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.

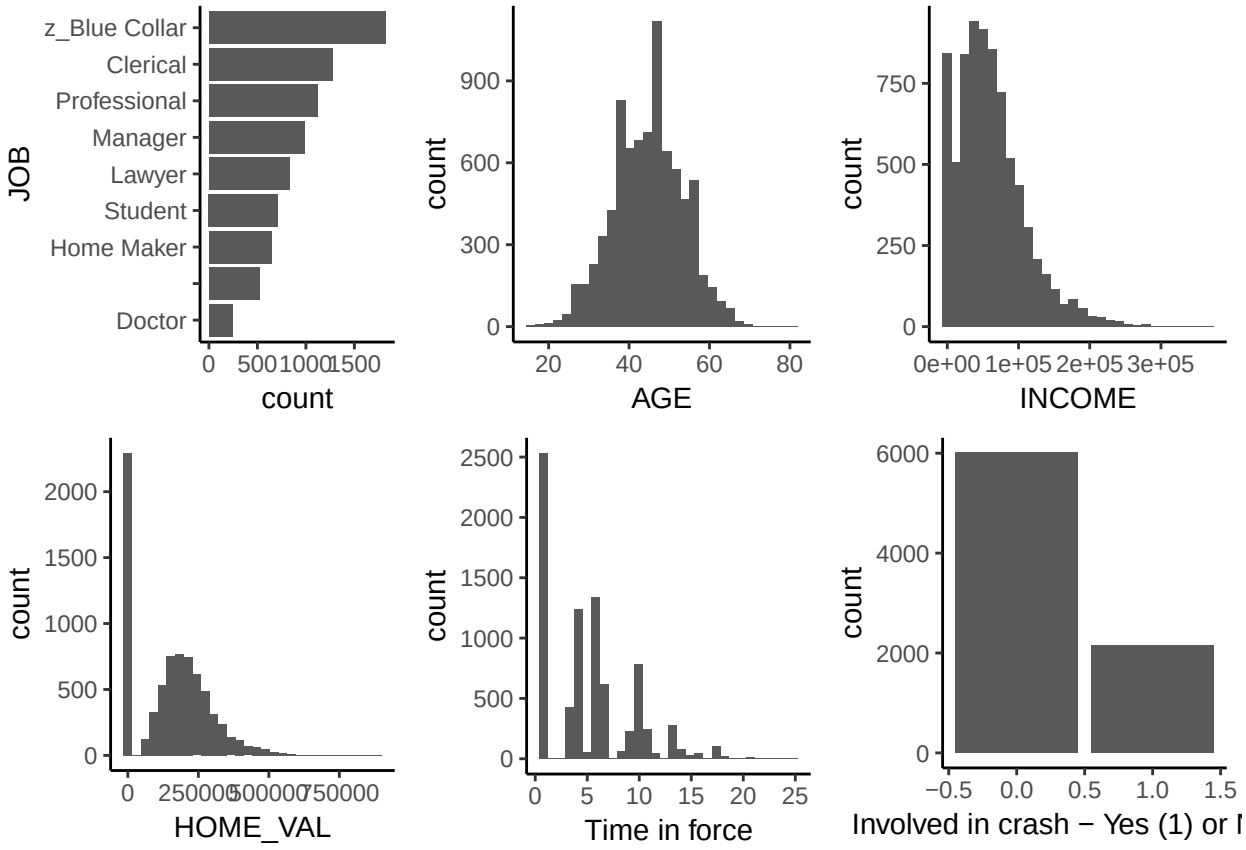


Figure 1: Customer portfolio

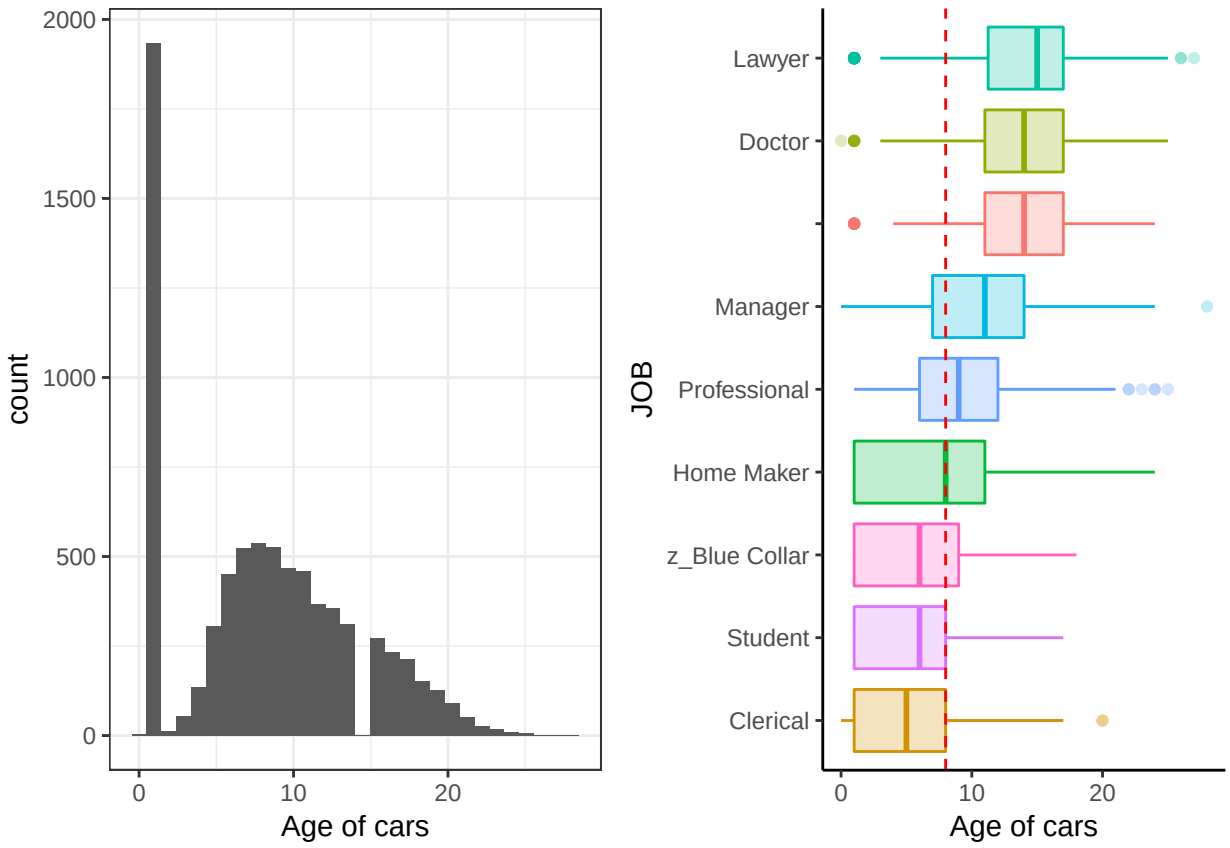


