Unit 2 Assignment: “INSURANCE LOGISTIC REGRESSION PROJECT”

MSDS 411 Section 56, spring 2019

Prabhat Thakur 05/10/2019

Introduction

The purpose of this project is to build a predictive model using an auto insurance company’s customer data to predict the probability of a customer making insurance claim and in case of claim predict the claim amount (loss to insurance company). The data contains historical information about insurance customer including previous claim history, demographic and financial information like age, family, education, job, income, auto and driving history like car details, traffic violations. For arriving to a best predictive model for probability of claim, multiple logistics models are fit to the data and evaluated. To predict the claim amount (loss), linear regression model is fit to the data as well. The final model is selected based on fit statistics and metric such as LOG LIKELIHOOD, AIC, or ROC CURVE.

Bonus Problem

I have attempted following bonus problems:

1. (20 points): Used additional logistics regression function “lrm” from RMS R package.
2. (5 Points) Use at least one PROBIT MODEL when building your logistic models.
3. (20 Points) Recreate as much of the program as you can in a second program. It has been added in the last section of R code file.

Data Exploration

The data for this assignment contains information of auto insurance customers. There are 8161 observations in the training dataset and 2141 in the test dataset. The Training data set has 26 columns which includes 1 observation identification number INDEX, 2 response variables TRAGET_FLAG and TARGET_AMT for car crash and cost respectively, and 23 potential predictor variables for building the logistic regression model for TRAGET_FLAG prediction and linear regression model for TARGET_AMT. The test dataset contains same columns but TRAGET_FLAG and TARGET_AMT values are blank. The trained models will be used to predict TRAGET_FLAG and TARGET_AMT for test dataset observations.

Main objective of this assignment is to create logistic regression model to predict probability of a customer claim. The dataset consists both continuous and categorical variables with low and high cardinality. For the binary target, about 35% of the 8,161 observations have a claim. The continuous target has a heavy tail characteristic of a loss distribution as shown in Figure 1.

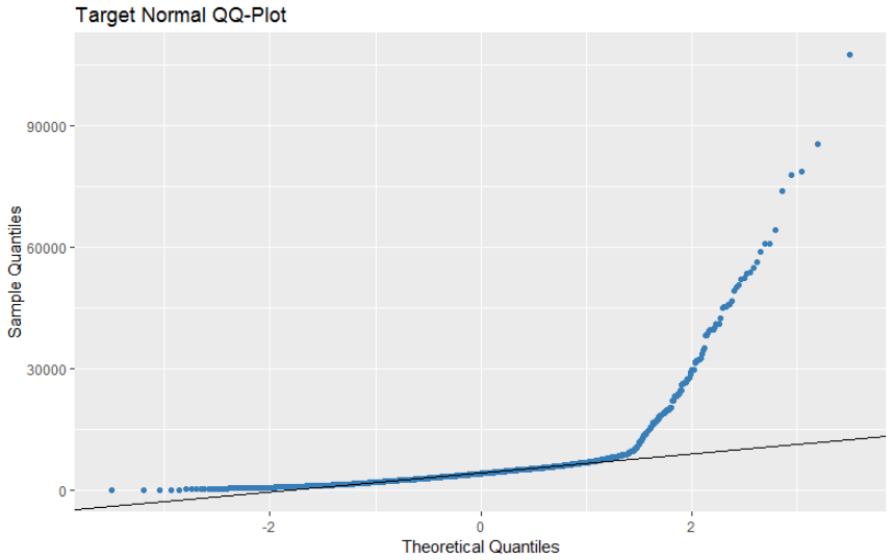

Figure 1 – QQ-Plots of Continuous Target

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.

To explore predictor variables from training and test data set together, both datasets were merged for data exploration purpose. Below are the boxplot and histogram of continuous predictor variables. Figure 2, shows the distribution of the continuous predictor variables. Based on the box plots, it is likely that the testing data likely comes from the same population as the training data. Figure 3 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. 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.

