Auto Insurance prediction

Sri Seshadri

10/20/2017

Contents

1. Introduction

An insurance company is interested in predicting which customers are likely to be in an accident and what would be the likely payout. The company requires this prediction to price the insurance policy. A predictive model is required to be deployed at point of request for quote or sale. The insurance company has been collecting data on which a predictive model would be trained and tested.

1.1 Analysis Process

The following process steps were used for building a predicitve models:

- Exploratory Data Analysis
  - Perform data quality checks, quantify missing data.
  - Check for systemic loss in data
  - Understand relationships amongst predictors and between target variables and predictors.
  - Create attribure or indicator variables to aid data cleaning.
  - Filter out clean data for feature selection and model building.
- Feature Selection
  - Subset complete records to model wins in season
  - Use different modeling techniques to select candidate predictors.
  - If data is missing for candidate predictors, identify imputing methods.
- Model Building
  - Test models that were build using complete records on the entire data set with imputed data.
  - Compare models based on Deviance, ROC and MAE
  - Check if models make physical sense.
- Initial model deployment
  - Deploy model to predict wins on out of sample data.
  - Discuss models and results with subject matter experts.
  - Fine tune model and re-test
- Final model deployment

1.2 Executive summary

2. Data

The insurance comnpany has data collected from almost 8200 customers. The dictionary of the data is provided in the appendix A.1. Tables 1 and 2 show the summary statistics of numeric and non-numeric features of the data. It is seen that some features have missing values. The missing values may need to be imputed if the features are deemed important predictors of likelihood of customer involving in a crash or payout.

It is seen that the minimum age of car is -3. Which is not rational, the data is filled in with +3 assuming it is a typographical error. Also it is seen that there is white space in the JOB column of the data.

Table 1: Summary statistics

	min	Q1	median	Q3	max	mean	sd	n	missing
INDEX	1	2559.00	5133.00	7745.00	10302.00	5151.87	2978.89	8161	0
TARGET_FLAG	0	0.00	0.00	1.00	1.00	0.26	0.44	8161	0
TARGET_AMT	0	0.00	0.00	1036.00	107586.14	1504.32	4704.03	8161	0
KIDSDRIV	0	0.00	0.00	0.00	4.00	0.17	0.51	8161	0
AGE	16	39.00	45.00	51.00	81.00	44.79	8.63	8155	6

	min	Q1	median	Q3	max	mean	sd	n	missing
HOMEKIDS	0	0.00	0.00	1.00	5.00	0.72	1.12	8161	0
YOJ	0	9.00	11.00	13.00	23.00	10.50	4.09	7707	454
INCOME	0	28096.97	54028.17	85986.21	367030.26	61898.10	47572.69	7716	445
HOME_VAL	0	0.00	161159.53	238724.45	885282.34	154867.29	129123.78	7697	464
TRAVTIME	5	22.45	32.87	43.81	142.12	33.49	15.90	8161	0
BLUEBOOK	1500	9280.00	14440.00	20850.00	69740.00	15709.90	8419.73	8161	0
TIF	1	1.00	4.00	7.00	25.00	5.35	4.15	8161	0
OLDCLAIM	0	0.00	0.00	4636.00	57037.00	4037.08	8777.14	8161	0
CLM_FREQ	0	0.00	0.00	2.00	5.00	0.80	1.16	8161	0
MVR_PTS	0	0.00	1.00	3.00	13.00	1.70	2.15	8161	0
CAR_AGE	-3	1.00	8.00	12.00	28.00	8.33	5.70	7651	510

Table 2: Sanity check of non numeric variables

	# Unique	n	missing	Blanks
PARENT1	2	8161	0	0
MSTATUS	2	8161	0	0
SEX	2	8161	0	0
EDUCATION	5	8161	0	0
JOB	9	8161	0	526
CAR_USE	2	8161	0	0
CAR_TYPE	6	8161	0	0
RED_CAR	2	8161	0	0
REVOKED	2	8161	0	0
URBANICITY	2	8161	0	0

2.1 Exploratory Data Analysis (EDA)

In this section intresting observations in the data are noted and used to characterize the population of customers.

2.1.1 Customer profile

Figure 1 shows the customer portfolio of the insurance company. It is seen from figure 1 that majority of the customers hold blue collar and clerical roles. The income and home values are zero inflated, likely contributed by students and home makers. Nearly 25% of the customers have been involved in a car crash. Most of the customers are with the policy less than 2 years.

It is seen that majority of the customers own cars that are new; less than 2 years old. The older cars are owned by professionals and by the group who has their nature of job missing (potentially not disclosed). Also interestingly that cars that are between 14 and 15 years of age are missing from the population. As seen in histogram in figure 2.

Figure 3 explores the missing JOB data. The missing JOB values lines up with the population of white collar jobs. From the income, age and year on the job perspective. Also the zero inflation in income is contributed by students and home makers. The jobs above the red dashed lines the income boxplot is defined as white collar.

