Auto Insurance prediction

Sri Seshadri

10/20/2017

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

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

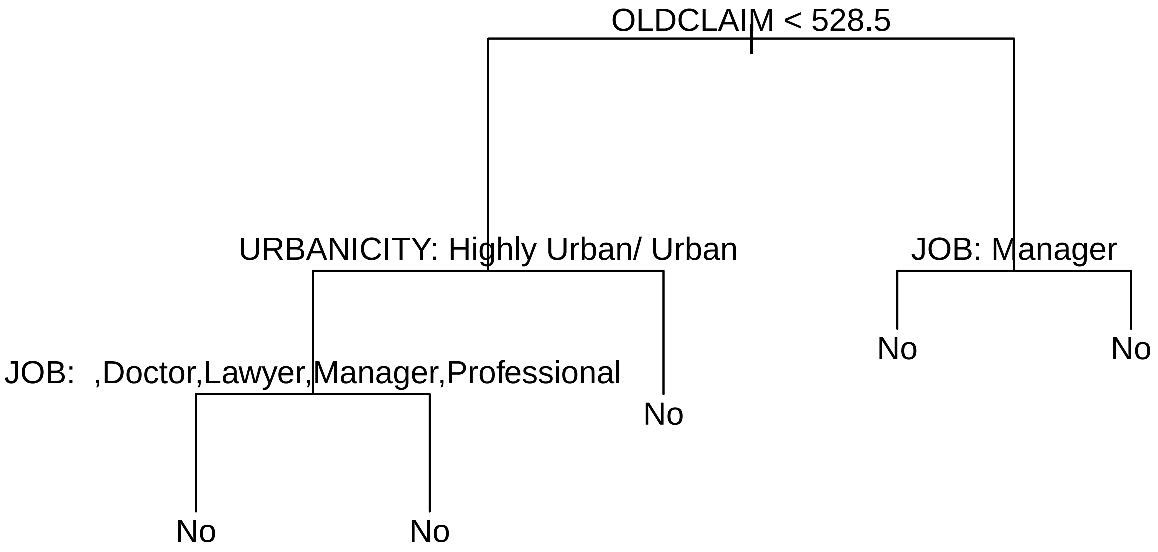
	min	Q1	median	Q3	max	mean	sd	n	missing
HOMEKIDS	0	0.00	0.00	1.00	5.00	0.72	1.12	8161	0
YOJ	0	9.00	11.00	13.00	23.00	10.50	4.09	7707	454
INCOME	0	28096.97	54028.17	85986.21	367030.26	61898.10	47572.69	7716	445
HOME_VAL	0	0.00	161159.53	238724.45	885282.34	154867.29	129123.78	7697	464
TRAVTIME	5	22.45	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