Figure 2: car age

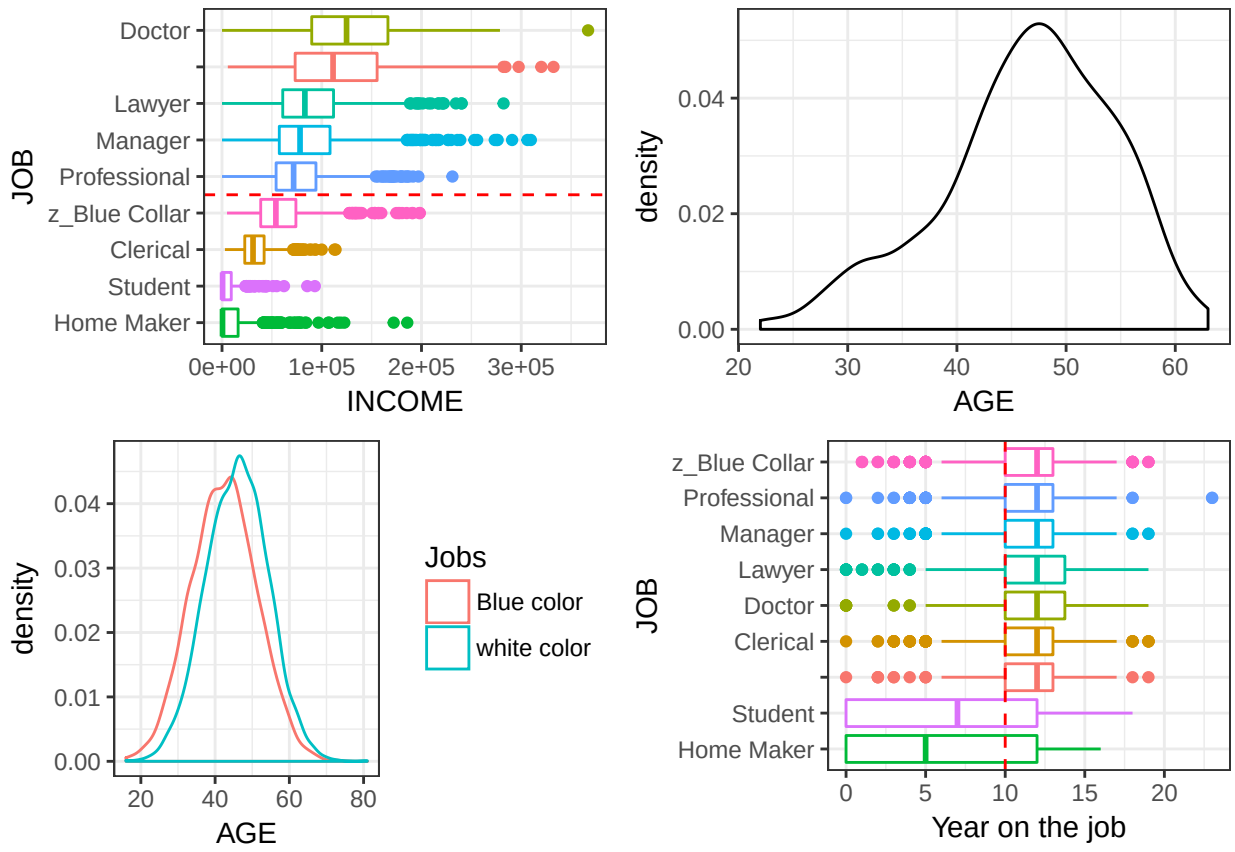


Figure 3: Missing JOB profile

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.