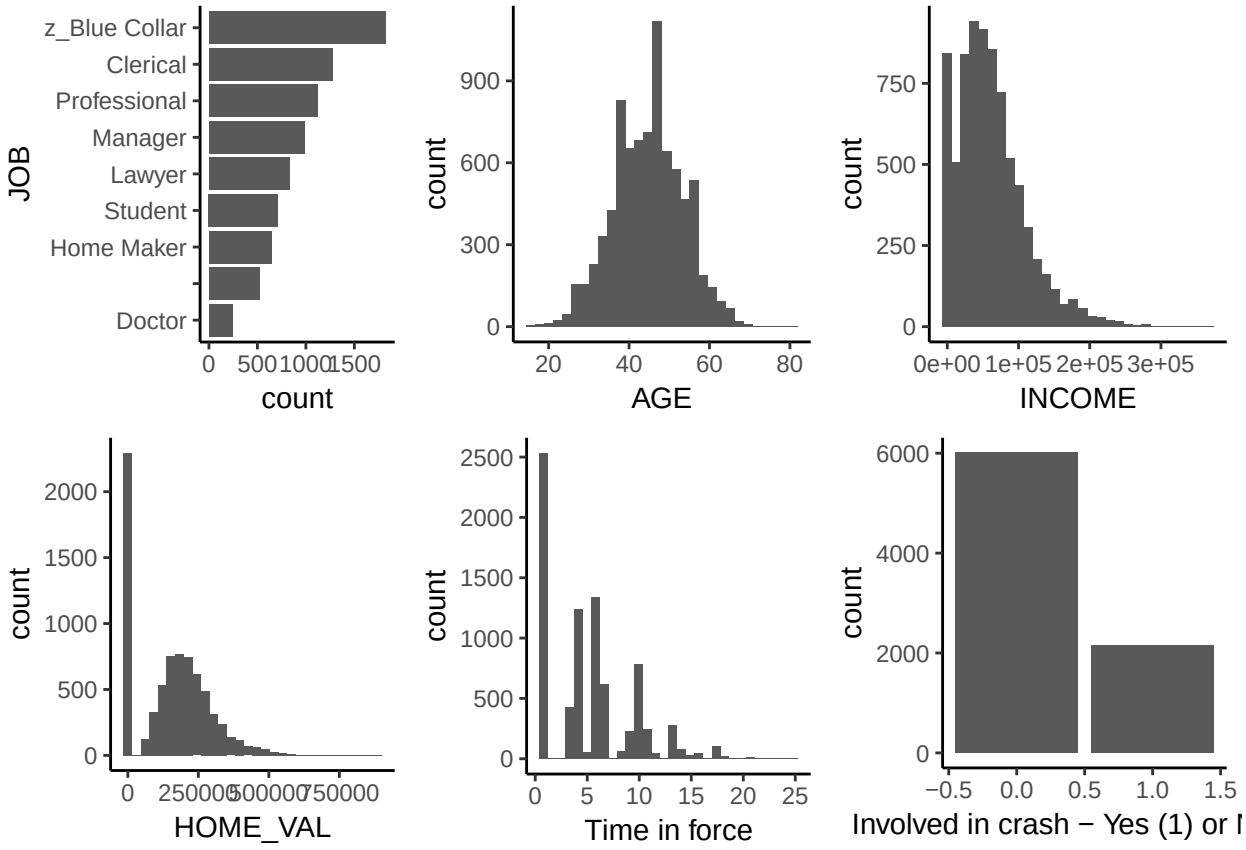


Figure 1: Customer portfolio

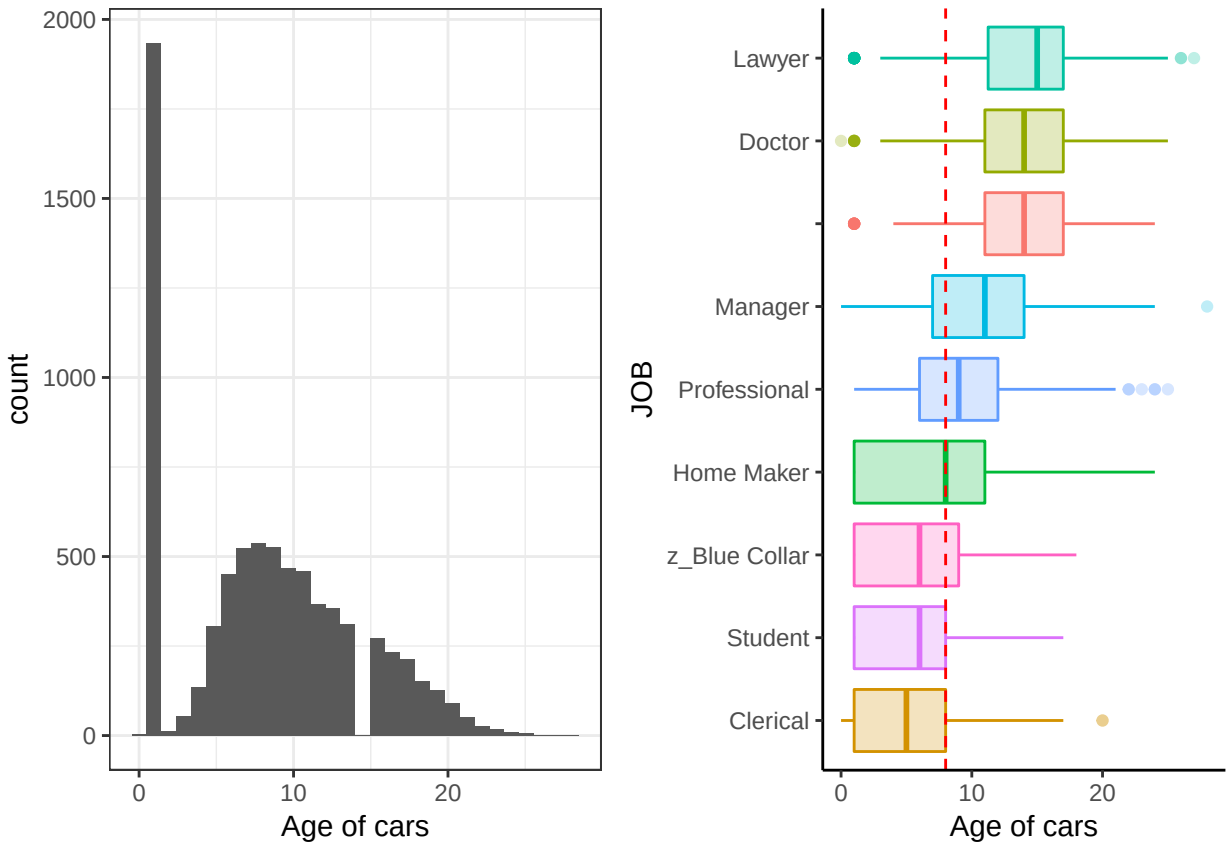


Figure 2: car age

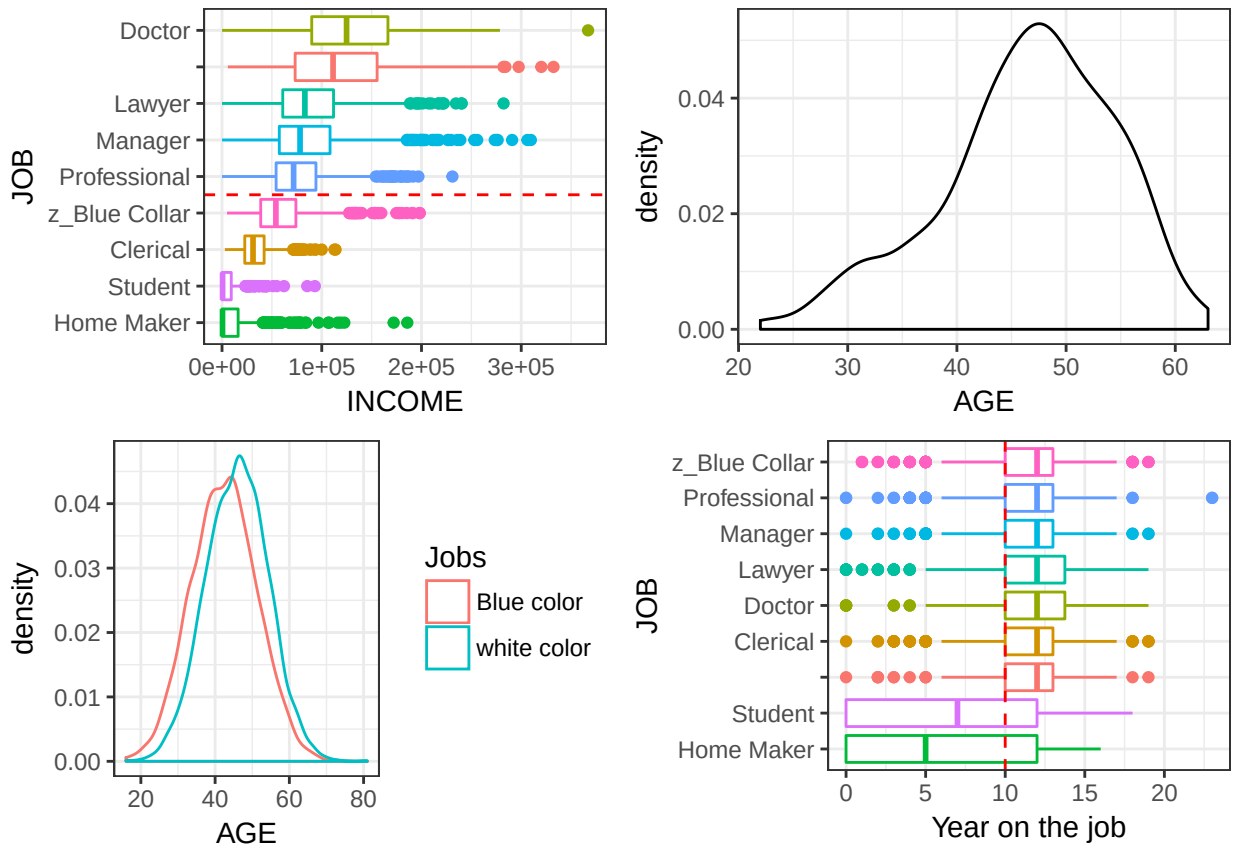


Figure 3: Missing JOB profile

2.1.2 Attribute variables

The categorical variables are transformed into indicator variables to be used in modeling. The missing data points are manifested as indicator variables.