3. Feature Selection

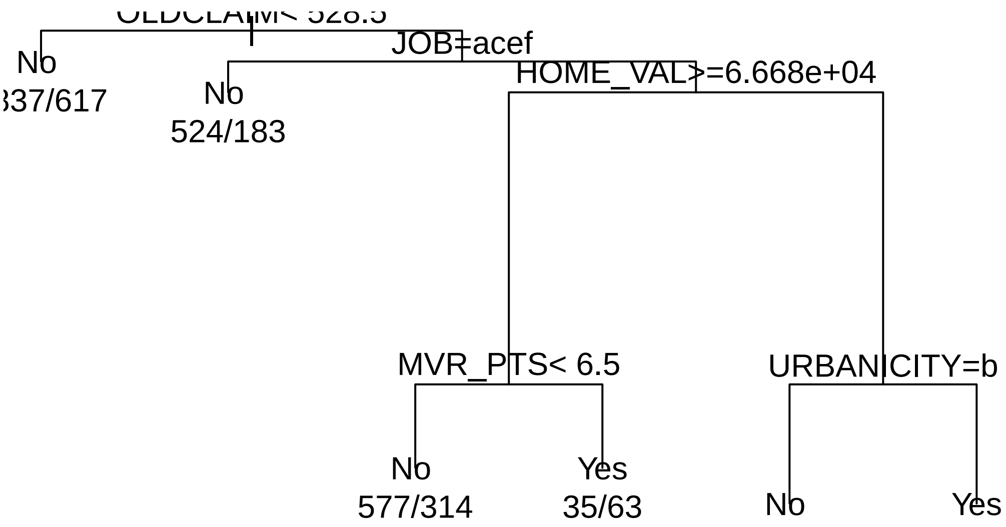
3.1 Training and Test data partition

3.2 Decision Tree

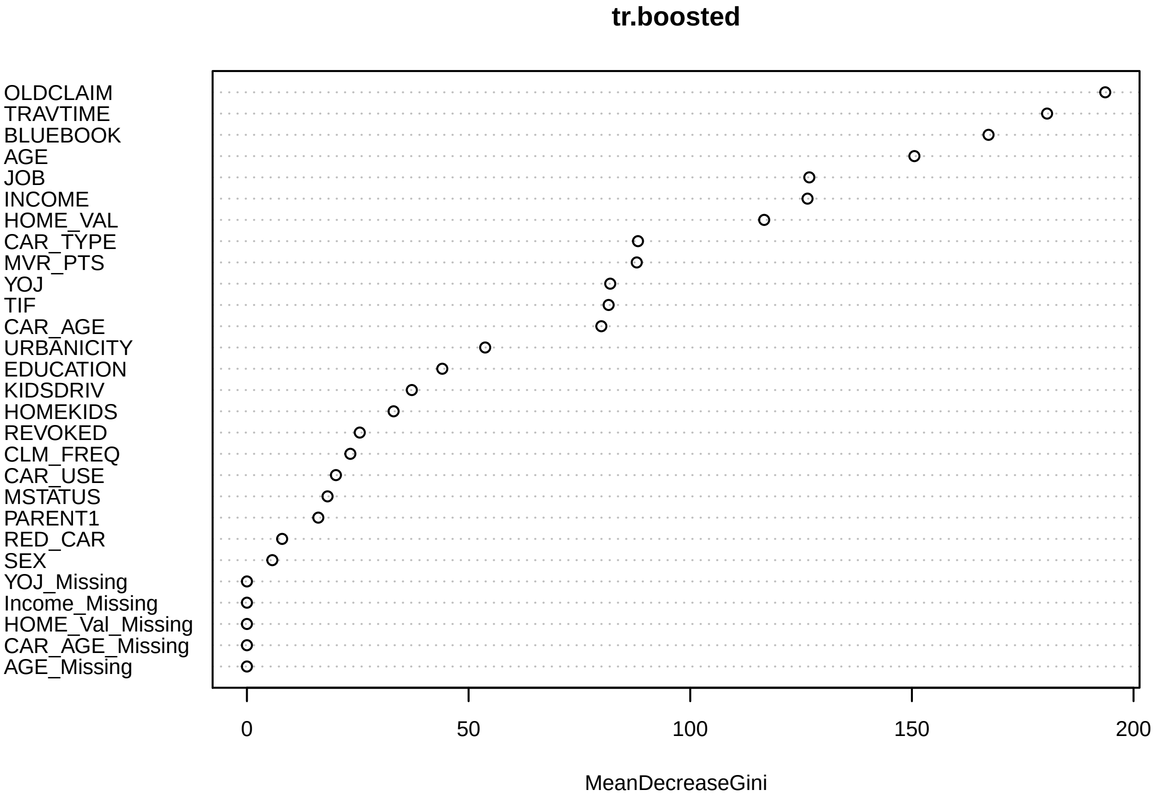
```
##
## Classification tree:
## tree::tree(formula = TARGET_FLAG2 ~ . - TARGET_FLAG, data = training)
## Variables actually used in tree construction:
## [1] "OLDCLAIM"   "URBANICITY" "JOB"
## Number of terminal nodes:  5
## Residual mean deviance:  1.016 = 4610 / 4539
## Misclassification error rate: 0.2645 = 1202 / 4544
```

