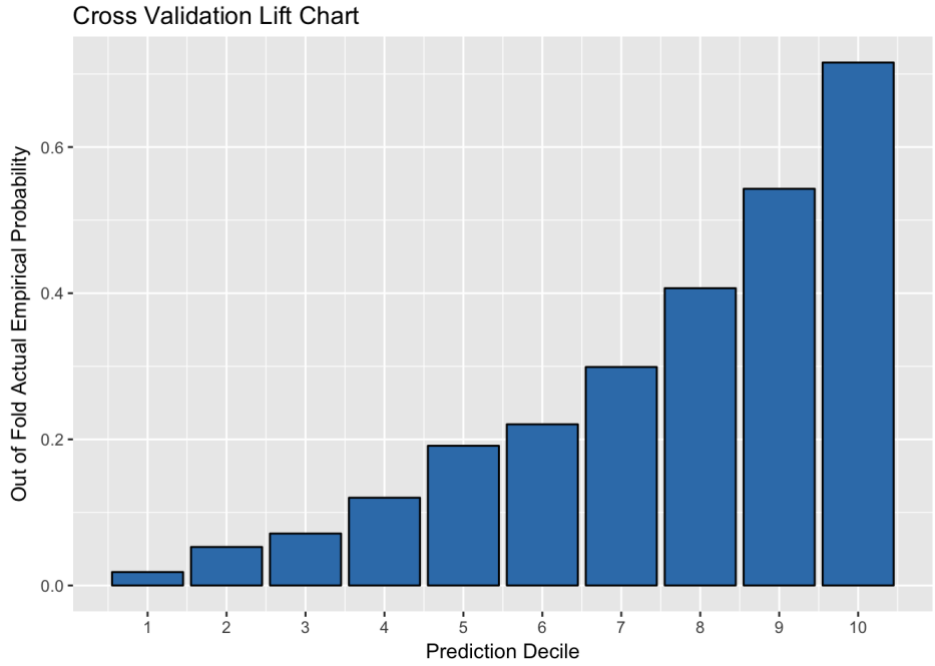

Figure 19 – Model 2 Lift

Model 3 is the same as Model 2 but rather than a logit-based model, Model 3 uses probit regression. The ROC curve shown in Figure 20 looks similar to Model 2, however it is often easier to interpret at a glance the coefficients generated by the logit model.

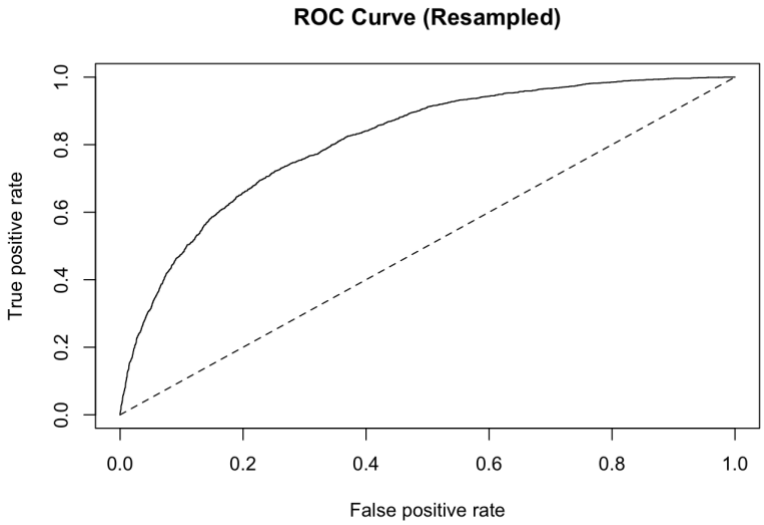

Figure 20 – Model 3 ROC Curve

The log-odds shown in Figure 21 show that for the lowest predicted group, the model is over-predicting. The true likelihood of a claim is greater for that group than the model is suggesting. This difference makes sense given the use of probit compared to logit. This makes the lift chart in Figure 22 less reliable from a modified lift ratio perspective but overall, it looks similar to that of Model 2. The value of the modified lift ratio becomes so high that ratio is impacted by small changes in the probability of a claim in the lowest-risk group.