2.1.3 Handling missing data

The complete cases would be used to determine key features required for modeling. If data for key features are missing, an imputation strategy would be determined.

3. Feature Selection

In this section we consider various feature selection methodologies such as 1. Decision Trees, 2. Peanalized model - Lasso

3.1 Training and Test data partition

In order to test if the feature selection are really useful, feature selections need to be cross validated on a hold out test data set. 80% of the data is used as training and the rest is used as hold out for testing. A stratified sampling method is used to capture identical percentage of samples involved in crash.

3.2 Decision Tree

Decision tree model is fitted on the training set to identify stand out splits in the data based on Gini index. The decision tree is shown in figure 4. Bagging technique is used to minimize variance in the model to ensure we have a reliable feature selection. The important features are shown in figure 5.

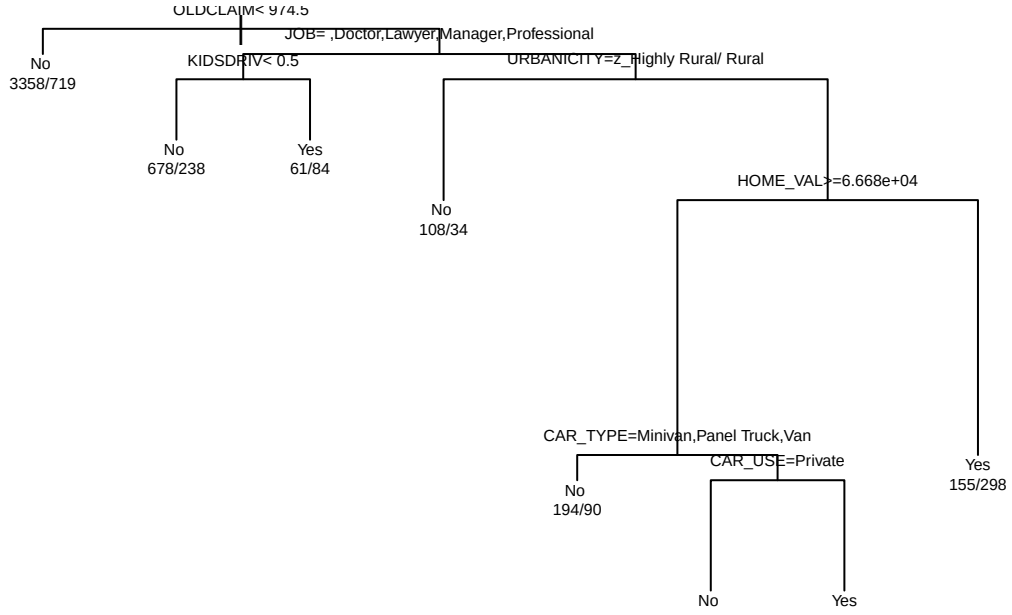


Figure 4: Decision Tree

```
##
## Call:
##  randomForest(formula = as.factor(TARGET_FLAG2) ~ . - TARGET_FLAG,      data = training, mtry = 28, r
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 28
##
##         OOB estimate of  error rate: 21.1%
## Confusion matrix:
##       No Yes class.error
## No  3495 312  0.08195429
## Yes  777 578  0.57343173
```

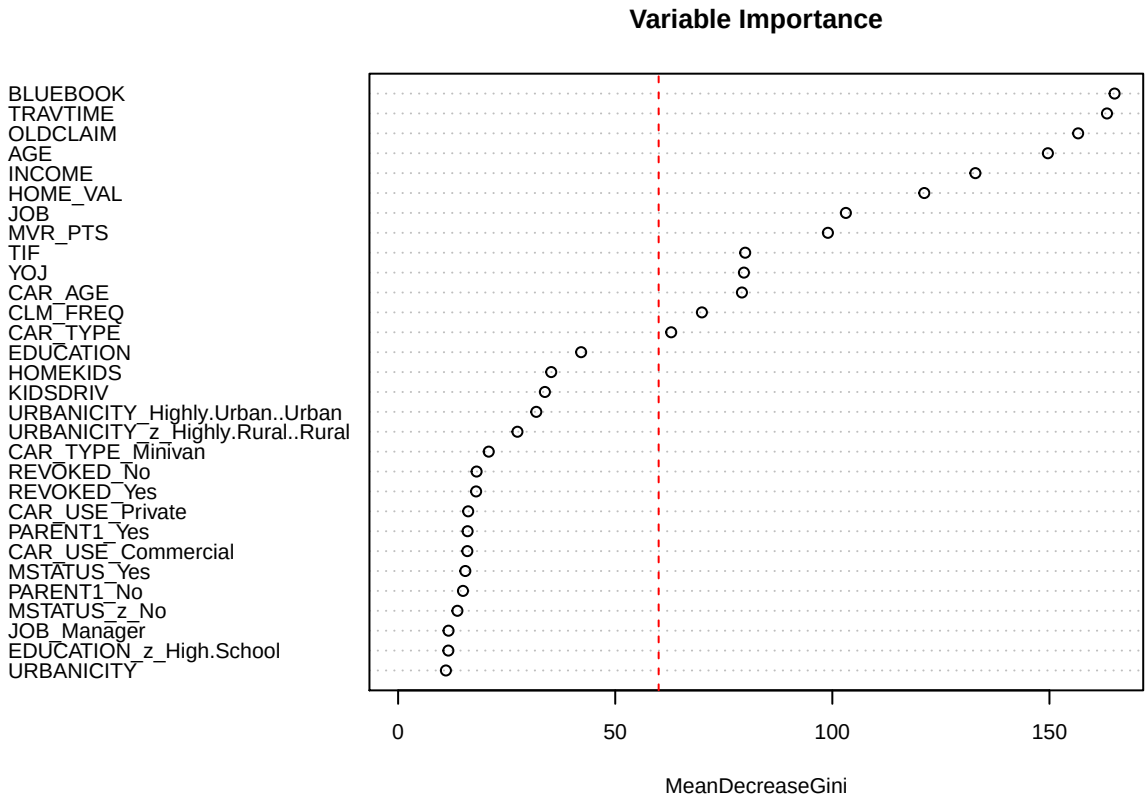


Figure 5: Variable Importance Plot and test prediction

The decision tree was used to predict the hold out test data and the AUC was found to be 80 % as shown in figure 6. Therefore we'll use the top 13 predictors as shown in figure 4. In the next section we will explore penalized models to gather important predictors.

3.3 Penalized model - Lasso

The variable selection property of Lasso is used to aid with automated variable selection. Again the model is trained on training data set and tested on a hold out dataset, as in the decision tree method. However a 10 fold cross validation method was used to identify the optimal penalization parameter - lambda. Figure 7 shows the coefficients that are not zero in the decreasing order of absolute value of coefficients. The predictors falling above the red dashed line, drawn at the elbow in figure 7 are deemed good predictors.