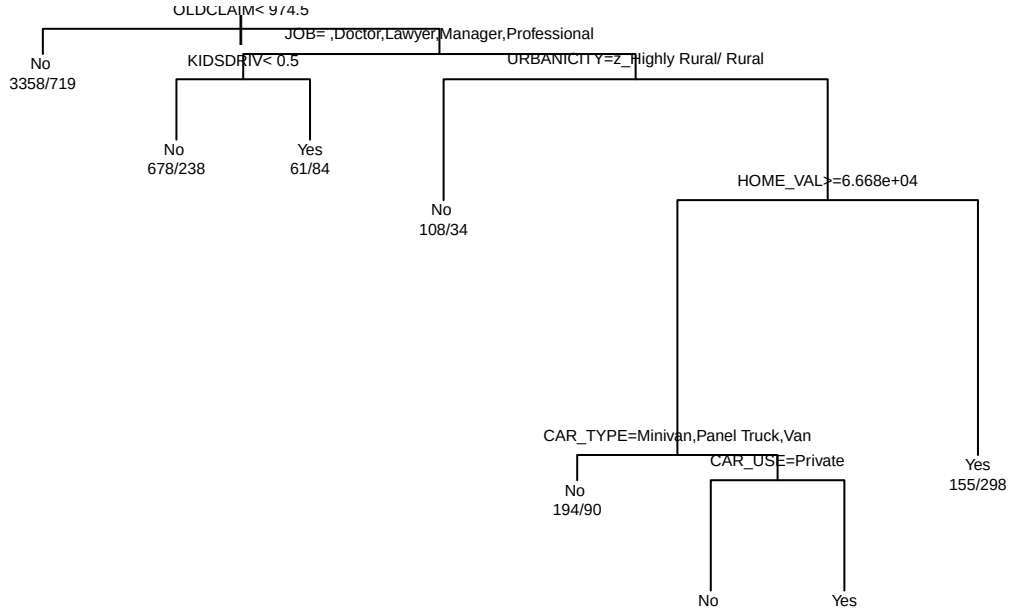


Figure 4: Decision Tree

```
##
## 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
```