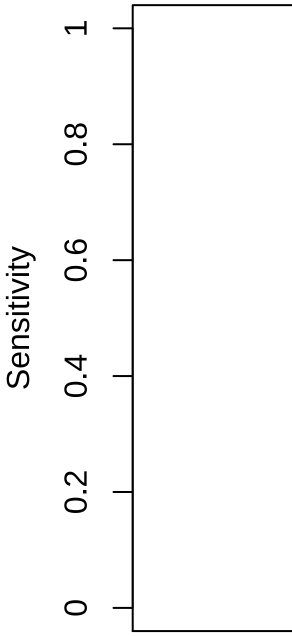

OLDCLAIM < 528.5
URBANICITY: Highly Urban/ Urban
JOB: Manager
JOB: ,Doctor,Lawyer,Manager,Professional
No
No
No
No
No

```
## Call:
## rpart::rpart(formula = TARGET_FLAG2 ~ . - TARGET_FLAG, data = training,
##     parms = list(split = "gini"))
##   n= 5714
##
##           CP nsplit rel error    xerror       xstd
## 1 0.02166225      0 1.0000000 1.0000000 0.02209347
## 2 0.01856764      4 0.9084881 0.9608753 0.02180832
## 3 0.01000000      5 0.8899204 0.9250663 0.02153313
##
## Variable importance
##   OLDCLAIM   CLM_FREQ    MVR_PTS        JOB   HOME_VAL  EDUCATION
##         27         27         12         10          6          5
## URBANICITY    MSTATUS    CAR_AGE     INCOME    PARENT1        AGE
##          5          2          2          1          1          1
##
## Node number 1: 5714 observations,    complexity param=0.02166225
##   predicted class=No   expected loss=0.2639132  P(node) =1
##     class counts:  4206  1508
##    probabilities: 0.736 0.264
##   left son=2 (3454 obs) right son=3 (2260 obs)
##   Primary splits:
##       OLDCLAIM   < 528.5    to the left,  improve=127.02110, (0 missing)
##       CLM_FREQ   < 0.5      to the left,  improve=125.78970, (0 missing)
##       URBANICITY splits as  RL,           improve=110.54770, (0 missing)
##       MVR_PTS    < 2.5      to the left,  improve= 67.24890, (0 missing)
##       HOME_VAL   < 213918.3 to the right, improve= 56.73779, (323 missing)
##   Surrogate splits:
##       CLM_FREQ < 0.5      to the left,  agree=0.999, adj=0.998, (0 split)
##       MVR_PTS  < 2.5      to the left,  agree=0.736, adj=0.332, (0 split)
##       AGE      < 29.5     to the right, agree=0.607, adj=0.007, (0 split)
##       KIDSDRIV < 3.5      to the left,  agree=0.605, adj=0.001, (0 split)
##       HOMEKIDS < 4.5      to the left,  agree=0.605, adj=0.001, (0 split)
##
## Node number 2: 3454 observations
##   predicted class=No   expected loss=0.1786335  P(node) =0.6044802
##     class counts:  2837   617
##    probabilities: 0.821 0.179
##
## Node number 3: 2260 observations,    complexity param=0.02166225
##   predicted class=No   expected loss=0.3942478  P(node) =0.3955198
##     class counts:  1369   891
##    probabilities: 0.606 0.394
##   left son=6 (707 obs) right son=7 (1553 obs)
##   Primary splits:
##       JOB      splits as  LRLRLLRRR,    improve=37.72876, (0 missing)
##       HOME_VAL < 66676.81 to the right, improve=29.22019, (150 missing)
##       REVOKED  splits as  LR,           improve=23.42995, (0 missing)
##       MVR_PTS  < 6.5      to the left,  improve=23.02408, (0 missing)
##       CAR_USE  splits as  RL,           improve=21.67330, (0 missing)
##   Surrogate splits:
##       EDUCATION splits as  RRLLR,        agree=0.887, adj=0.638, (0 split)
##       CAR_AGE   < 12.5     to the right, agree=0.747, adj=0.191, (0 split)
##       INCOME    < 82955.1  to the right, agree=0.720, adj=0.106, (0 split)
##       CAR_TYPE  splits as  RLRRRR,       agree=0.694, adj=0.021, (0 split)
##       BLUEBOOK  < 27375    to the right, agree=0.692, adj=0.016, (0 split)
##
## Node number 6: 707 observations
##   predicted class=No   expected loss=0.2588402  P(node) =0.1237312
##     class counts:   524   183
##    probabilities: 0.741 0.259
##
## Node number 7: 1553 observations,    complexity param=0.02166225
##   predicted class=No   expected loss=0.4558918  P(node) =0.2717886
##     class counts:   845   708
##    probabilities: 0.544 0.456
##   left son=14 (989 obs) right son=15 (564 obs)
##   Primary splits:
##       HOME_VAL   < 66676.81 to the right, improve=27.34753, (105 missing)