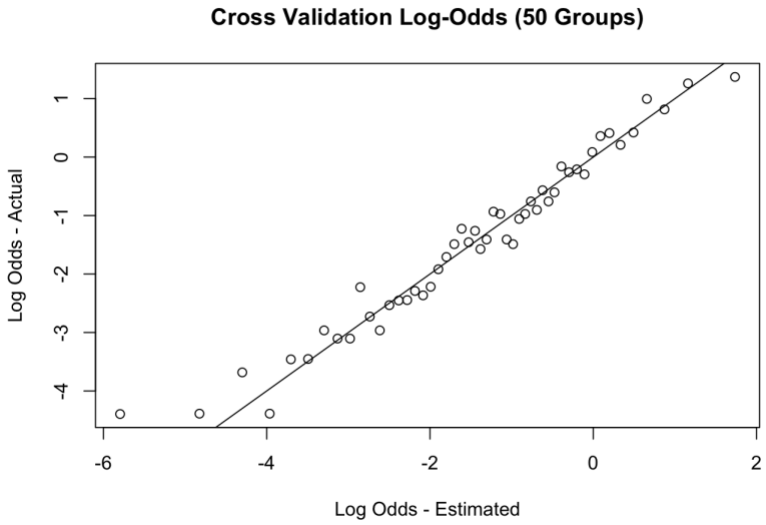

Figure 21 – Model 3 Log-Odds

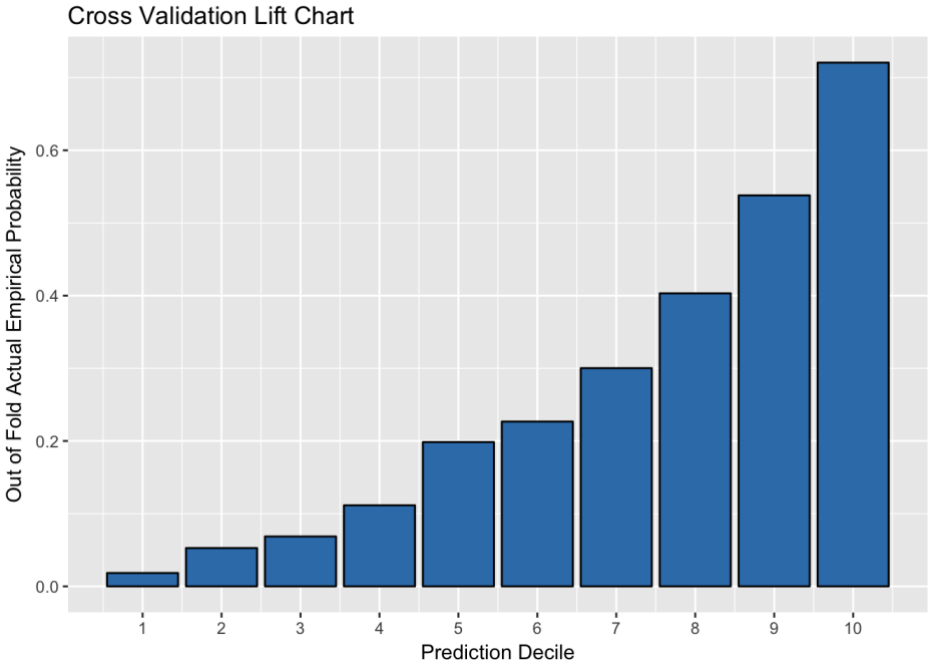

Figure 22 – Model 3 Lift

The first three models only considered main effects. Interactions between variables add complexity to the model but may also improve the fit. One way to observe whether interactions may add value is to analyze decision trees. When a variable appears to have different effects on different branches, an interaction term may apply. Observing different trees such as the one shown in Figure 23, indicates that interactions will be significant.