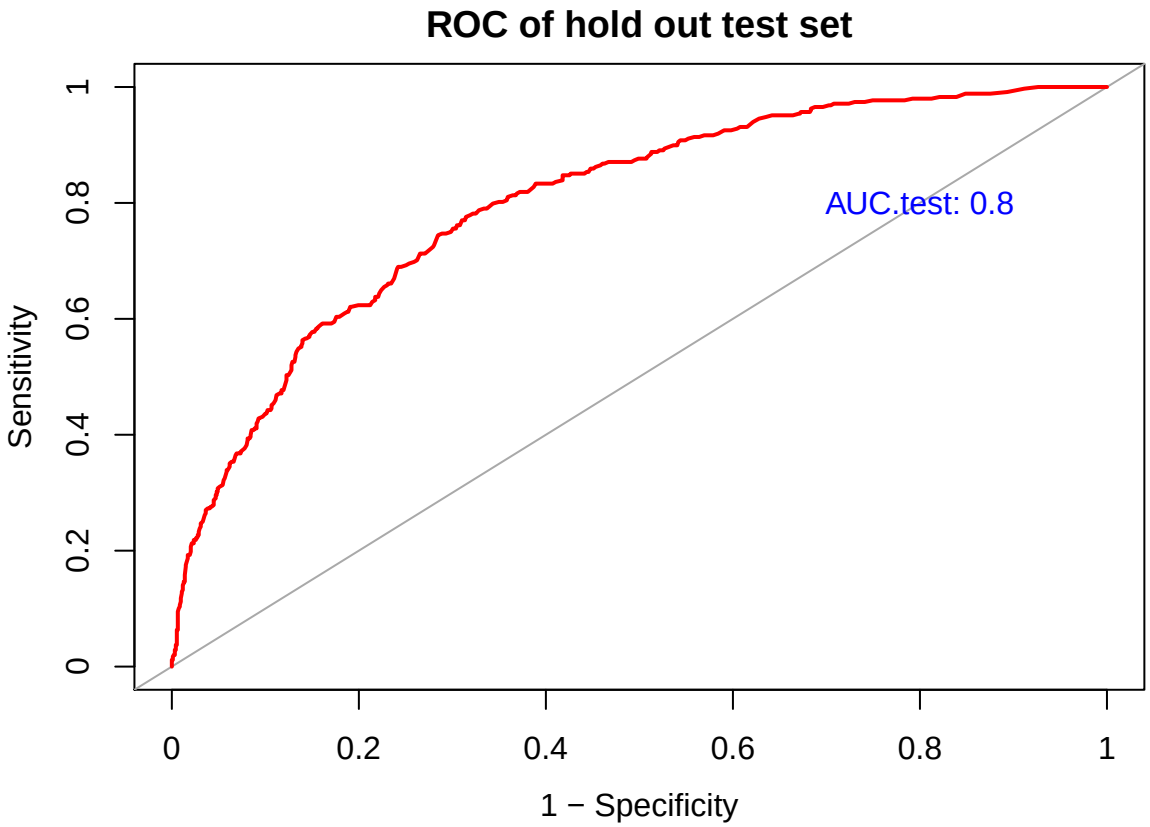


Figure 6: Prediction of bagged decision tree

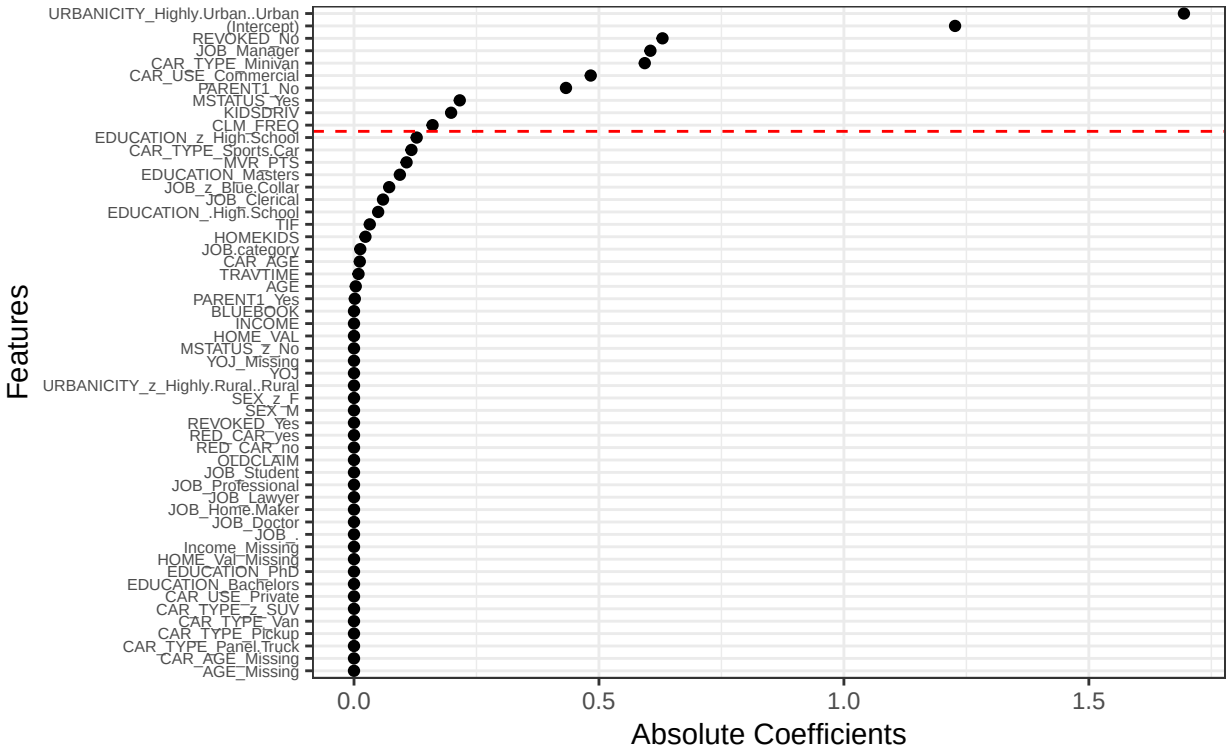


Figure 7: Feature selection - Lasso

3.3.1 Logistic regression as penalized models - Lasso

While the peanalized logistic regression model is used for feature selection, it may be used for prediction as well. All the predictors corresponding to non zero coefficients are considered as predictors. The ROC curves for the training and test samples are shown in figure 8 and the Gain chart is shown in figure 9.