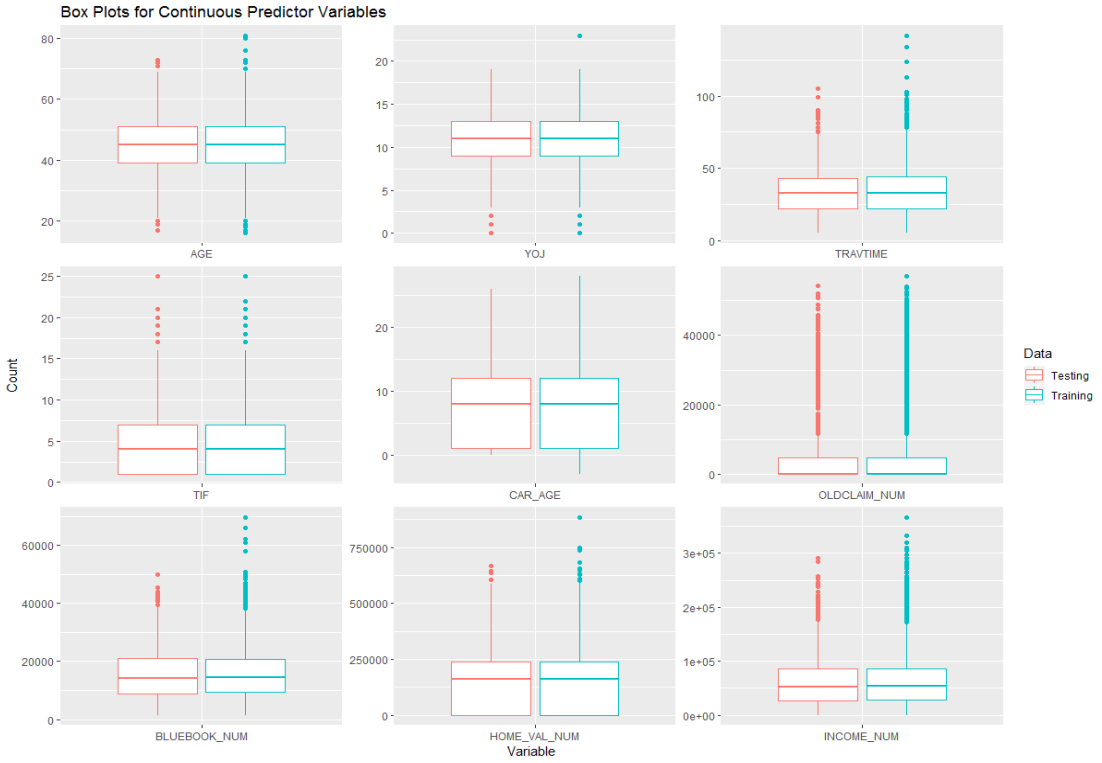

Figure 2 – Continuous Predictor Box Plots

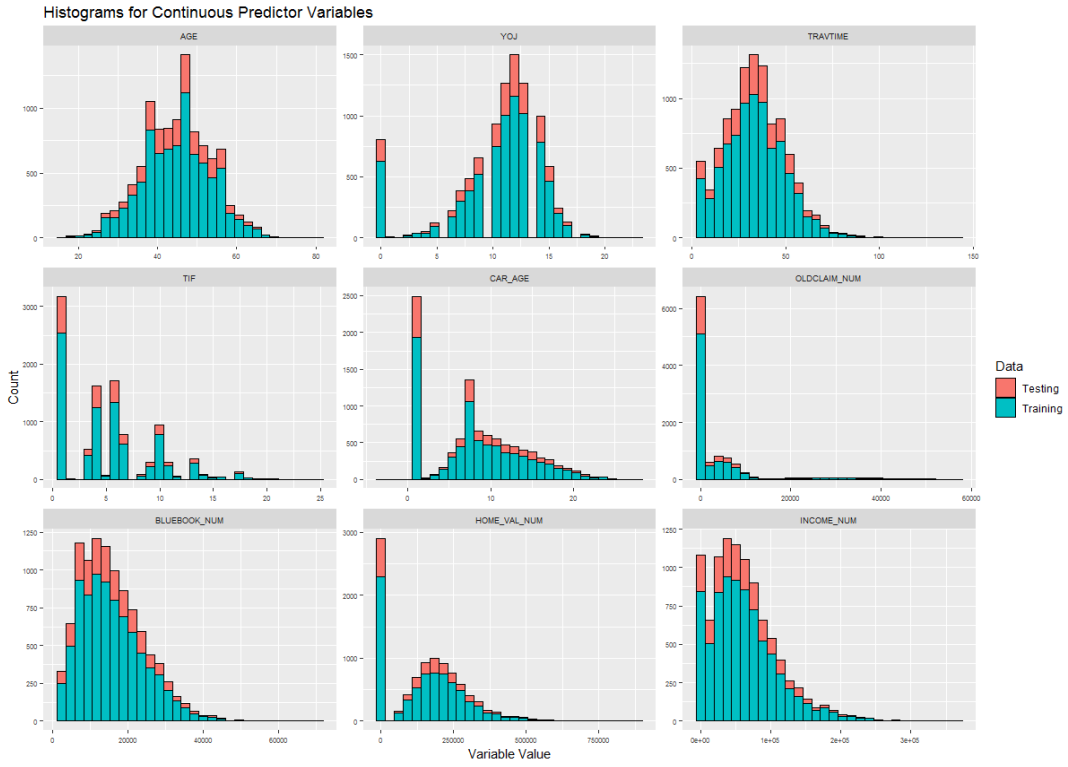

Figure 2 – Continuous Predictor Histograms

The correlation matrix heat map of continuous predictor correlation is shown in Figure 4. It shows that none of the variables have extremely high correlation with each other. The strongest correlation is between income and home value.