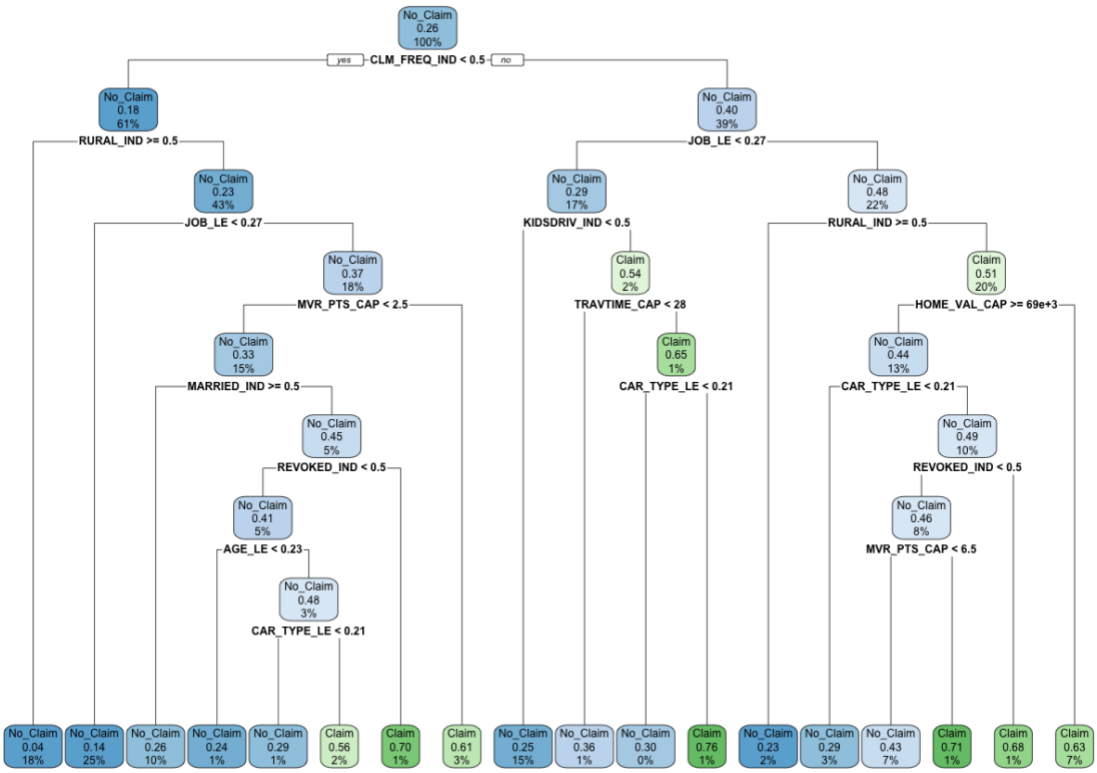

Figure 23 – Decision Tree for Interaction Detection

Model 4 keeps all of the variables included in Model 2 with the exception of the prior loss costs but then applies forwards stepwise selection to second-level variable interactions. Because this selection is applied to the training data, overfitting could occur. To account for this, only the first seven interaction pairs to enter the model are included. Figure 24 shows the ROC curve for the model. Based on the curve alone, the model appears similar to that of Model 2.

The log-odds shown in Figure 25 and the modified lift chart shown in Figure 26, are also similar to Model 2. This does not suggest that the increase in model complexity and risk of overfitting is worth implementing.

The use of likelihood encoding also makes interactions incorporating these variables even more difficult to interpret. For example the interaction of the kids driving indicator and the job likelihood encoded variable is not clear.