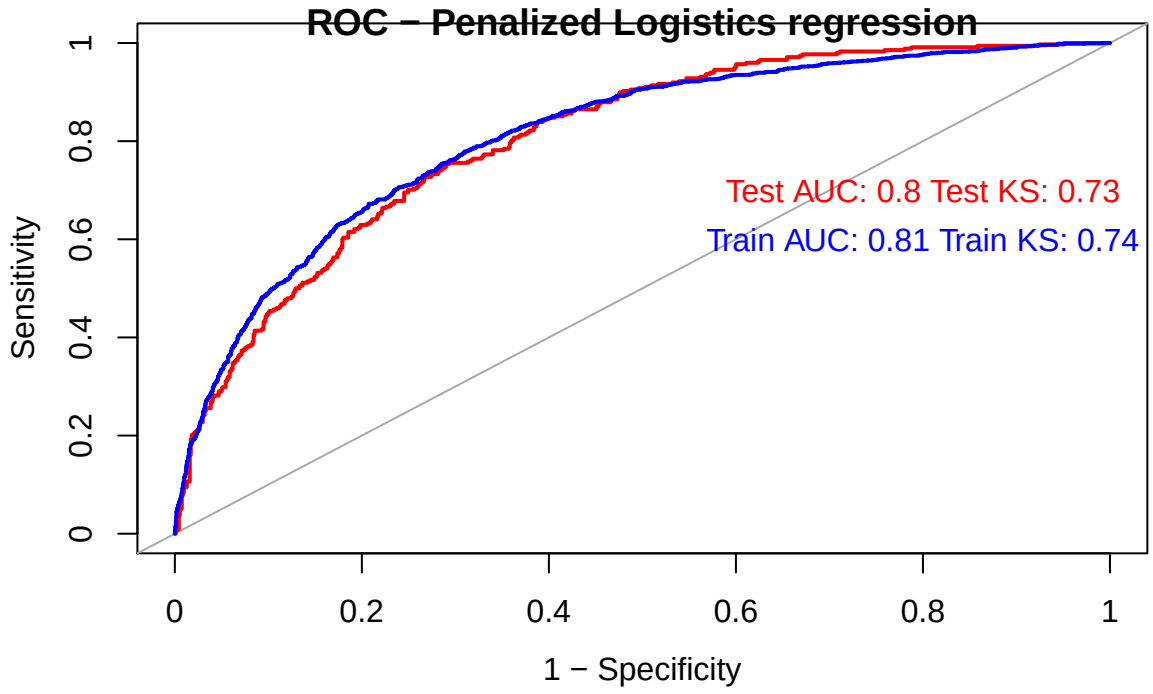
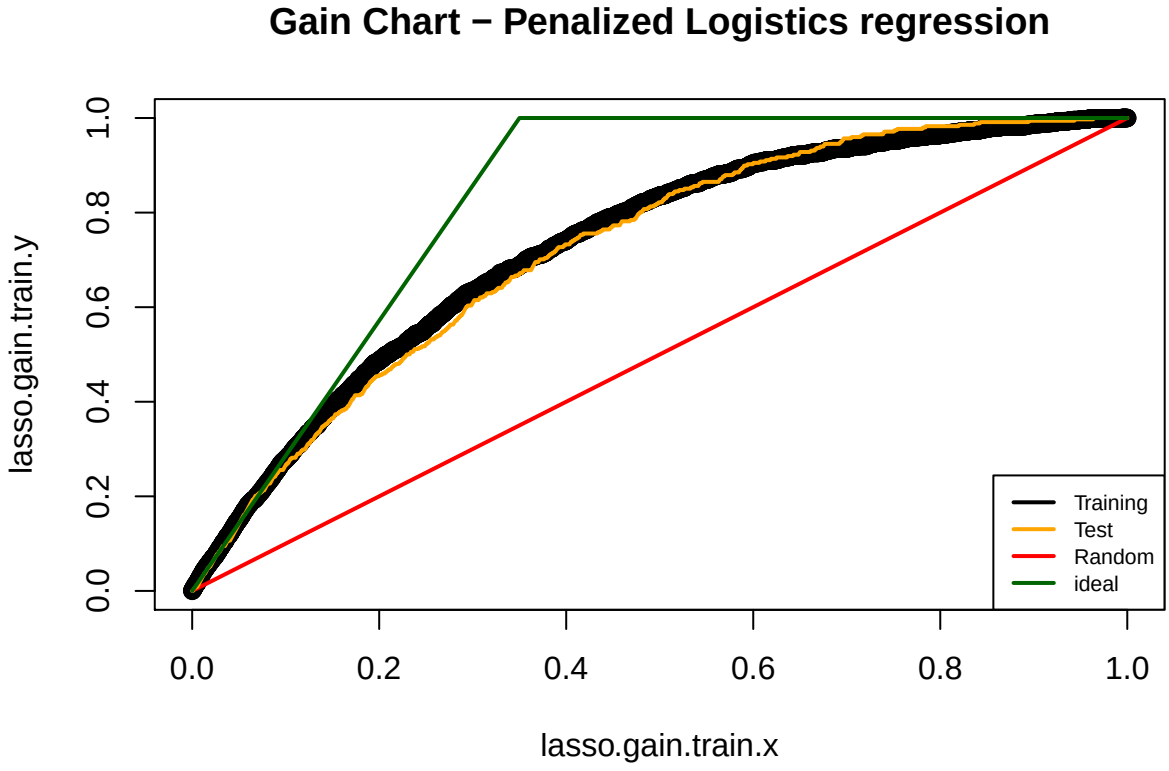


Figure 8: Lasso ROC curves for test and training data



It can

be seen from figure 8 and 9, lasso regression has 74% AUC and is better than a random model. The test performance is identical to the training performance. The misclassification is about 23%. However performance on the records with missing values is yet evaluated and will be discussed in the modeling section.

```
##            Reference
## Prediction    0    1
##          0 3616  902
##          1  191  453
```

Table 3: confusion matrix statistics - lasso logistic train

Accuracy	0.79
Kappa	0.34
AccuracyLower	0.78
AccuracyUpper	0.80
AccuracyNull	0.74
AccuracyPValue	0.00
McnemarPValue	0.00

```
##            Reference
## Prediction   0   1
##          0 886 240
##          1  52 108
```

Table 4: confusion matrix statistics - lasso logistic test

	0.77
Accuracy Kappa	0.31
AccuracyLower	0.75
AccuracyUpper	0.80
AccuracyNull	0.73
AccuracyPValue	0.00
McnemarPValue	0.00
## 3.4. Features	selection conclusion.
The following pre	dictors are deemed useful for model- ing pur- poses after the above feature selec- tion process.

	0.77
Accuracy Kappa	0.31
AccuracyLower	0.75
AccuracyUpper	0.80
AccuracyNull	0.73
AccuracyPValue	0.00
McnemarPValue	0.00

Table: Selected Features from Decision Tree and Lasso regression

BLUEBOOK TRAVTIME OLDCLAIM
AGE INCOME HOME_VAL
JOB MVR_PTS TIF
YOJ CAR_AGE CLM_FREQ
URBANICITY_Highly.Urban..Urban REVOKED_No CAR_TYPE_Minivan CAR_USE_Commercial
PARENT1_No MSTATUS_Yes
KIDSDRIV

4. Modeling

4.1. Logistic regression - using all features selected in section 3

A logistic regression with the above selected predcitors (section 3.4) is fitted. Indicator variables for JOB is used for modeling. The summary of the fit is shown below. It can be seen from the Chi-squared goodness of fit that the model is adequate. There are some predictors whose regression coefficients are not significantly different from 0. In the next section a model without predictors with regression coeffiecients that are not significantly different from 0 would be explored.

The model has an AUC of 81% and KS statistic whose p value is zero, indicating an adequate fit. The model has misclassification of 20% for both the training and the test samples.