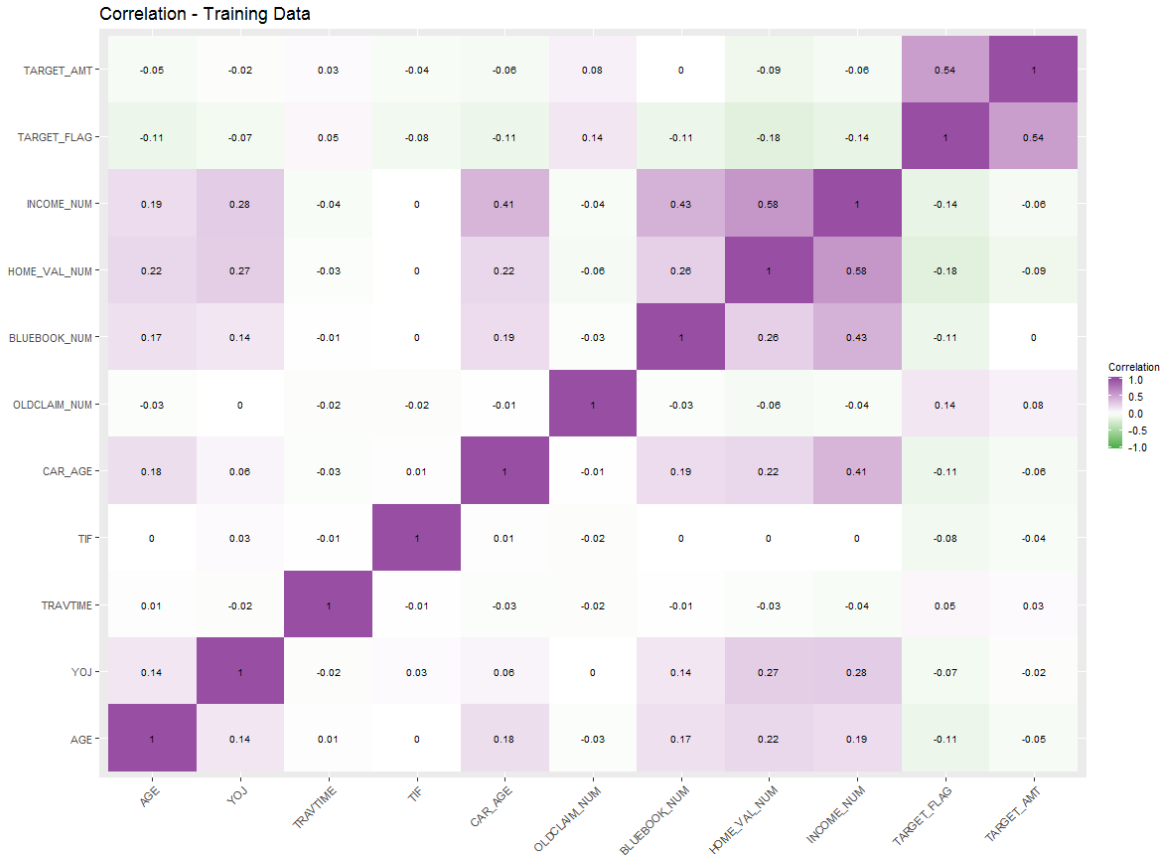

Figure 4 – Continuous Predictor Correlation

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 5 and 6 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.