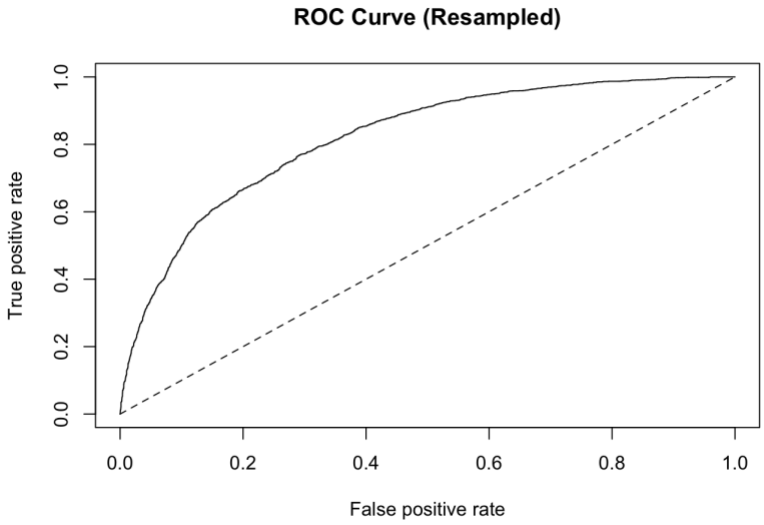

Figure 24 – Model 4 ROC Curve

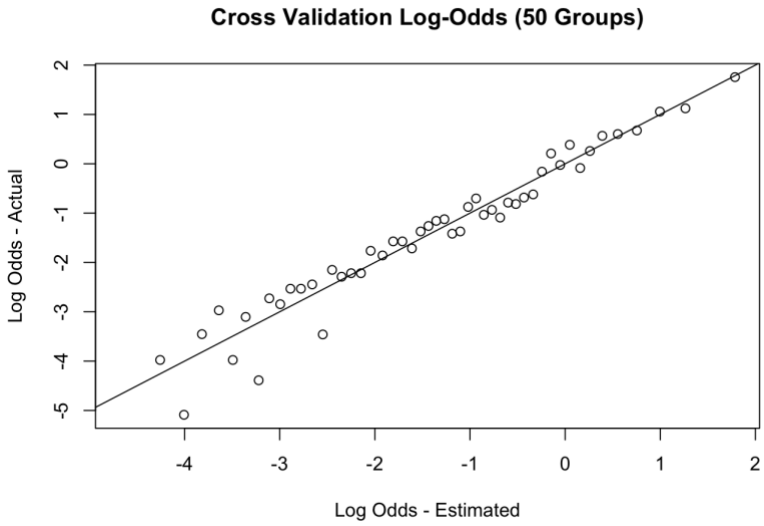

Figure 25 – Model 4 Log-Odds

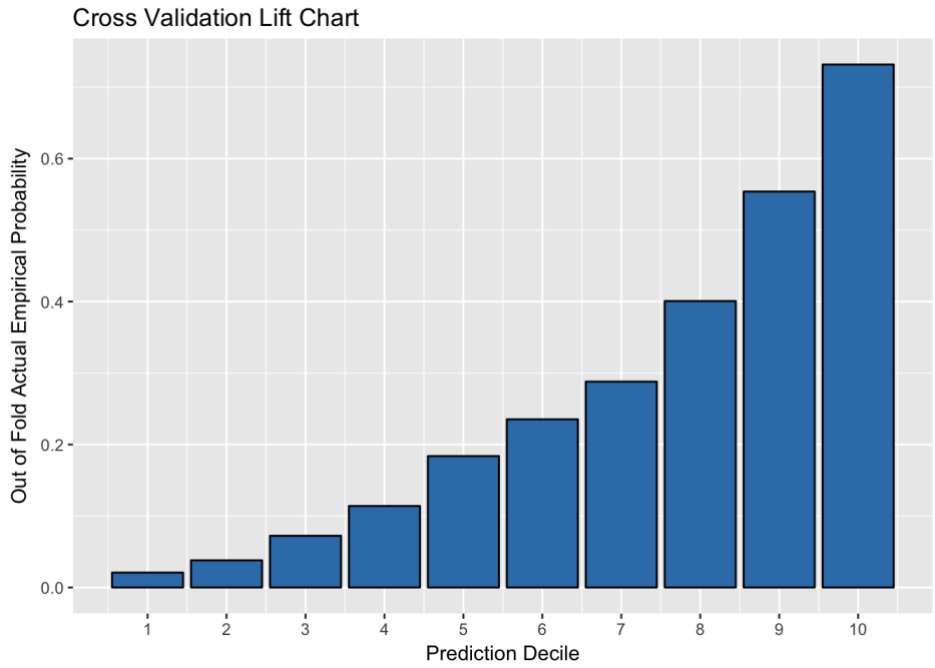

Figure 26 – Model 4 Lift

Model 5 represents the use of the continuous target using ordinary least squares regression. Backwards stepwise selection is used to select variables in the model. In the case of the continuous target, the likelihood encoded variables are not used and the regular