The model is very identical to the Lasso regression. Choosing the variables the top 10 features from the Lasso variable selection along with features from the decision tree does not yield a model that's different from the Lasso itself.

```
##
## Call:
## glm(formula = TARGET_FLAG ~ ., family = binomial, data = training.logistic)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5611  -0.7159  -0.4098   0.6132   3.1197
##
## Coefficients: (2 not defined because of singularities)
##                                  Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)                    -9.598e-01  3.063e-01  -3.134 0.001724 **
## BLUEBOOK                       -2.573e-05  5.136e-06  -5.009 5.48e-07 ***
## AGE                            -6.339e-03  4.584e-03  -1.383 0.166708
## YOJ                            -6.790e-03  1.040e-02  -0.653 0.513890
## URBANICITY_Highly.Urban..Urban  2.212e+00  1.380e-01  16.035  < 2e-16 ***
## CAR_USE_Commercial              6.378e-01  9.771e-02   6.527 6.69e-11 ***
## KIDSDRIV                        3.512e-01  7.006e-02   5.012 5.37e-07 ***
## TRAVTIME                        1.531e-02  2.362e-03   6.482 9.03e-11 ***
## INCOME                         -2.250e-06  1.367e-06  -1.646 0.099758 .
## MVR_PTS                         1.233e-01  1.706e-02   7.227 4.93e-13 ***
## CAR_AGE                        -1.992e-02  8.108e-03  -2.456 0.014034 *
## REVOKED_No                     -8.805e-01  1.170e-01  -7.526 5.22e-14 ***
## PARENT1_No                     -4.979e-01  1.231e-01  -4.044 5.25e-05 ***
## OLDCLAIM                       -5.777e-06  4.936e-06  -1.170 0.241828
## HOME_VAL                       -1.215e-06  4.352e-07  -2.792 0.005240 **
## TIF                            -5.167e-02  9.171e-03  -5.635 1.75e-08 ***
## CLM_FREQ                        2.023e-01  3.526e-02   5.736 9.68e-09 ***
## CAR_TYPE_Minivan               -8.280e-01  9.421e-02  -8.789  < 2e-16 ***
## MSTATUS_Yes                    -3.514e-01  1.045e-01  -3.361 0.000776 ***
## JOB_.                          -4.831e-01  1.810e-01  -2.669 0.007611 **
## JOB_Clerical                    1.437e-01  1.287e-01   1.116 0.264352
## JOB_Doctor                     -4.345e-01  2.850e-01  -1.525 0.127318
## JOB_Home.Maker                 -2.067e-01  1.866e-01  -1.108 0.267922
## JOB_Lawyer                     -2.361e-01  1.783e-01  -1.324 0.185534
## JOB_Manager                    -1.046e+00  1.608e-01  -6.506 7.73e-11 ***
## JOB_Professional               -2.567e-01  1.368e-01  -1.876 0.060607 .
## JOB_Student                    -3.779e-02  1.648e-01  -0.229 0.818615
## JOB_z_Blue.Collar                      NA         NA      NA       NA
## JOB.category                           NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5943.0  on 5161  degrees of freedom
## Residual deviance: 4628.8  on 5135  degrees of freedom
## AIC: 4682.8
##
## Number of Fisher Scoring iterations: 5
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: TARGET_FLAG
##
## Terms added sequentially (first to last)
##
##
##                              Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                          5161     5943.0
## BLUEBOOK                      1    65.15      5160     5877.9 6.930e-16
## AGE                           1    54.87      5159     5823.0 1.289e-13
## YOJ                           1     5.07      5158     5817.9 0.0243864
```

```
## URBANICITY_Highly.Urban..Urban  1   339.42      5157     5478.5 < 2.2e-16
## CAR_USE_Commercial              1   156.67      5156     5321.8 < 2.2e-16
## KIDSDRIV                        1    48.42      5155     5273.4 3.440e-12
## TRAVTIME                        1    42.13      5154     5231.3 8.553e-11
## INCOME                          1    71.17      5153     5160.1 < 2.2e-16
## MVR_PTS                         1   130.70      5152     5029.4 < 2.2e-16
## CAR_AGE                         1    14.56      5151     5014.9 0.0001359
## REVOKED_No                      1    71.47      5150     4943.4 < 2.2e-16
## PARENT1_No                      1    67.99      5149     4875.4 < 2.2e-16
## OLDCLAIM                        1     3.42      5148     4872.0 0.0644200
## HOME_VAL                        1    29.48      5147     4842.5 5.651e-08
## TIF                             1    30.30      5146     4812.2 3.699e-08
## CLM_FREQ                        1    34.23      5145     4778.0 4.898e-09
## CAR_TYPE_Minivan                1    75.86      5144     4702.1 < 2.2e-16
## MSTATUS_Yes                     1    11.71      5143     4690.4 0.0006231
## JOB_.                           1     0.75      5142     4689.6 0.3870173
## JOB_Clerical                    1    10.88      5141     4678.8 0.0009745
## JOB_Doctor                      1     0.01      5140     4678.8 0.9033436
## JOB_Home.Maker                  1     0.01      5139     4678.8 0.9382485
## JOB_Lawyer                      1     2.57      5138     4676.2 0.1090987
## JOB_Manager                     1    43.84      5137     4632.3 3.559e-11
## JOB_Professional                1     3.52      5136     4628.8 0.0606790
## JOB_Student                     1     0.05      5135     4628.8 0.8185791
## JOB_z_Blue.Collar               0     0.00      5135     4628.8
## JOB.category                    0     0.00      5135     4628.8
##
## NULL
## BLUEBOOK                       ***
## AGE                            ***
## YOJ                            *
## URBANICITY_Highly.Urban..Urban ***
## CAR_USE_Commercial             ***
## KIDSDRIV                       ***
## TRAVTIME                       ***
## INCOME                         ***
## MVR_PTS                        ***
## CAR_AGE                        ***
## REVOKED_No                     ***
## PARENT1_No                     ***
## OLDCLAIM                       .
## HOME_VAL                       ***
## TIF                            ***
## CLM_FREQ                       ***
## CAR_TYPE_Minivan               ***
## MSTATUS_Yes                    ***
## JOB_.
## JOB_Clerical                   ***
## JOB_Doctor
## JOB_Home.Maker
## JOB_Lawyer
## JOB_Manager                    ***
## JOB_Professional               .
## JOB_Student
## JOB_z_Blue.Collar
```

```
## JOB.category
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## [1] "Goodness of fit test... P-Value 0"
```