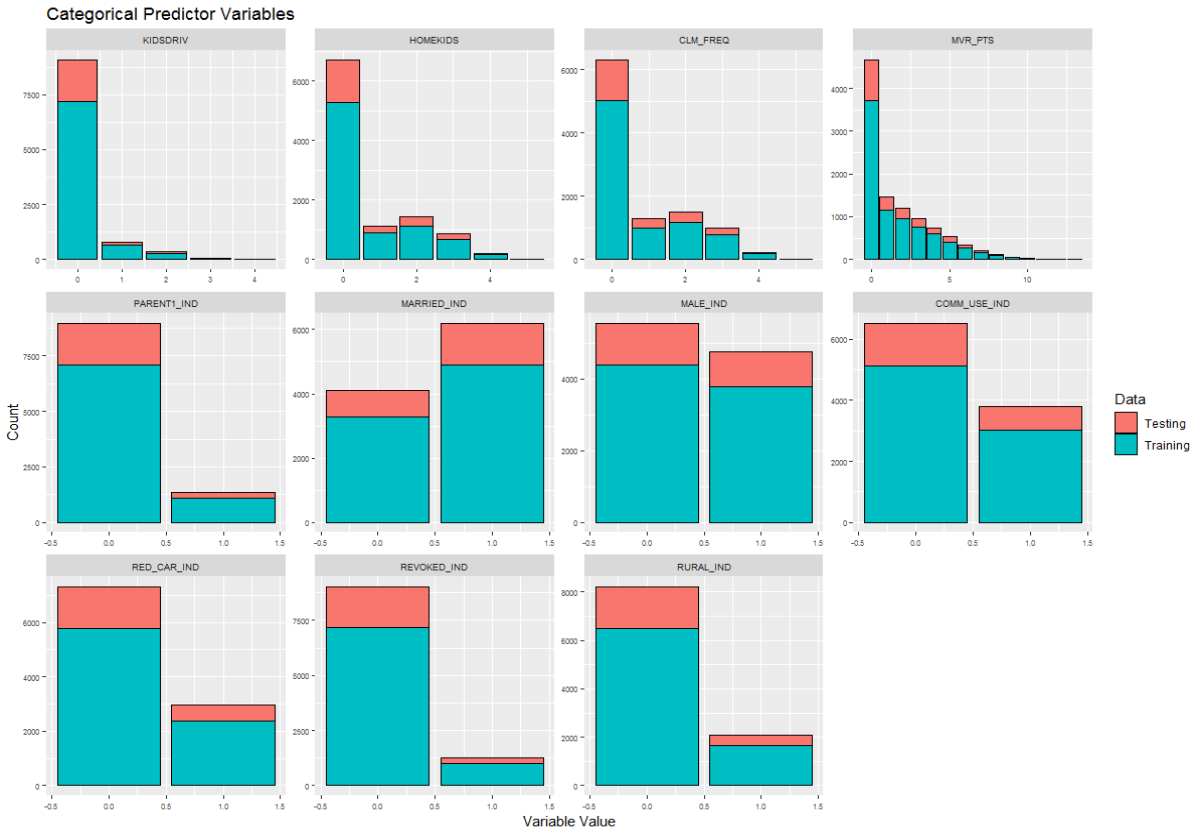

Figure 5 – Categorical Predictor Variables Population

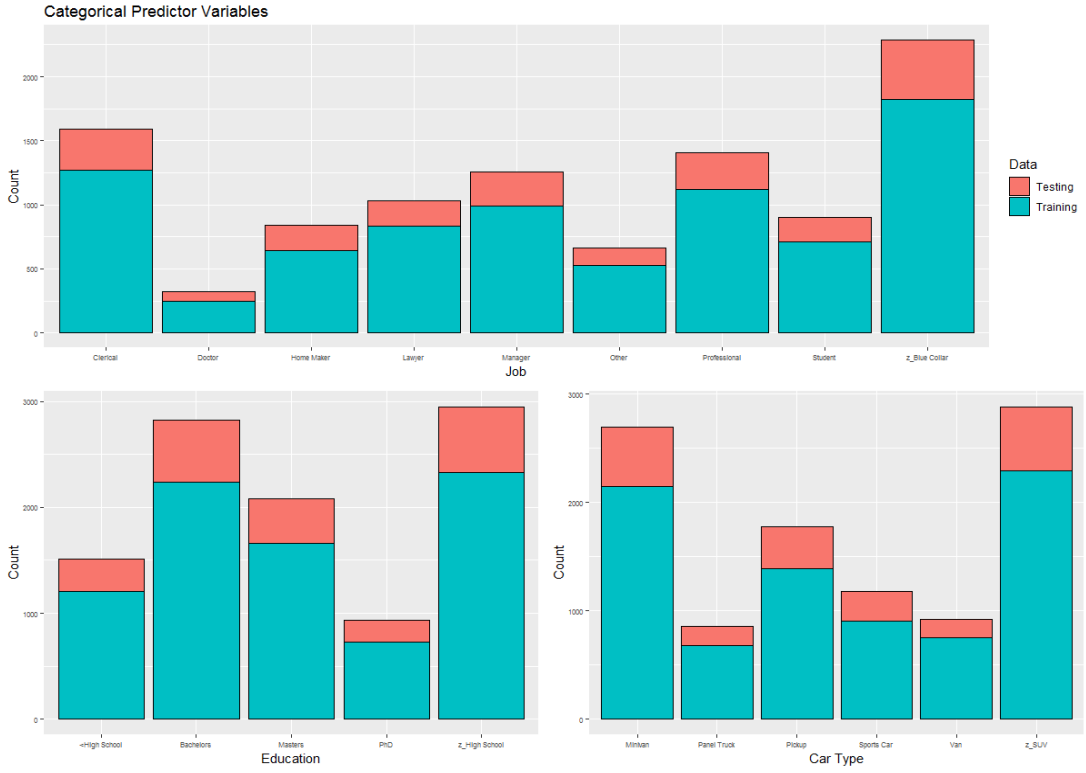

Figure 6 – Categorical Predictor Variables Population

Figure 7 shows the portion of values 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. These missing values are imputed in the Data Preparation section.