variables are used in their place. Figure 27 shows the diagnostic plots for the model. The QQ plot, and scale-location plot indicate that the assumptions of ordinary least squares regression are violated.

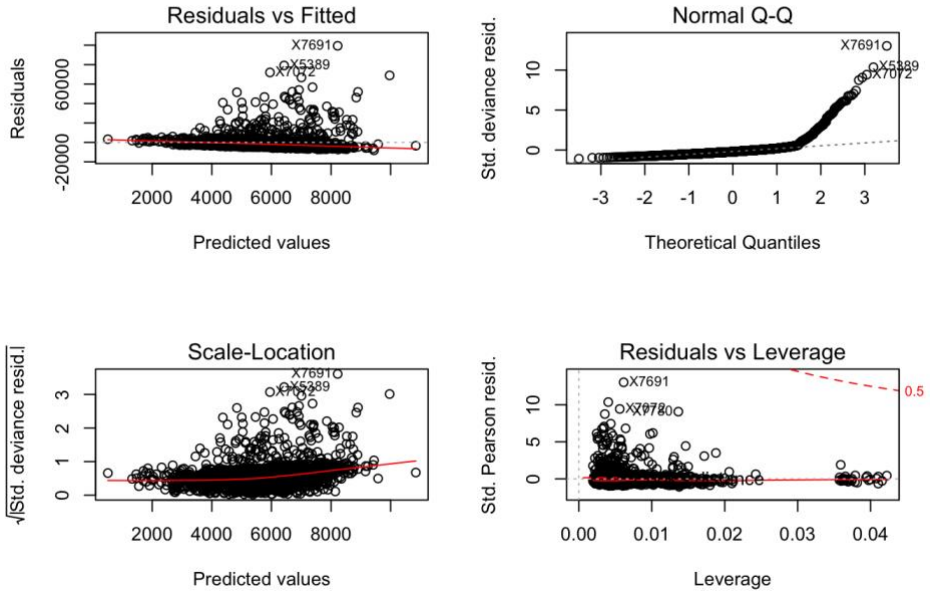

Figure 27 – Model 5 Diagnostic Plots

Figure 28 shows the modified lift for Model 5. The model does segment claims to large and small sized claims but the chart is from monotonically increasing, indicating poor ranking overall.

