Significant Predictors – of target customers

Predict who will be interested in buying caravan insurance policy using different regression methods.

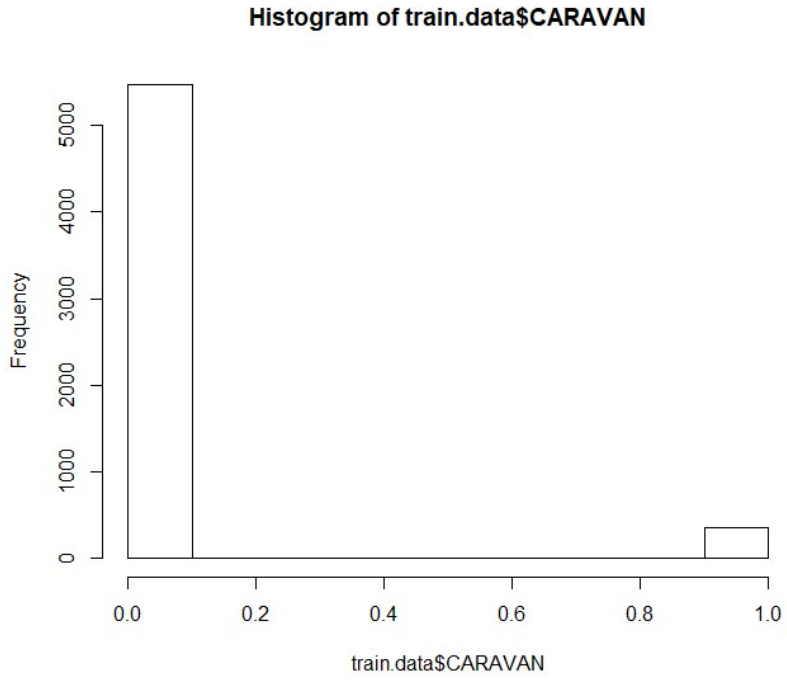
Dataset: Insurance company benchmark dataset
Training Dataset : ticdata2000.txt - 5822 rows
Test Dataset : ticeval2000.txt - 4000 rows
Features : 86 features

Dataset contains type of customer, customer income and other policy the customer holds and demographic details.

Exploratory data analysis:

Caravan is not common so it's expected to have lower number of policies sold.
Number of caravan policies purchased in the dataset - **348**

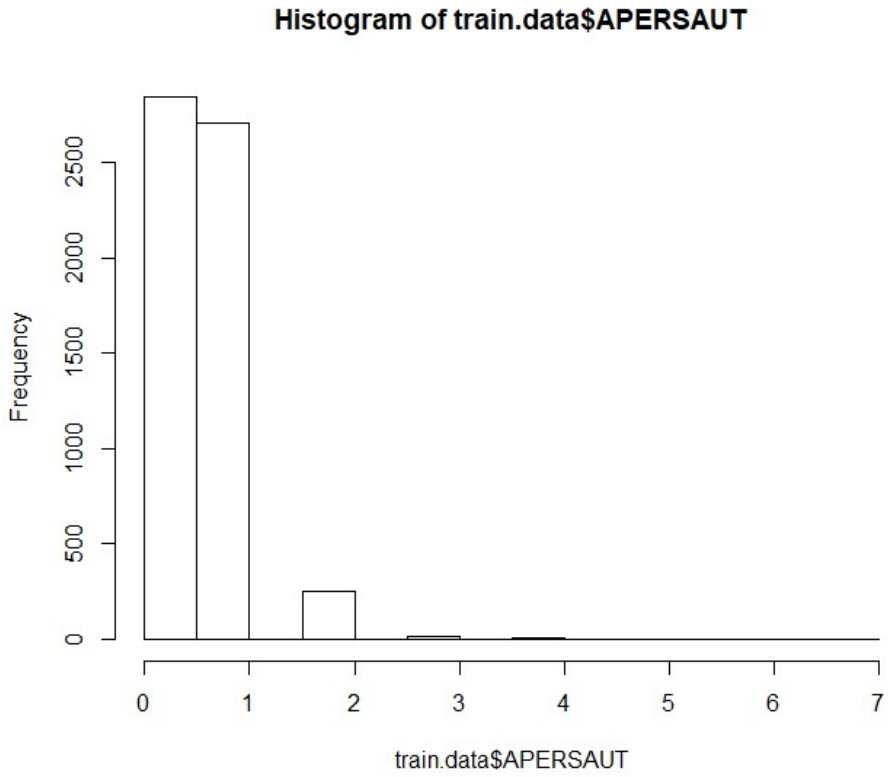
```
> carvanpol<-table(train.data$CARAVAN)
> carvanpol

   0    1
5474  348
```