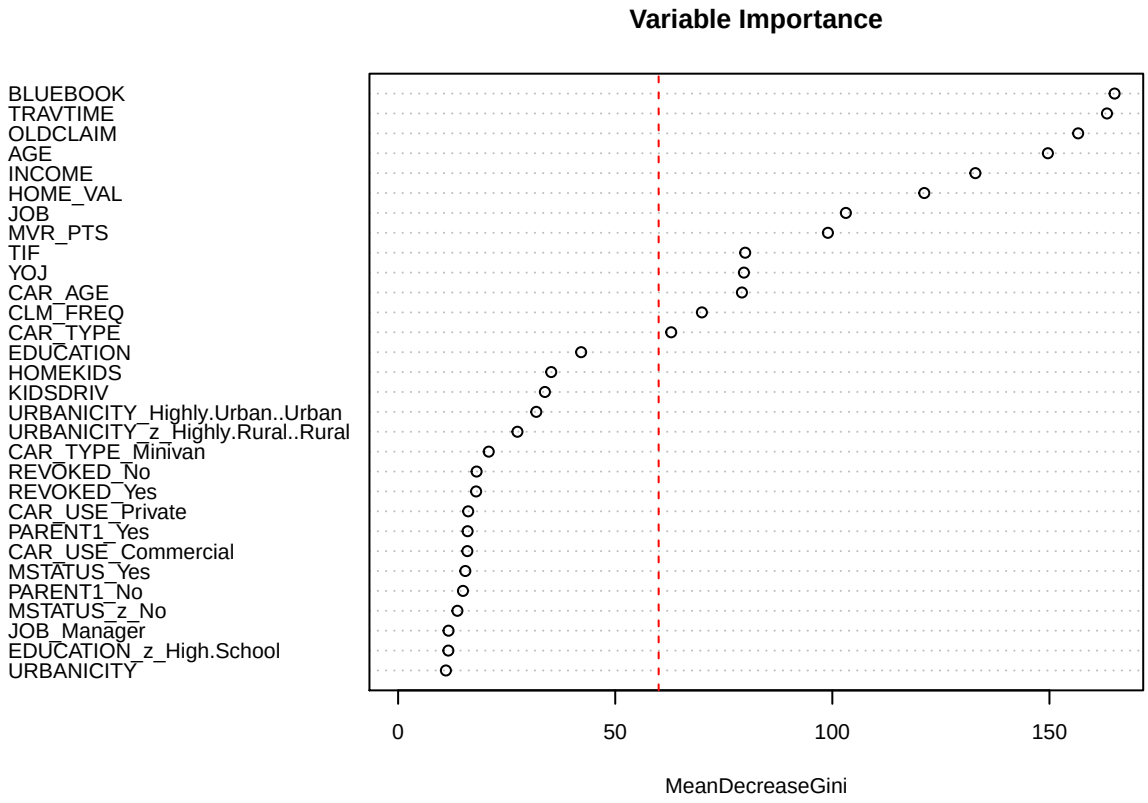


Figure 5: Variable Importance Plot and test prediction

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. The below plot shows the coefficients that are not zero in the decreasing order of absolute value of coefficients.