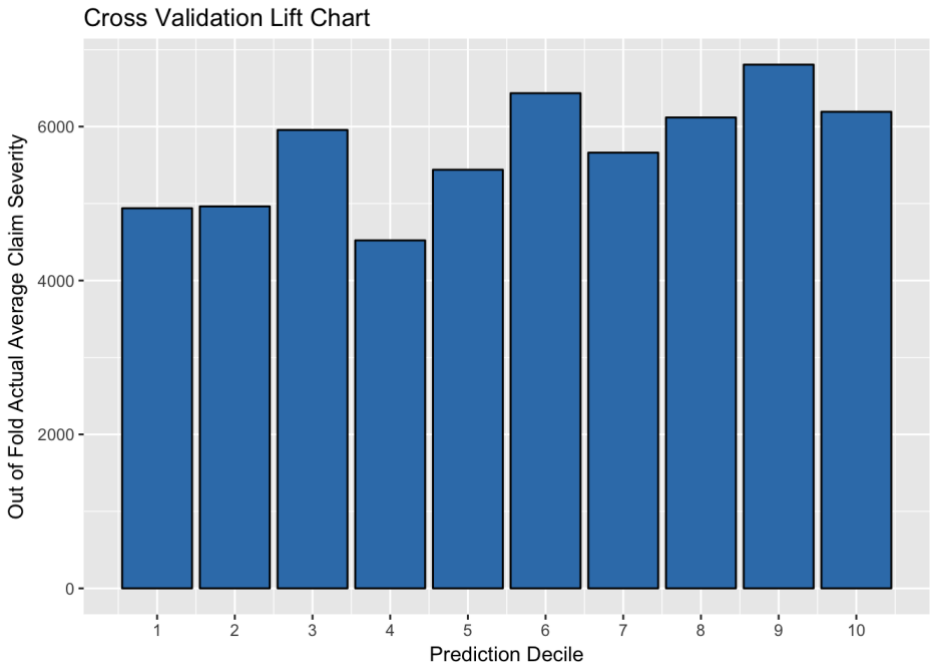

Figure 28 – Model 5 Lift

Model 6 uses the same techniques as Model 5 with the exception of using a Gamma GLM with a log-link function. The scale-location plot shown in Figure 29 shows that this model fits the entire population better.

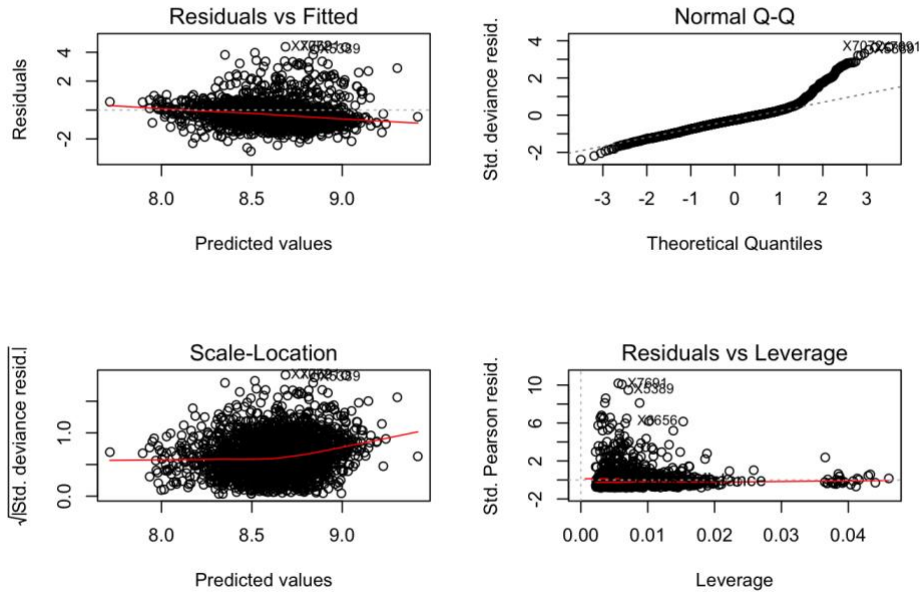

Figure 29 – Model 6 Diagnostic Plots

Figure 30 displays the modified lift chart. The Gamma GLM ranks the claims better than Model 5 however, the lift is less dramatic than the logistic regression. For measuring overall risk, it is easier to segment policies on the basis of whether a claim will occur rather than the cost of the claim.