Comparing the number of car policies which is so commonly owned vehicle category among customers.

```
> carpol<-table(train.data$APERSAUT)
> carpol
```

```
   0    1    2    3    4    6    7
2845 2712  246   12    5    1    1
```

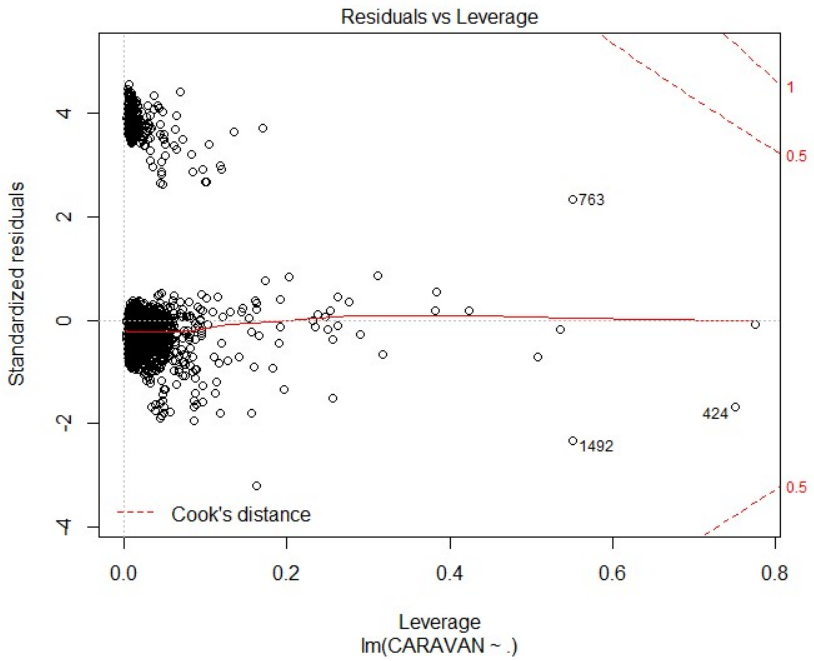


Step1: Linear Regression

Using linear regression to identify most significant predictors for the response variable caravan policy. Below mentioned are the significant predictors for caravan policy as residuals square is minimum.

```
Adjusted R-squared:  0.05916
```

```
Call:
lm(formula = CARAVAN ~ ., data = train.data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.67293 -0.08720 -0.04593 -0.00639  1.04628

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.7685381  0.4298406   1.788 0.073835 .
MGEMLEEF     0.0107473  0.0049596   2.167 0.030279 *
PPERSAUT     0.0102787  0.0026346   3.901 9.67e-05 ***
PLEVEN      -0.0155397  0.0064753  -2.400 0.016433 *
PGEZONG      0.1937254  0.0793370   2.442 0.014644 *
PWAOREG      0.0647933  0.0256913   2.522 0.011696 *
PBRAND       0.0132643  0.0035906   3.694 0.000223 ***
ALEVEN       0.0372344  0.0154024   2.417 0.015661 *
AGEZONG     -0.4050642  0.1898715  -2.133 0.032938 *
APLEZIER     0.3633887  0.0885318   4.105 4.11e-05 ***
```