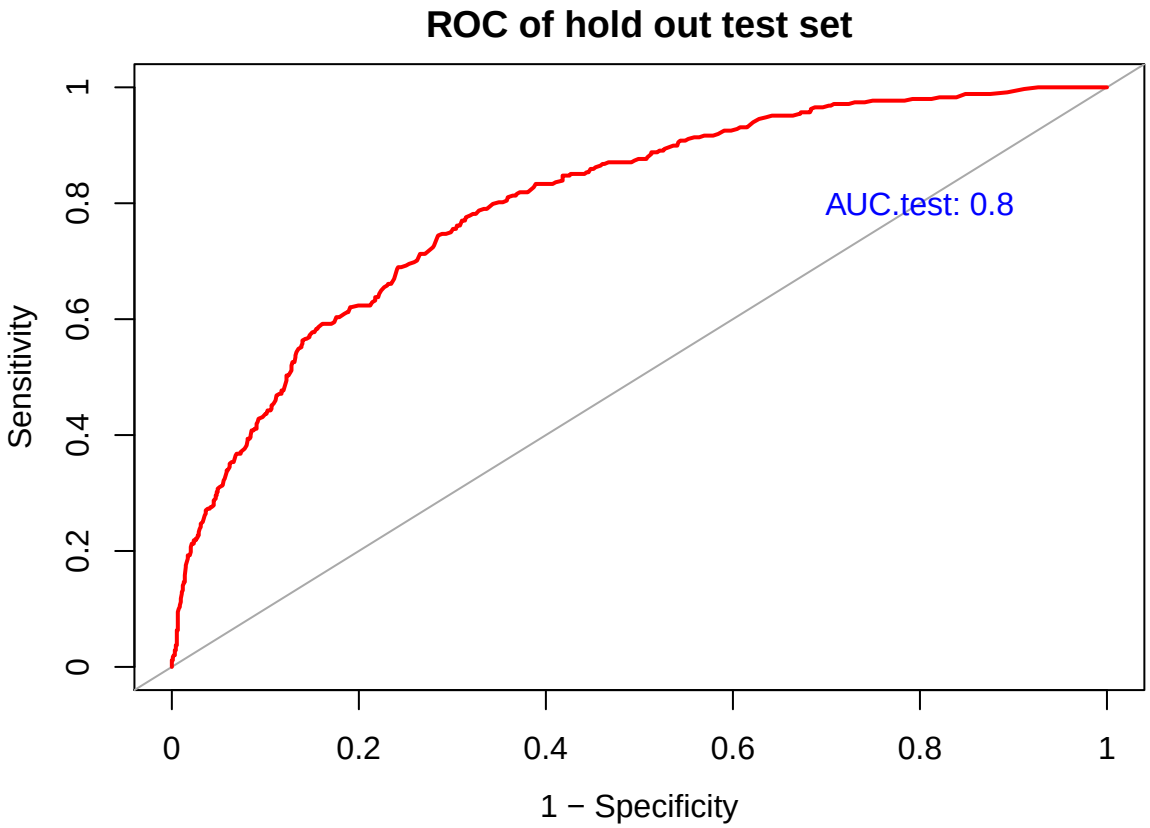


Figure 6: Prediction of bagged decision tree

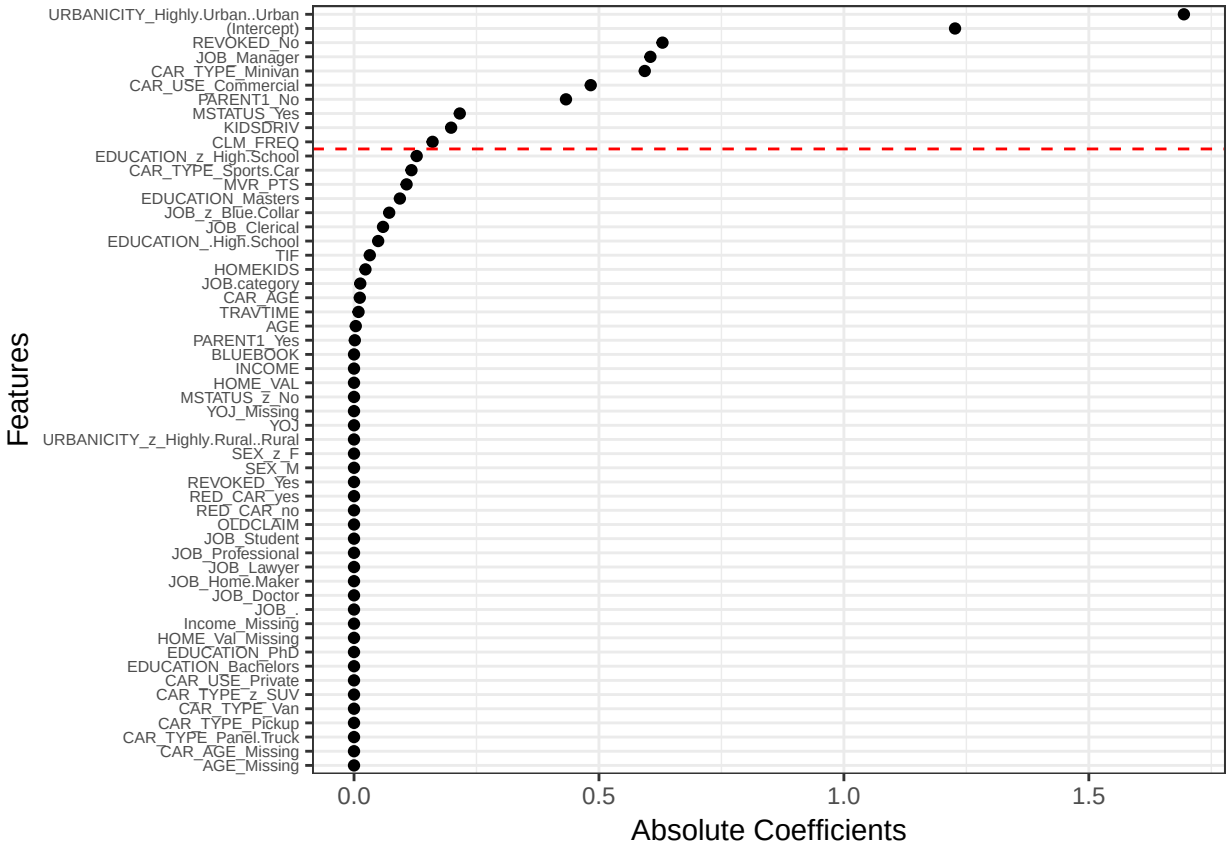


Figure 7: Feature selection - Lasso

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. The ROC curves for the training and test samples are shown below.