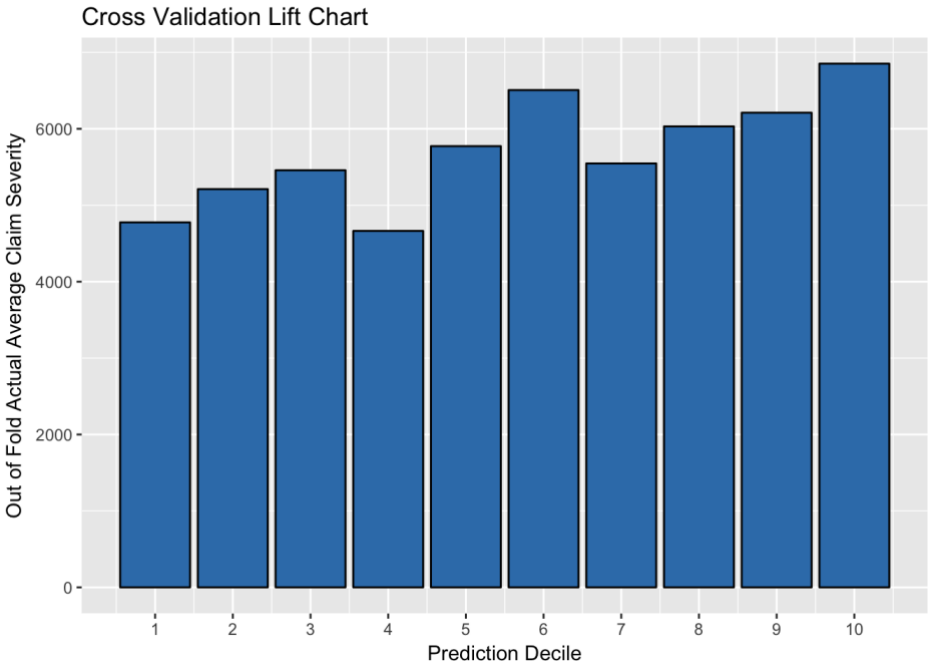

Figure 30 – Model 6 Lift

Select Models

Fit statistics for the binary target models are included in Figure 31. Model 4 has the best fit from AIC, AUC, and KS statistics. However, that model is significantly more complex and less interpretable with its use of interactions with likelihood encoding. That complexity does not result in a considerably better fit. Instead, Model 2 excluding prior claim cost due to multicollinearity is selected as the final model. Model 2 has interpretable coefficients and shows a similarly strong ability to segment risk. The fit statistics and plots show similar results to Model 2 as a whole.

Model	AIC	AUC	KS	Modified Lift
1	8953.3	0.624	0.267	2.30
2	7310.4	0.811	0.472	37.59
3	7314.5	0.812	0.470	53.58
4	7191.4	0.819	0.477	41.75
2 (excl. Prior Cost)	7319.4	0.812	0.468	36.89

Figure 31 – Binary Target Fit Statistics

Model	AIC	RMSE	Modified Lift
5	44633	7675.38	1.25
6	41240	7723.70	1.43

Figure 32 – Continuous Target Fit Statistics

The fit statistics for the continuous target models are shown in Figure 32. The use of the Gamma GLM results in a better fit. As a result, Model 6 is chosen as the final model.

The ROC, log-odds, and modified lift plots for the final selected logistic regression is shown in Figures 33-35. As discussed above, the results are close to that of the unmodified Model 2.