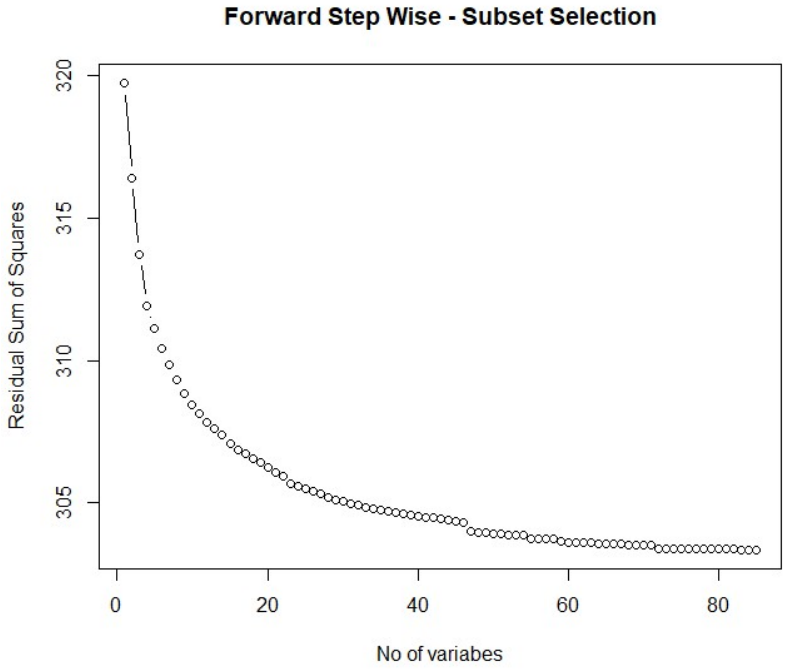


Step2: Subset Selection – Forward

It starts with zero predictors and keep adding one predictor which has lower RSS at each stage.

Here let's find forward models with all 86 predictors which influence CARAVAN policies and then choose the best one based on the residual sum of errors.

```
> fwd.models<-regsubsets(CARAVAN~.,train.data,nvmax=86,method="forward"
)
> fmod<-summary(fwd.models)$which
>
> dim(fmod)
[1] 15 86
```



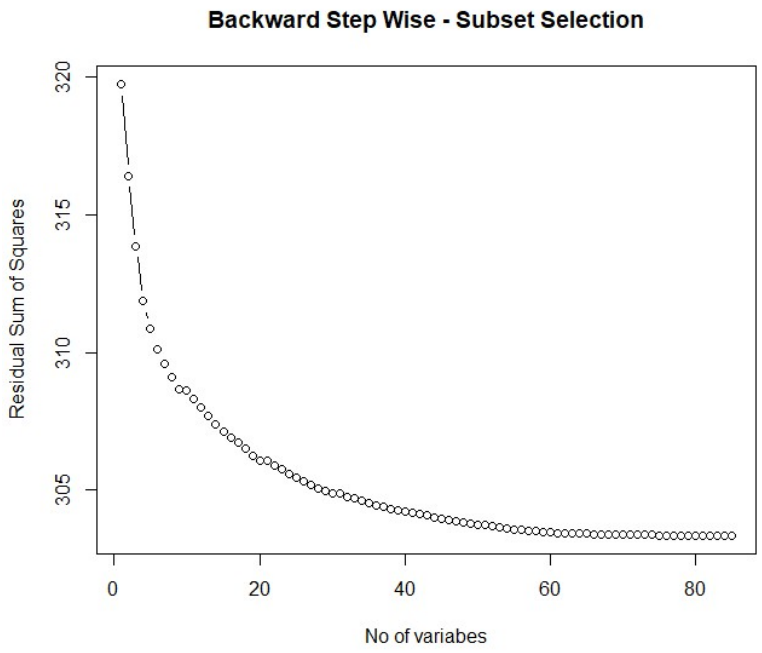
Error slowly reduces by including all features. Let's see what are first 10 predictors to chosen by the forward subset selection method and its weight.

```
> coef(fwd.models,10)
  (Intercept)       MRELGE     MOPLLAAG     MBERBOER     MKOOPKLA      PWAPART     PPERSAUT       PBRAND     APLEZIER      AFIETS
-0.019845474  0.005933600 -0.005914735 -0.008419351  0.004227842  0.011079245  0.010545098  0.006845541  0.285530029  0.039278555
    ABYSTAND
 0.078294696
```

Step3: Backward subset selection method

It starts with full model and removes each predictor at each stage which does not have any influence on the response variable.

```
> back.models<-regsubsets(CARAVAN~.,train.data,nvmax=86,method="backwar
d")
>
> bmod<-summary(back.models)$which
> dim(bmod)
[1] 85 86
```



This model states that 60 predictor model is optimal and its where the RSS gets constant.