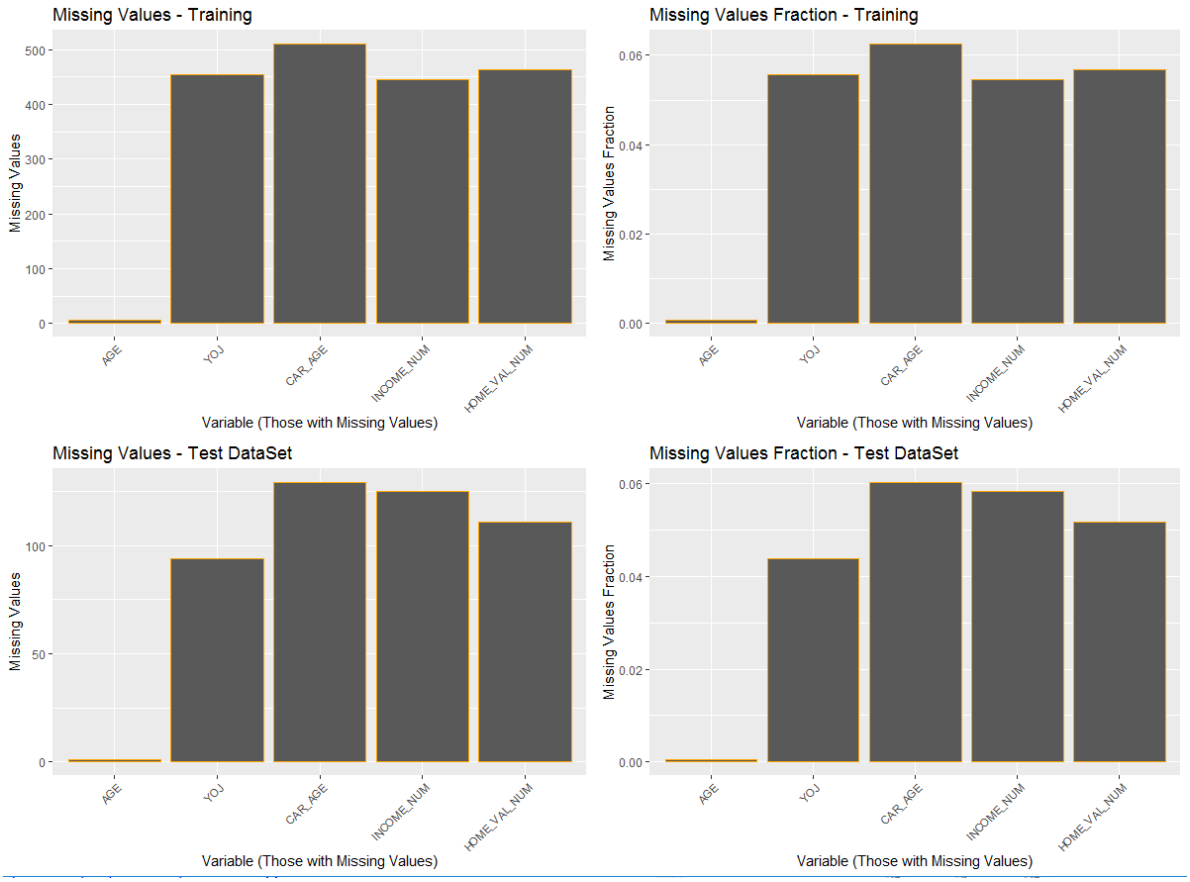

Figure 7 – Missing Value in Training and Test dataset

Data Preparation

Each of the variables with missing values are imputed using the means of the data. To impute missing values of Years on Job, Home value, and Income, mean was calculated at Jobs level to derive more realistic values. Similarly for Car age, mean was calculated by car type. 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 zero to avoid unintuitive results.

New indicator variables created for home value zero and income of zero to server for common values seen in continuous variables. 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.

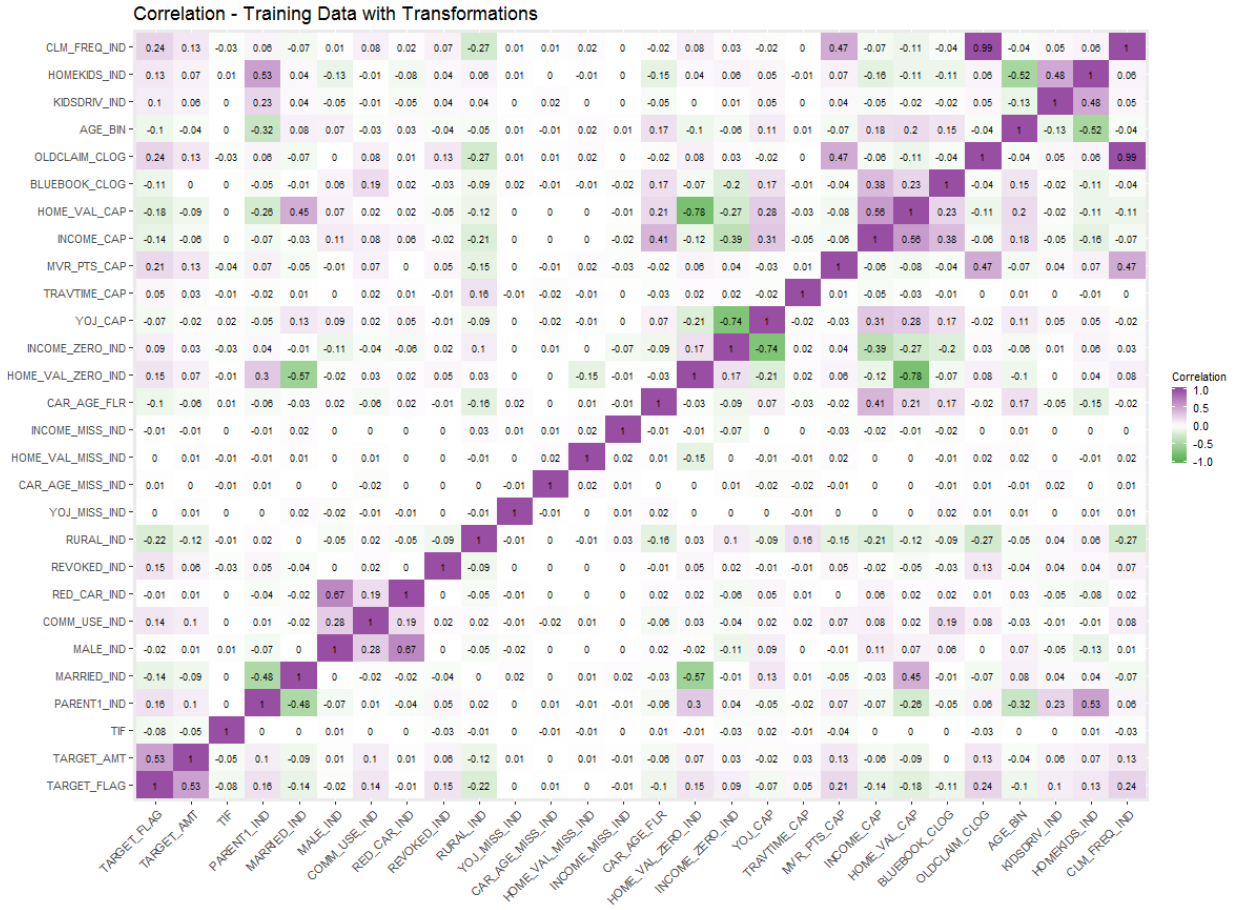

Figure 8 – Correlation - Training Data with Transformations

Build Models

Due to imbalance nature of binary target, 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.