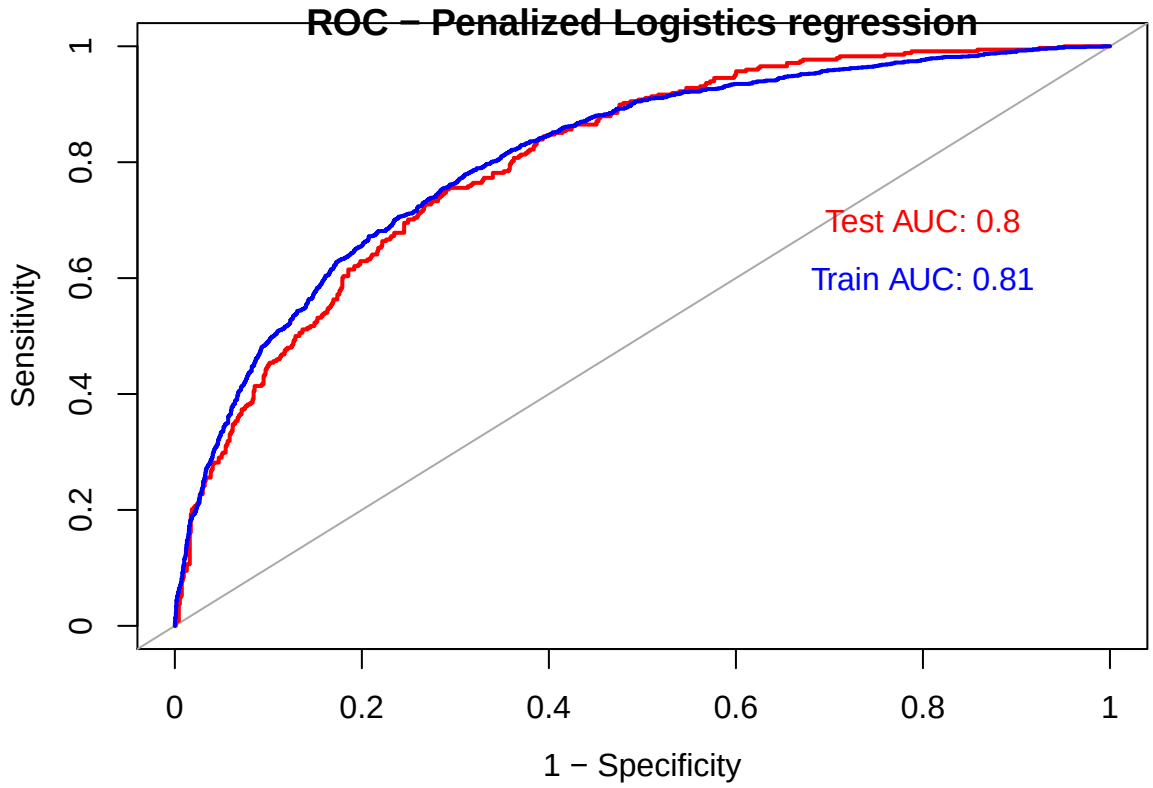
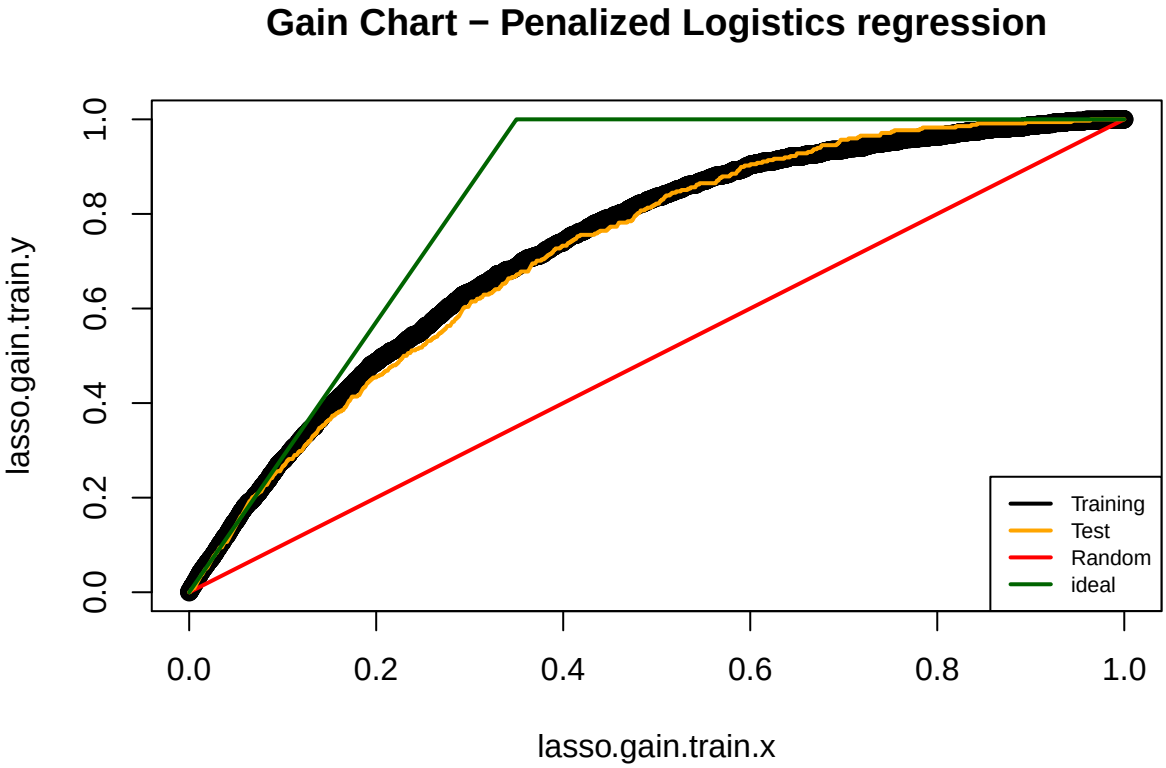