Let's verify just top 10 predictors which influence the caravan policy buying.

```
> coef(back.models,10)
  (Intercept)       MRELGE     MOPLLAAG     MBERBOER      PWALAND     PPERSAUT       PBRAND      ALEVEN     APLEZIER      AFIETS
-9.266593e-05  6.781188e-03 -7.379527e-03 -8.546117e-03 -1.960119e-02  1.108606e-02  1.089528e-02  7.334644e-03  2.849877e-01  4.044235e-02
    ABYSTAND
 8.003541e-02
>
```

Step 4: LASSO Model

This model reduces the beta coefficients of the predictors and it shrinks to zero if the predictor is insignificant wrt to the response variable and we have hyper parameter to learn lambda which is done using cross validation method to find minimum lambda with which regression model is regularized. These models work on the Gaussian data so we have to scale the data.

```
> lasso.fit<-cv.glmnet(s.train_data[,-86],s.train_data[,"CARAVAN"],nfol
ds=10,alpha=1)
> head(lasso.fit$lambda)
[1] 0.15089675 0.13749150 0.12527714 0.11414786 0.10400728 0.09476756
```

```
> cv.lambda<-lasso.fit$lambda.min
> cv.lambda
[1] 0.01115236
```

<u>Coefficients of predictors determined by LASSO:</u>

30 predictors are retained by the LASSO to predict the CARAVAN policy.

```
> lasso.pred.coef
86 x 1 sparse Matrix of class "dgCMatrix"
                        1
(Intercept)  3.757981e-16
MGEMLEEF     9.344540e-03
MGODRK      -5.379169e-03
MGODPR       5.544119e-03
MGODGE      -6.161908e-03
MRELGE       1.971297e-02
MRELSA      -5.395473e-03
MOPLHOOG     2.665323e-02
MOPLLAAG    -2.696071e-02
MBERBOER    -2.576627e-02
MBERMIDD     1.174253e-02
MHHUUR      -1.717856e-02
MAUT1        1.253703e-02
MINK7512     6.786253e-03
MINK123M    -1.697601e-02
MINKGEM      1.868166e-02
MKOOPKLA     2.078806e-02
PWAPART      3.333610e-02
PWALAND     -1.777139e-02
PPERSAUT     1.174234e-01
PWERKT      -2.418797e-03
PGEZONG      1.190605e-02
PWAOREG      2.006287e-02
PBRAND       4.935285e-02
APERSAUT     6.349467e-03
ATRACTOR    -7.536562e-03
ALEVEN       3.107723e-04
AZEILPL      6.034406e-03
APLEZIER     8.720832e-02
AFIETS       2.242371e-02
ABYSTAND     2.918466e-02
```