Model1: 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 9 shows the ROC curve. Based on the shape of the curve, there is opportunity for a more complex model to better fit the data.

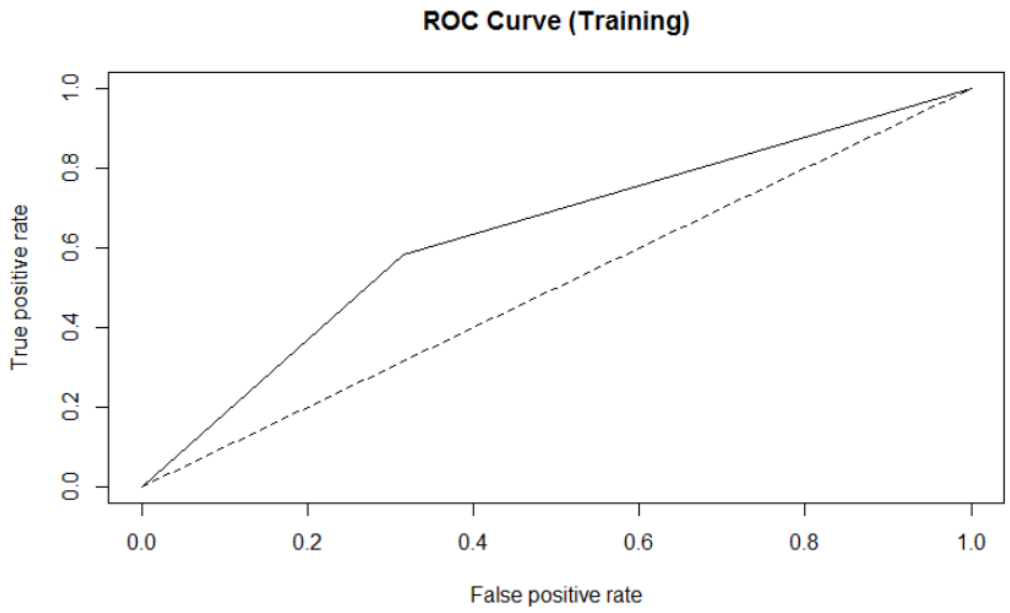

Figure 9 – Model 1 ROC Curve

Model2: This model 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 10 shows the ROC curve for Model 2. The shape of the curve indicates that the model fits the data well.

This model includes all of the transformed predictor variables. an indicator for rural areas, an indicator for the presence of a prior claim, an indicator for being a single parent, car type, an indicator for a revoked license, home value, travel time, time in force, MVR points, an indicator for kids driving,

age, an indicator for commercial use, the bluebook value, an indicator for being married, 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 49 due to its correlation with claim frequency. As a result, it should not be included in a final model.

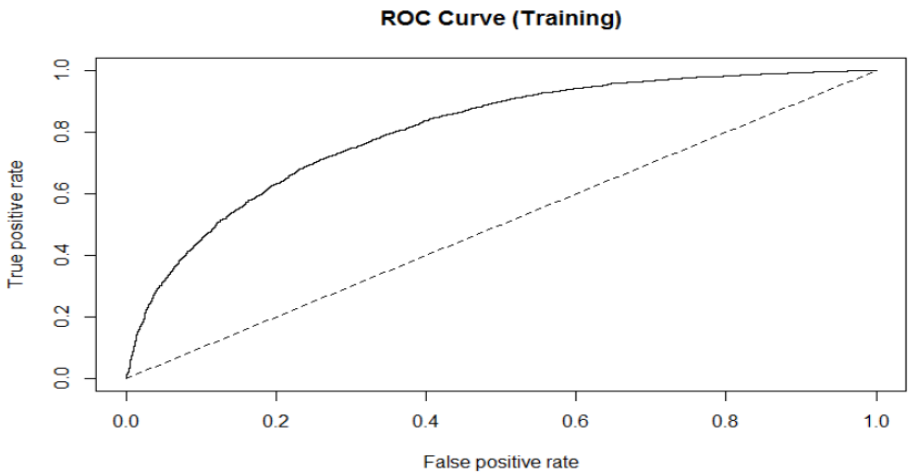

Figure 10 – Model 2 ROC Curve

Model3 (Bonus Model): is the same as Model 2 but rather than a logit-based model, Model 3 uses probit regression. The ROC curve shown in Figure 11 looks similar to Model 2, however it is often easier to interpret at a glance the coefficients generated by the logit model.