Figure 8: Lasso ROC curves for test and training data




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

Table 3: confusion matrix statistics - lasso logistic train

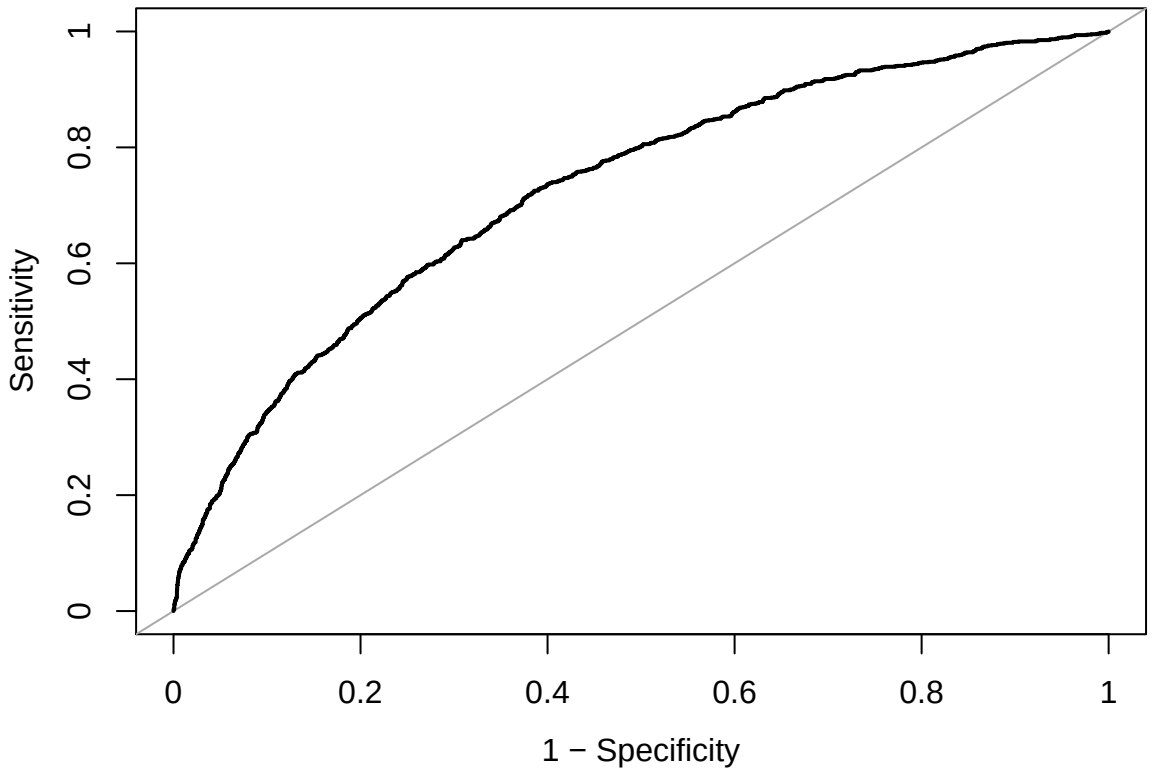
Accuracy	0.7882604
Kappa	0.3419255
AccuracyLower	0.7768572
AccuracyUpper	0.7993367
AccuracyNull	0.7375048
AccuracyPValue	0.0000000
McnemarPValue	0.0000000