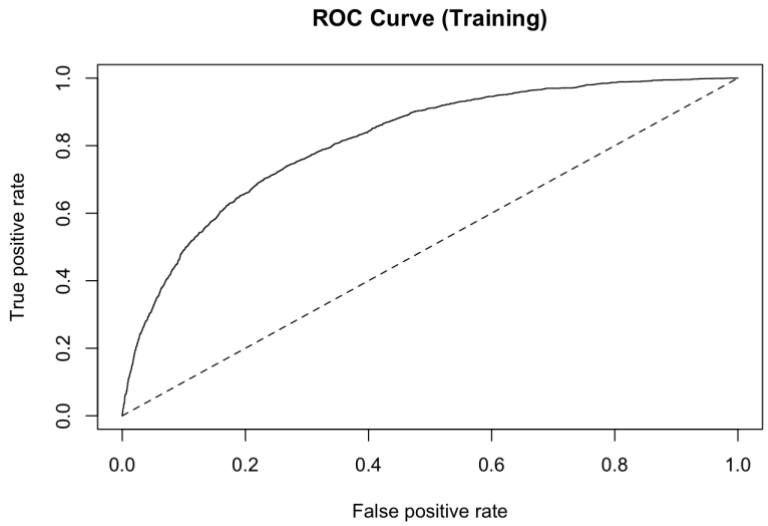

Figure 33 – Modified Model 2 ROC Curve

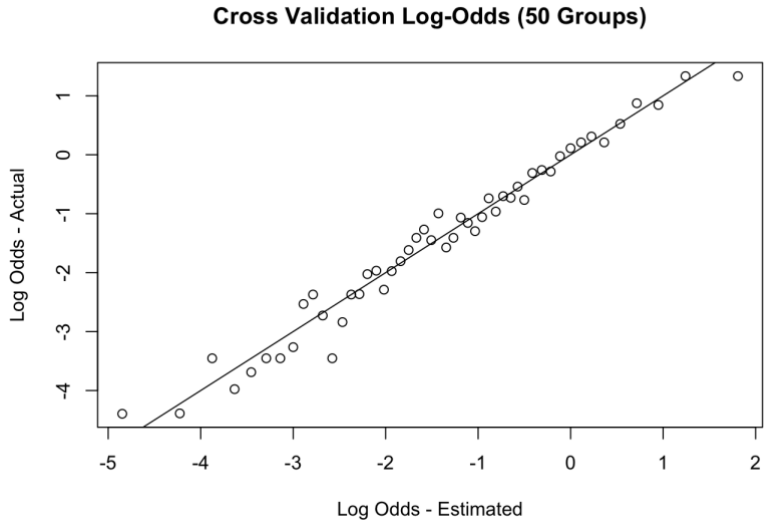

Figure 34 – Modified Model 2 Log-Odds

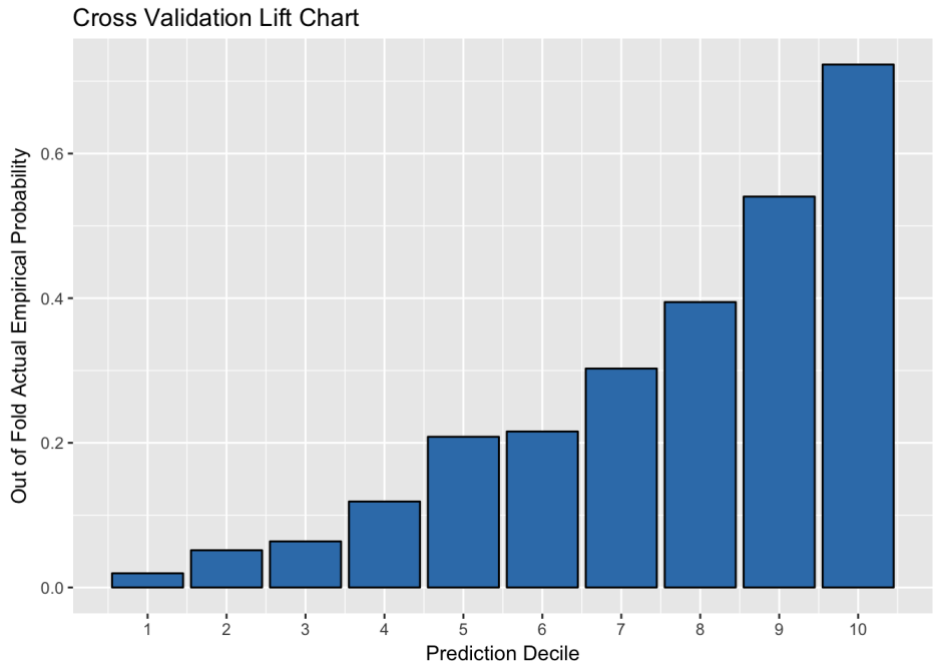

Figure 35 – Modified Model 2 Lift

Conclusion

To predict the probability that a customer had a claim, four models were fit to historical insurance data. After imputing missing values and transforming predictors, the final selected model is able to segment the customers based on their risk such that the riskiest ten percent of customers are approximately 40 times more likely to have a claim than the safest ten percent. An additional model is used to predict the loss amount given the presence of a claim. Using the two models together can allow for better pricing to match price to risk or be used in underwriting to reject the riskiest customer applications.