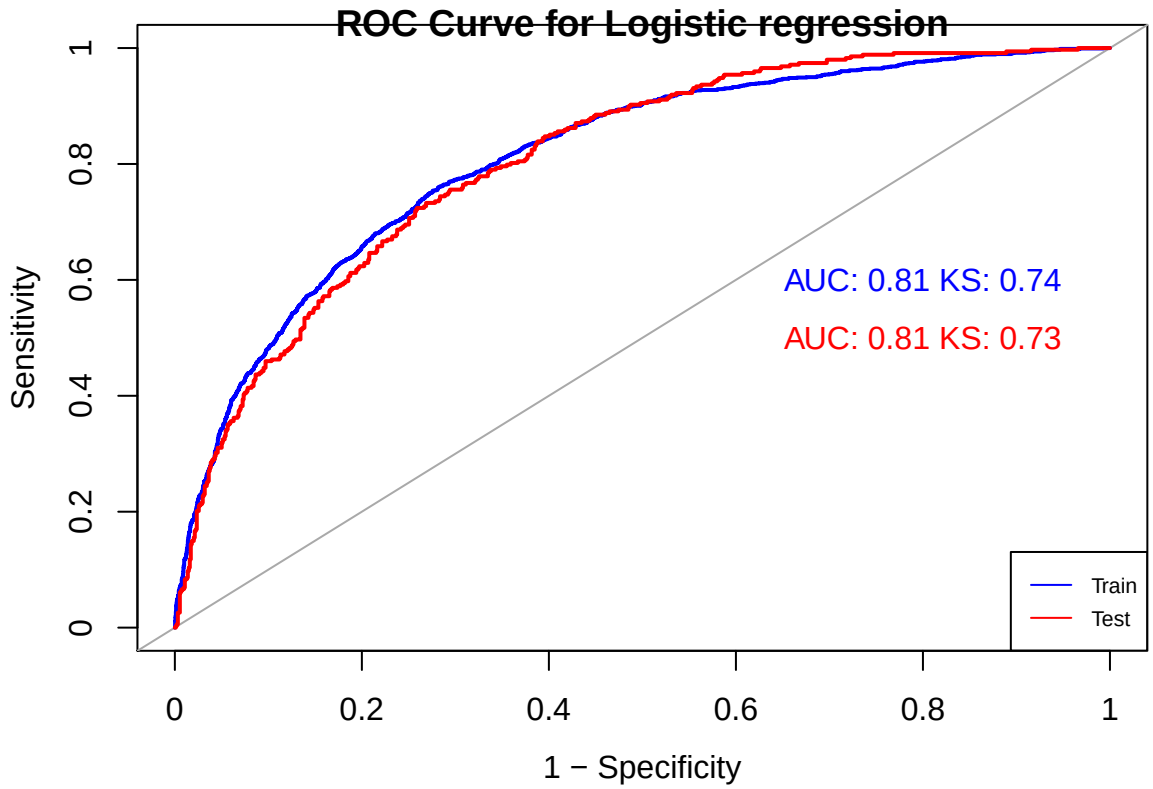


Figure 9: ROC curve for Logistic regression - full chosen predictors

```
##            Reference
## Prediction    0    1
##          0 3536  784
##          1  271  571
```

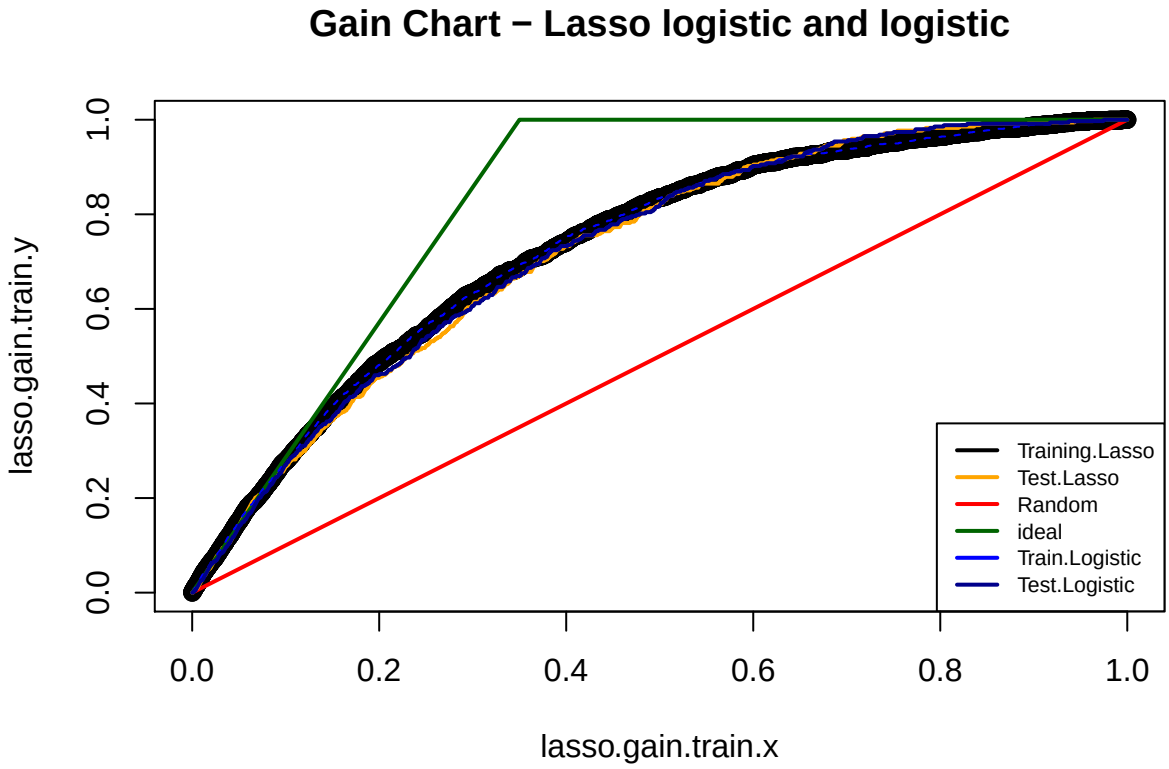
Table 5: Confusion matrix statistics for Losgistic regression - Training sample

Accuracy	0.80
Kappa	0.40
AccuracyLower	0.78
AccuracyUpper	0.81
AccuracyNull	0.74
AccuracyPValue	0.00
McnemarPValue	0.00

```
##            Reference
## Prediction   0   1
##          0 869 208
##          1  69 140
```

Table 6: Confusion matric statistics for Logistic regression - Test sample

Accuracy	0.78
Kappa	0.38
AccuracyLower	0.76
AccuracyUpper	0.81
AccuracyNull	0.73
AccuracyPValue	0.00
McnemarPValue	0.00



###

4.1.1 Model performance on data including the missing values.

The performance of the model on complete cases seem to be good and identical to the Lasso logistic regression model. However, the model does require features that has data missing. We'll use the KNN imputation method to impute data.

```
##            Reference
## Prediction    0    1
##          0 1106  241
##          1   81  204
```

Table 7: Confusion matrix stats for Logistic regression on Imputed test data

Accuracy	0.8026961
Kappa	0.4395879
AccuracyLower	0.7825448
AccuracyUpper	0.8217522
AccuracyNull	0.7273284
AccuracyPValue	0.0000000

McnemarPValue	0.0000000

The model performance is identical to what was seen in above sections.

4.2 Logistic regression - using features selected using decision tree alone.

4.2.1 Indicator variables

As seen in figure 4, customers with house prices less than $66680 and claims greater 974.5 are more likely to be in a crash. It'll be useful to create indicator variables denoting this for modeling. Also the JOBs are categorized as blue collar and white collar. White collar being indicated as 1 and blue collar as 0.

4.2.2 Model results

It is seen from the model results that the goodness of fit is good. Also the signs of coefficients makes theoretical sense. So does models discussed above.

```
##
## Call:
## glm(formula = TARGET_FLAG ~ ., family = binomial, data = training.dt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7702  -0.7668  -0.5634   0.8979   2.6877
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -8.212e-01  2.251e-01  -3.648 0.000264 ***
## BLUEBOOK        -1.551e-05  4.865e-06  -3.187 0.001438 **
## TRAVTIME         7.018e-03  2.190e-03   3.205 0.001350 **
## AGE             -1.159e-02  4.138e-03  -2.800 0.005116 **
## INCOME          -4.784e-06  1.143e-06  -4.184 2.86e-05 ***
## MVR_PTS          1.145e-01  1.684e-02   6.798 1.06e-11 ***
## TIF             -4.068e-02  8.856e-03  -4.594 4.35e-06 ***
## YOJ             -2.530e-03  9.030e-03  -0.280 0.779295
## CAR_AGE         -1.929e-02  6.910e-03  -2.792 0.005246 **
## CLM_FREQ         3.084e-02  5.108e-02   0.604 0.545987
## HOME_VAL_Bucket  6.377e-01  7.562e-02   8.433  < 2e-16 ***
## OLDCLAIM_Bucket  7.950e-01  1.305e-01   6.090 1.13e-09 ***
## JOB.category     2.247e-01  9.543e-02   2.355 0.018529 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5531.6  on 4825  degrees of freedom
## Residual deviance: 4963.1  on 4813  degrees of freedom
##   (1295 observations deleted due to missingness)
## AIC: 4989.1
##
## Number of Fisher Scoring iterations: 4
```