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

Table 4: confusion matrix statistics - lasso logistic test

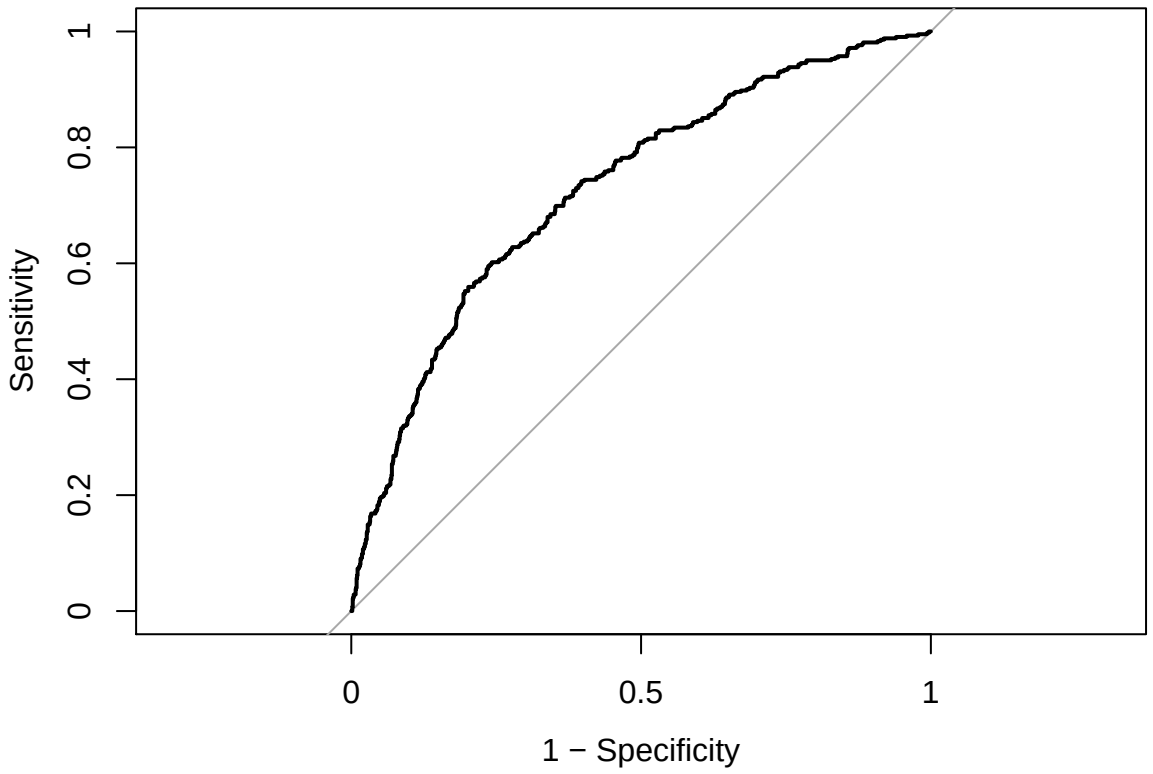
Accuracy	0.7729393
Kappa	0.3070814
AccuracyLower	0.7490442
AccuracyUpper	0.7955800
AccuracyNull	0.7293935
AccuracyPValue	0.0001983
McnemarPValue	0.0000000

```
##
## Call:
## glm(formula = y ~ ., family = binomial, data = dffit[complete.cases(dffit),
##     ])
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9016  -0.7759  -0.5674   0.8798   2.8372
##
## Coefficients: (1 not defined because of singularities)
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)         9.437e-02  2.465e-01   0.383 0.701826
## BLUEBOOK           -1.393e-05  4.857e-06  -2.868 0.004127 **
## TRAVTIME            6.743e-03  2.174e-03   3.102 0.001921 **
## OLDCLAIM            3.820e-06  4.129e-06   0.925 0.354935
## AGE                -1.070e-02  4.228e-03  -2.531 0.011380 *
## INCOME             -1.012e-06  1.278e-06  -0.792 0.428305
## HOME_VAL           -2.509e-06  3.476e-07  -7.217 5.31e-13 ***
## MVR_PTS             1.592e-01  1.659e-02   9.593  < 2e-16 ***
## TIF                -4.059e-02  8.863e-03  -4.579 4.66e-06 ***
## YOJ                -1.068e-02  9.786e-03  -1.091 0.275124
## CAR_AGE            -4.690e-03  7.811e-03  -0.600 0.548179
## CLM_FREQ            2.574e-01  3.328e-02   7.734 1.04e-14 ***
## JOB_.              -1.874e-01  1.786e-01  -1.049 0.294209
## JOB_Clerical       -4.283e-01  1.109e-01  -3.863 0.000112 ***
## JOB_Doctor         -8.293e-01  2.688e-01  -3.085 0.002033 **
## JOB_Home.Maker     -5.877e-01  1.598e-01  -3.679 0.000234 ***
## JOB_Lawyer         -6.660e-01  1.535e-01  -4.338 1.44e-05 ***
## JOB_Manager        -1.077e+00  1.484e-01  -7.254 4.04e-13 ***
## JOB_Professional   -5.625e-01  1.235e-01  -4.557 5.20e-06 ***
## JOB_Student        -5.777e-01  1.518e-01  -3.806 0.000141 ***
## JOB_z_Blue.Collar          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: 5585.9  on 4826  degrees of freedom
## Residual deviance: 4963.0  on 4807  degrees of freedom
## AIC: 5003
##
## Number of Fisher Scoring iterations: 4
```




```
## Area under the curve: 0.7257
```

```
##                test.y
## predictedClass    0    1
##              0 1134  339
##              1   65   83
```




```
## Area under the curve: 0.7282
```