Step5: Ridge Regression

Similar to LASSO but it does not eliminate the predictors it just shrinks the beta coefficient of the predictors so with beta coefficients we can analyze most significant predictors for the response variable CARAVAN.

```
> ridge.fit<-cv.glmnet(s.train_data[,-86],s.train_data["CARAVAN"],nfol
ds=10,alpha=0)
> cv.lambda<-ridge.fit$lambda.min
> cv.lambda
[1] 0.4716604
```

```
> ridge.pred.coef<-predict(ridge.fit,s=cv.lambda,type="coefficients")
> ridge.pred.coef
86 x 1 sparse Matrix of class "dgCMatrix"
                        1
(Intercept)  2.614920e-16
MOSTYPE     -1.491873e-03
MAANTHUI    -7.618006e-03
MGEMOMV     -2.441277e-03
MGEMLEEF     1.725879e-02
MOSHOOFD    -5.643873e-03
MGODRK      -1.045428e-02
MGODPR       9.485674e-03
MGODOV       7.418567e-03
MGODGE      -8.133299e-03
MRELGE       1.194378e-02
MRELSA      -1.044481e-02
MRELOV      -2.915132e-03
MFALLEEN    -3.748966e-03
MFGEKIND    -6.329806e-03
MFWEKIND     4.831440e-03
MOPLHOOG     2.588488e-02
MOPLMIDD     4.302374e-03
MOPLLAAG    -2.238729e-02
MBERHOOG     4.856414e-03
MBERZELF     1.061485e-03
MBERBOER    -2.273206e-02
MBERMIDD     1.688268e-02
MBERARBG    -4.851935e-03
MBERARBO     1.265014e-03
MSKA         4.373148e-03
MSKB1       -1.859388e-03
MSKB2       -2.925969e-03
MSKC         6.776325e-03
MSKD        -6.481744e-03
MHHUUR      -1.205594e-02
MHKOOP       1.050163e-02
MAUT1        1.276624e-02
MAUT2        1.550621e-03
MAUT0       -5.631485e-03
MZFONDS      3.790825e-03
MZPART      -5.600205e-03
MINKM30     -2.146253e-03
MINK3045     5.712806e-03
MINK4575     1.433546e-03
MINK7512     1.443022e-02
MINK123M    -1.833734e-02
MINKGEM      1.601423e-02
MKOOPKLA     1.718901e-02
PWAPART      2.396745e-02
PWABEDR     -3.115532e-03
PWALAND     -1.067052e-02
PPERSAUT     6.191382e-02
PBESAUT     -1.873294e-03
PMOTSCO     -4.404261e-03
PVRAAUT     -5.160558e-03
PAANHANG     6.807708e-03
PTRACTOR     1.326990e-04
PWERKT      -4.624816e-03
PBROM       -1.572114e-03
```

```
PLEVEN       -1.245489e-02
PPERSONG     -1.004213e-03
PGEZONG       1.441602e-02
PWAOREG       2.021310e-02
PBRAND        3.647613e-02
PZEILPL      -4.264305e-03
PPLEZIER      2.570689e-02
PFIETS        9.907600e-03
PINBOED      -8.780072e-03
PBYSTAND      1.315715e-02
AWAPART       1.527546e-02
AWABEDR       8.397858e-04
AWALAND      -9.795174e-03
APERSAUT      4.615619e-02
ABESAUT      -4.086401e-03
AMOTSCO       2.936516e-03
AVRAAUT      -3.255344e-03
AAANHANG      6.779448e-04
ATRACTOR     -8.912427e-03
AWERKT       -3.505759e-03
ABROM        -3.853250e-03
ALEVEN        1.545713e-02
APERSONG     -3.848880e-03
AGEZONG       2.042716e-03
AWAOREG       2.944940e-03
ABRAND        2.670622e-03
AZEILPL       1.289000e-02
APLEZIER      5.066845e-02
AFIETS        1.505660e-02
AINBOED       7.943098e-03
ABYSTAND      1.899729e-02
```

Conclusion:

By analyzing the beta coefficients of different regression methods, the top 10 predictors for the response variable CARAVAN is as below. The most common predictors in all the models are PPERSAUT, APLEZIER, PBRAND.

Linear Regression	Forward Method	Backward Method	LASSO	Ridge
MGEMLEEF	MRELGE	MRELGE	PPERSAUT	PPERSAUT
PPERSAUT	MOPLLAAG	MOPLLAAG	APLEZIER	APLEZIER
PLEVEN	MBERBOER	MBERBOER	PBRAND	APERSAUT
PGEZONG	MKOOPKLA	PWALAND	PWAPART	PBRAND
PWAOREG	PWAPART	PPERSAUT	ABYSTAND	MOPLHOOG
PBRAND	PPERSAUT	PBRAND	MOPLHOOG	PPLEZIER
ALEVEN	PBRAND	ALEVEN	AFIETS	PWAPART
AGEZONG	APLEZIER	APLEZIER	MKOOPKLA	PWAOREG
APLEZIER	AFIETS	AFIETS	PWAOREG	ABYSTAND
	ABYSTAND	ABYSTAND	MRELGE	MGEMLEEF

PPERSAUT- Contribution car policies
APLEZIER- Number of boat policies
PBRAND- Contribution fire policies

Customers with cars and boats along with fire policy are tend to own the caravan as well so they are the target customers for the CARAVAN policy.