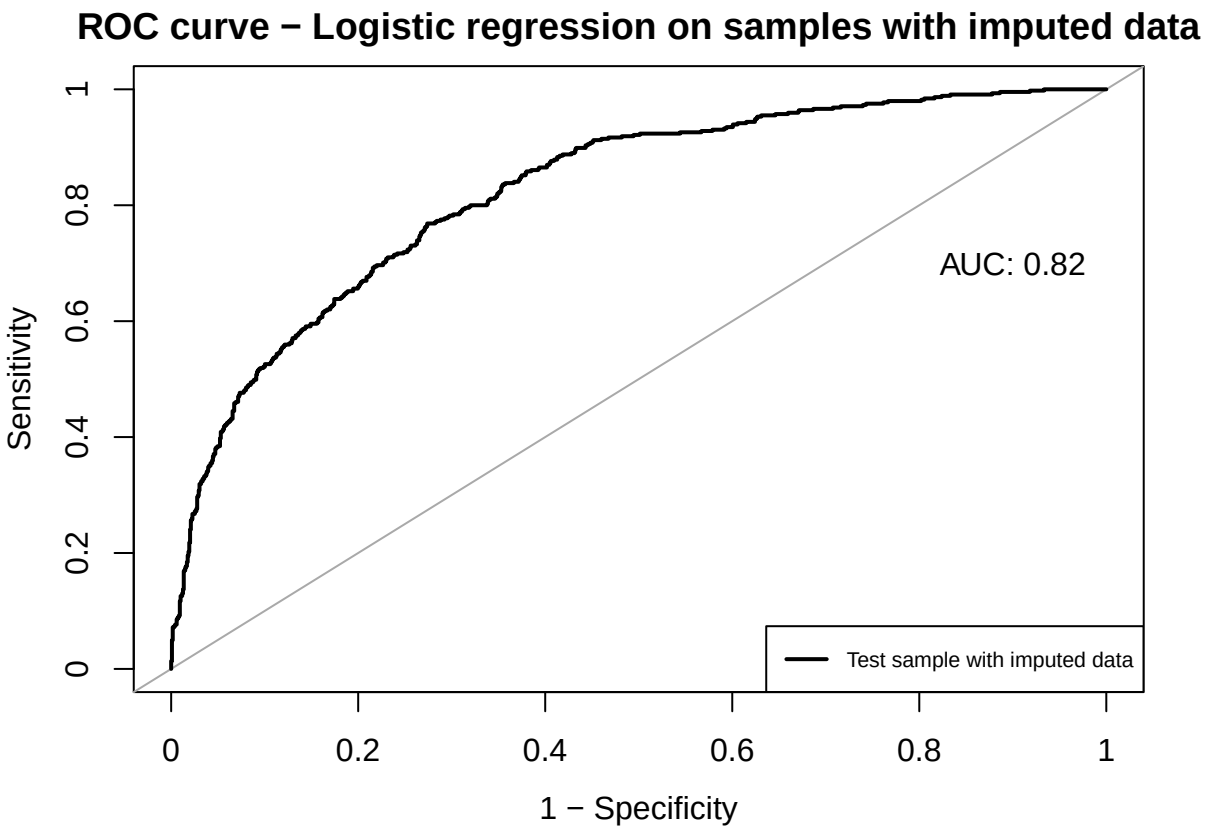


Figure 10: Model performance on imputed test data

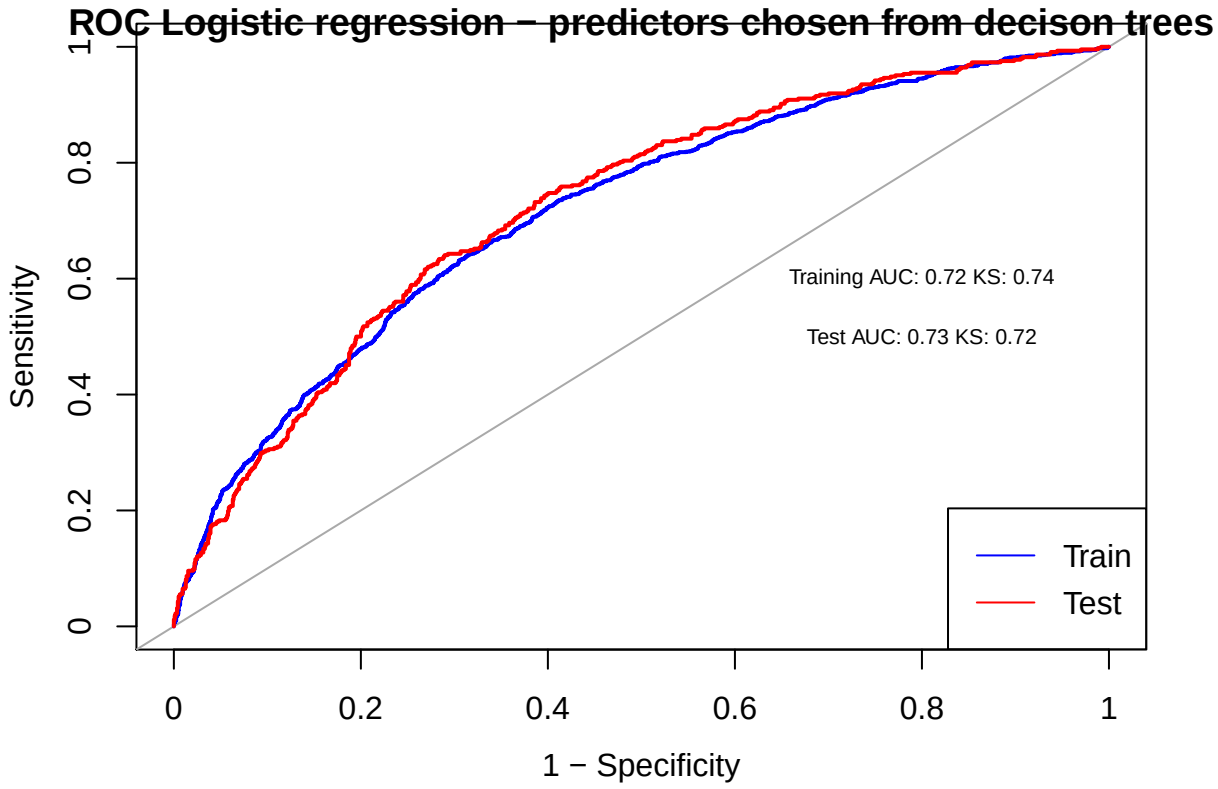


Figure 11: ROC Logistic regression- predictors from Decision Tree

```
## [1] "Goodness of fit test p Value : 0"
```

```
##            Reference
## Prediction    0    1
##          0 3406  991
##          1  165  264
```

Table 8: Confusion matrix stats for Logistic regression on predictor selection from decision trees - Training set

Accuracy	0.7604642
Kappa	0.2086947
AccuracyLower	0.7481636
AccuracyUpper	0.7724488
AccuracyNull	0.7399503
AccuracyPValue	0.0005605
McnemarPValue	0.0000000

```
##            Reference
## Prediction    0    1
##          0 1109  365
##          1   65   83
```

Table 9: Confusion matrix stats for Logistic regression on predictor selection from decision trees - Test set

Accuracy	0.7348952
Kappa	0.1638213
AccuracyLower	0.7126939
AccuracyUpper	0.7562431
AccuracyNull	0.7237978
AccuracyPValue	0.1656200
McnemarPValue	0.0000000