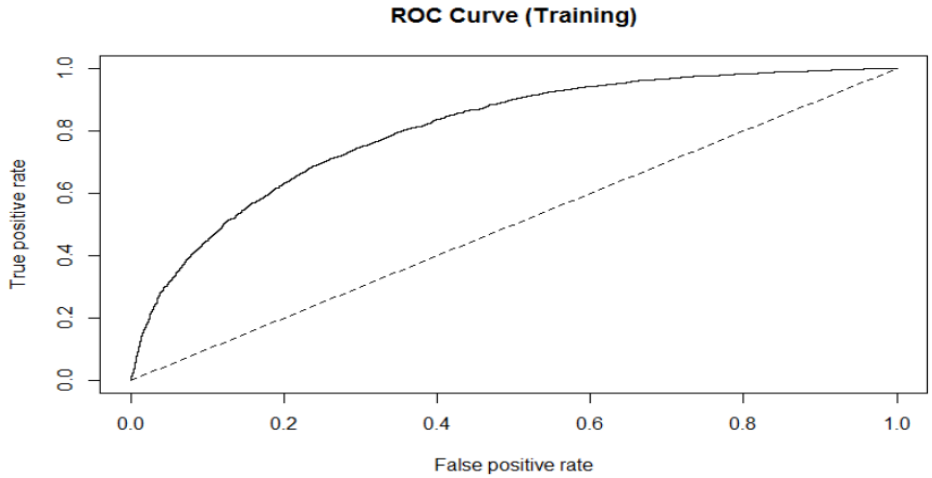

Figure 11 – Model 3 ROC Curve

Model4 (Bonus Model): is the same as Model 2 but rather using a different logistic regression function lrm from R package "rms". Below is the model output. This models AIC is more than model 2 and model 3.

```
Logistic Regression Model

 lrm(formula = TARGET_FACTOR ~ TIF + PARENT1_IND + MARRIED_IND +
     MALE_IND + COMM_USE_IND + REVOKED_IND + RURAL_IND + AGE_IMPUTE +
     YOJ_MISS_IND + CAR_AGE_MISS_IND + HOME_VAL_MISS_IND + INCOME_MISS_IND +
     CAR_AGE_FLR + HOME_VAL_ZERO_IND + INCOME_ZERO_IND + YOJ_CAP +
     TRAVTIME_CAP + MVR_PTS_CAP + INCOME_CAP + HOME_VAL_CAP +
     OLDCLAIM_CLOG + KIDSDRIV_IND + HOMEKIDS_IND, data = train_2)

                       Model Likelihood     Discrimination    Rank Discrim.
                          Ratio Test           Indexes           Indexes
 Obs           8161    LR chi2    1890.94    R2        0.302    C         0.797
  No_Claim     6008    d.f.            23    g         1.535    Dxy       0.595
  Claim        2153    Pr(> chi2) <0.0001    gr        4.639    gamma     0.595
 max |deriv| 1e-08                          gp        0.230    tau-a     0.231
                                            Brier     0.151
```

```
> AIC(glm_fit4)
[1] 7575.018
```

Model5: This model represents the use of the continuous target using ordinary least squares regression. Backwards stepwise selection is used to select variables in the model. Figure 12 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.

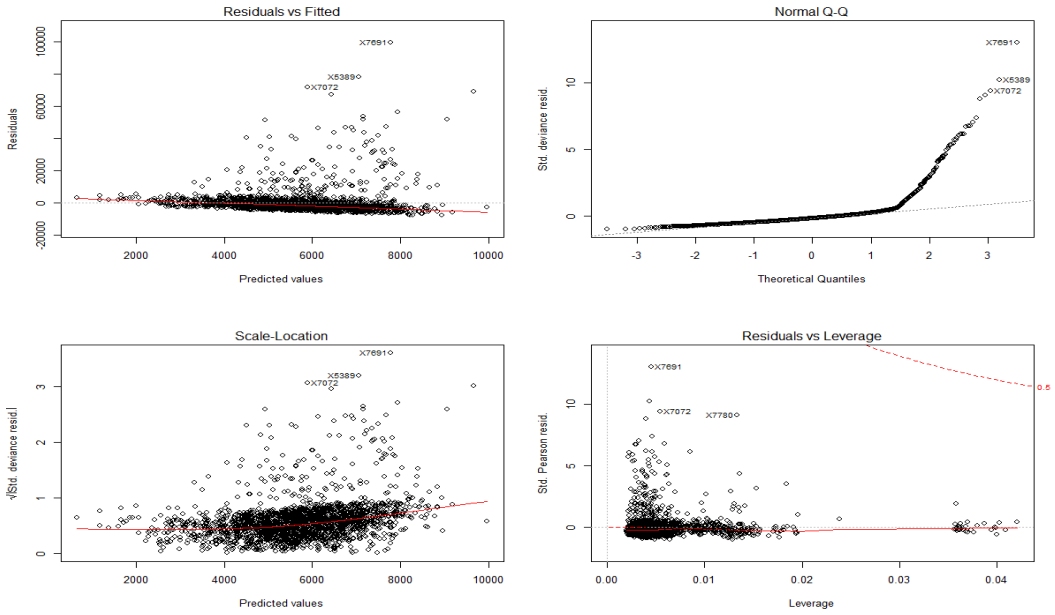

Figure 12 – Model 5 Linear Regression model Diagnostic Plots

Select Model

Fit statistics for the binary target models are included in below table. Model 2 excluding prior claim cost due to multicollinearity (Final Model in the table) 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. The fit statistics and ROC plot (Figure 13) show similar results to Model 2 as a whole.