##       CAR_TYPE   splits as  LRRRRR,       improve=19.28322, (0 missing)
##       MSTATUS    splits as  LR,           improve=18.92105, (0 missing)
##       URBANICITY splits as  RL,           improve=18.33362, (0 missing)
##       CAR_USE    splits as  RL,           improve=17.73996, (0 missing)
##   Surrogate splits:
##       MSTATUS splits as  LR,           agree=0.787, adj=0.416, (105 split)
##       JOB     splits as  -L-L--LRL,    agree=0.734, adj=0.269, (0 split)
##       PARENT1 splits as  LR,           agree=0.712, adj=0.209, (0 split)
##       INCOME  < 12977.77 to the right, agree=0.659, adj=0.063, (0 split)
##       AGE     < 27.5     to the right, agree=0.648, adj=0.034, (0 split)
##
## Node number 14: 989 observations,    complexity param=0.01856764
##   predicted class=No   expected loss=0.3811931  P(node) =0.1730837
##     class counts:   612   377
##    probabilities: 0.619 0.381
##   left son=28 (891 obs) right son=29 (98 obs)
##   Primary splits:
##       MVR_PTS  < 6.5      to the left,  improve=14.895760, (0 missing)
##       CAR_TYPE splits as  LRRRRR,       improve=14.454750, (0 missing)
##       CAR_USE  splits as  RL,           improve=11.995880, (0 missing)
##       BLUEBOOK < 13000    to the right, improve= 8.519499, (0 missing)
##       INCOME   < 73848.25 to the right, improve= 7.107499, (54 missing)
##
## Node number 15: 564 observations,    complexity param=0.02166225
##   predicted class=Yes  expected loss=0.4131206  P(node) =0.09870494
##     class counts:   233   331
##    probabilities: 0.413 0.587
##   left son=30 (70 obs) right son=31 (494 obs)
##   Primary splits:
##       URBANICITY splits as  RL,           improve=22.189690, (0 missing)
##       REVOKED    splits as  LR,           improve= 9.439760, (0 missing)
##       CAR_TYPE   splits as  LRRRLL,       improve= 7.664257, (0 missing)
##       CAR_AGE    < 13.5     to the right, improve= 5.480076, (37 missing)
##       JOB        splits as  -R-R--LLR,    improve= 4.646520, (0 missing)
##   Surrogate splits:
##       AGE      < 21.5     to the left,  agree=0.879, adj=0.029, (0 split)
##       TRAVTIME < 69.67052 to the right, agree=0.878, adj=0.014, (0 split)
##       OLDCLAIM < 606.5    to the left,  agree=0.878, adj=0.014, (0 split)
##
## Node number 28: 891 observations
##   predicted class=No   expected loss=0.352413  P(node) =0.1559328
##     class counts:   577   314
##    probabilities: 0.648 0.352
##
## Node number 29: 98 observations
##   predicted class=Yes  expected loss=0.3571429  P(node) =0.01715086
##     class counts:    35    63
##    probabilities: 0.357 0.643
##
## Node number 30: 70 observations
##   predicted class=No   expected loss=0.2142857  P(node) =0.01225061
##     class counts:    55    15
##    probabilities: 0.786 0.214
##
## Node number 31: 494 observations
##   predicted class=Yes  expected loss=0.3603239  P(node) =0.08645432
##     class counts:   178   316
##    probabilities: 0.360 0.640
```



```
##
## Call:
##  randomForest(formula = TARGET_FLAG2 ~ . - TARGET_FLAG, data = training,      mtry = 28, na.action =
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 28
##
##         OOB estimate of  error rate: 22.54%
## Confusion matrix:
##       No Yes class.error
## No  3036 306  0.09156194
## Yes  718 484  0.59733777
```


tr.boosted
OLDCLAIM
TRAVTIME
BLUEBOOK
AGE
JOB
INCOME
HOME_VAL
CAR_TYPE
MVR_PTS
YOJ
TIF
CAR_AGE
URBANICITY
EDUCATION
KIDSDRIV
HOMEKIDS
REVOKED
CLM_FREQ
CAR_USE
MSTATUS
PARENT1
RED_CAR
SEX
YOJ_Missing
Income_Missing
HOME_Val_Missing
CAR_AGE_Missing
AGE_Missing
0
50
100
150
200
MeanDecreaseGini


Sensitivity
1
0.8
0.6
0.4
0.2
0