Model	AIC	BIC	-2*Log Like	AUC	KS
Model1	8953.342	8967.356	8949.342	0.63358	0.2605
Model2	7482.364	7615.5	7444.364	0.80347	0.4492
Model3	7487.429	7620.564	7449.429	0.80341	0.4454
Model4	7575.018	7743.189	7527.018		0.4409
Final Model	7493.1	7619.225	7457.097	0.80246	0.4454

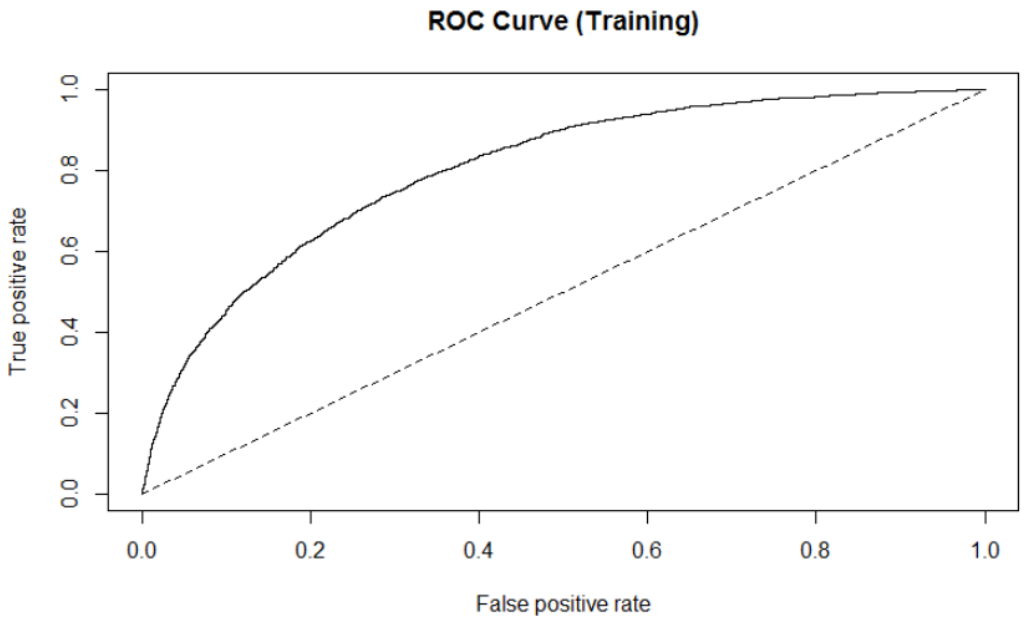

Figure 13 – Final Model ROC Curve

The liner regression model for continues target variable TARGET_AMT AIC is 44634, RMSE is 7754.103 , Rsquared is 0.006711214 , and MAE is 3750.548.

Following is the regression equation of final model which is used to generate scores (predict target flag) on test data set.

P_TARGET_FLAG_1 = 2.749264

*- 2.252701 * RURAL_IND*

*- 0.000001138021 * HOME_VAL_CAP*

*+ 0.1030372 * MVR_PTS_CAP*

*+ 0.9464012 * COMM_USE_IND*

*- 0.3950574 * BLUEBOOK_CLOG*

*+ 0.8738710 * REVOKED_IND*

*+ 0.2473310 * HOMEKIDS_IND*

*+ 0.01621714 * TRAVTIME_CAP*

*- 0.05285844 * TIF*

*+ 1.855246 * CLM_FREQ_IND*

*- 0.000006211482 * INCOME_CAP*

*- 0.5613587 * MARRIED_IND*

*+ 0.5632398 * KIDSDRIV_IND*

*- 0.2516945 * MALE_IND*

*- 0.02293248 * CAR_AGE_FLR*

*- 0.1618346 * OLDCLAIM_CLOG*

*+ 0.2890876 * INCOME_ZERO_IND*

*+ 0.2038418 * PARENT1_IND*

P_TARGET_FLAG = exp(P_TARGET_FLAG_1) / (1 + exp(P_TARGET_FLAG_1))

Below is the equation for scores for TRAGET_AMT on test data set. Note: Predicted target amount has be multiplied with predicted claim probability.

*P_TARGET_AMT = P_TARGET_FLAG * (- 6579.28293*

*- 897.49565 * MARRIED_IND*

*+ 589.44890 * MALE_IND*

*- 697.04293 * REVOKED_IND*

*- 71.56917 * CAR_AGE_FLR*

*- 676.67639 * HOME_VAL_ZERO_IND*

*+ 137.39614 * MVR_PTS_CAP*

*+ 1418.87337 * BLUEBOOK_CLOG*

*+ 1372.14895 * PhD_IND*

*- 551.62170 * HighSchool_IND*

*- 2471.08475 * Doctor_IND*

*- -1155.45783 * Manager_IND)*

Coefficient value and there sings are in sync with the predictor variable impact on the claim and claim amount.

Conclusion

Data preparation is an important step of predictive model building process. Proper handling of missing, outliers and skewed values in training and testing dataset is very important. Subject knowledge can be helpful when doing data exportation and preparation. 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. 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.