Sales Forecasting For Insurance Company

2023-09-15

PROJECT OVERVIEW

The insurance company benchmark data set gives information on customers. Specifically, it contains 86 variables on product-usage data and socio-demographic data derived from zip area codes. There are 5,822 customers in the training set and another 4,000 in the test set. The data were collected to answer the following questions: We will predict who will be interested in buying a caravan insurance policy and give an explanation why they did (The data can be found in the ISLR2 package >data(Caravan).)

#First we will develop a model using the linear model.

```
library(ISLR2)
```

```
## Warning: package 'ISLR2' was built under R version 4.3.2
```

```
data(Caravan)
which(is.na(Caravan) == TRUE)
```

```
## integer(0)
```

```
Caravan$Purch <- as.numeric(Caravan$Purchase == "Yes")
set.seed(123)
indis <- sample(1:nrow(Caravan), round(40/100*nrow(Caravan)), replace = FALSE)
caravan_train <- Caravan[indis, ]
caravan_test <- Caravan[-indis, ]
lm.fit <- lm(Purch~., data = caravan_train)
lm_pred <- predict(lm.fit, caravan_test )
```

```
## Warning in predict.lm(lm.fit, caravan_test): prediction from rank-deficient
## fit; attr(*, "non-estim") has doubtful cases
```

```
summary(lm.fit)
```

```
##
## Call:
## lm(formula = Purch ~ ., data = caravan_train)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -1.43e-14 -3.82e-17 -3.40e-18  3.12e-17  2.86e-15
##
## Coefficients: (1 not defined because of singularities)
##               Estimate Std. Error    t value Pr(>|t|)
## (Intercept) -1.471e-15  1.020e-15 -1.443e+00  0.14927
## MOSTYPE     -6.740e-18  5.416e-18 -1.244e+00  0.21350
## MAANTHUI     2.188e-17  1.911e-17  1.145e+00  0.25236
## MGEMOMV      1.276e-17  1.692e-17  7.540e-01  0.45090
## MGEMLEEF    -2.541e-17  1.141e-17 -2.227e+00  0.02605 *
## MOSHOOFD     3.002e-17  2.430e-17  1.235e+00  0.21680
## MGODRK       5.714e-18  1.309e-17  4.360e-01  0.66261
## MGODPR      -3.748e-18  1.398e-17 -2.680e-01  0.78872
## MGODOV      -4.636e-18  1.263e-17 -3.670e-01  0.71367
## MGODGE       3.010e-18  1.347e-17  2.230e-01  0.82320
## MRELGE      -2.235e-17  1.815e-17 -1.232e+00  0.21825
## MRELSA      -2.003e-17  1.712e-17 -1.170e+00  0.24209
## MRELOV      -2.288e-17  1.830e-17 -1.251e+00  0.21119
## MFALLEEN     2.389e-17  1.558e-17  1.533e+00  0.12539
## MFGEKIND     2.294e-17  1.603e-17  1.431e+00  0.15262
## MFWEKIND     1.911e-17  1.670e-17  1.144e+00  0.25258
## MOPLHOOG     5.737e-18  1.622e-17  3.540e-01  0.72357
## MOPLMIDD     8.711e-18  1.696e-17  5.140e-01  0.60758
## MOPLLAAG     1.197e-17  1.740e-17  6.880e-01  0.49129
## MBERHOOG     5.431e-18  1.067e-17  5.090e-01  0.61092
## MBERZELF     1.060e-17  1.235e-17  8.590e-01  0.39064
## MBERBOER     9.485e-18  1.202e-17  7.890e-01  0.43022
## MBERMIDD    -6.922e-18  1.076e-17 -6.430e-01  0.52025
## MBERARBG     6.816e-19  1.043e-17  6.500e-02  0.94791
## MBERARBO    -3.128e-18  1.062e-17 -2.950e-01  0.76837
## MSKA        -2.622e-18  1.215e-17 -2.160e-01  0.82917
## MSKB1        9.648e-18  1.203e-17  8.020e-01  0.42265
## MSKB2        7.137e-18  1.080e-17  6.610e-01  0.50870
## MSKC         5.469e-18  1.175e-17  4.650e-01  0.64180
## MSKD         1.190e-17  1.108e-17  1.074e+00  0.28300
## MHHUUR       4.150e-17  9.575e-17  4.330e-01  0.66480
## MHKOOP       3.609e-17  9.566e-17  3.770e-01  0.70598
## MAUT1       -7.823e-18  1.759e-17 -4.450e-01  0.65660
## MAUT2       -8.115e-18  1.605e-17 -5.050e-01  0.61326
## MAUT0       -4.379e-18  1.686e-17 -2.600e-01  0.79515
## MZFONDS      1.282e-16  1.114e-16  1.151e+00  0.24993
## MZPART       1.301e-16  1.113e-16  1.169e+00  0.24236
## MINKM30     -6.950e-18  1.212e-17 -5.730e-01  0.56658
## MINK3045    -1.202e-17  1.158e-17 -1.038e+00  0.29937
## MINK4575    -1.122e-17  1.170e-17 -9.590e-01  0.33742
## MINK7512    -1.648e-17  1.214e-17 -1.357e+00  0.17476
## MINK123M     1.009e-17  1.592e-17  6.340e-01  0.52627
## MINKGEM      3.760e-18  1.045e-17  3.600e-01  0.71892
## MKOOPKLA    -8.139e-18  5.338e-18 -1.525e+00  0.12749
## PWAPART     -4.012e-17  3.843e-17 -1.044e+00  0.29662
## PWABEDR      4.510e-17  6.316e-17  7.140e-01  0.47526
## PWALAND      1.135e-17  1.159e-16  9.800e-02  0.92198
```

```
## PPERSAUT    -9.956e-18  6.477e-18 -1.537e+00  0.12443
## PBESAUT     -1.303e-17  5.546e-17 -2.350e-01  0.81424
## PMOTSCO      8.642e-17  3.608e-17  2.395e+00  0.01669 *
## PVRAAUT      1.121e-16  3.790e-16  2.960e-01  0.76747
## PAANHANG     4.212e-17  1.134e-16  3.710e-01  0.71036
## PTRACTOR     2.454e-17  3.192e-17  7.690e-01  0.44204
## PWERKT       1.470e-17  4.015e-16  3.700e-02  0.97079
## PBROM       -2.974e-18  4.041e-17 -7.400e-02  0.94133
## PLEVEN       5.267e-17  1.693e-17  3.111e+00  0.00189 **
## PPERSONG    -6.245e-18  7.079e-17 -8.800e-02  0.92972
## PGEZONG     -1.916e-16  1.631e-16 -1.174e+00  0.24032
## PWAOREG      3.169e-17  7.446e-17  4.260e-01  0.67044
## PBRAND      -1.844e-17  8.651e-18 -2.131e+00  0.03318 *
## PZEILPL      9.778e-17  3.419e-16  2.860e-01  0.77489
## PPLEZIER    -6.785e-17  7.740e-17 -8.770e-01  0.38082
## PFIETS       1.083e-16  1.234e-16  8.780e-01  0.38015
## PINBOED     -4.027e-17  9.167e-17 -4.390e-01  0.66048
## PBYSTAND    -1.168e-17  8.039e-17 -1.450e-01  0.88453
## AWAPART      6.529e-17  7.519e-17  8.680e-01  0.38530
## AWABEDR     -1.765e-16  1.904e-16 -9.270e-01  0.35397
## AWALAND     -4.921e-17  4.144e-16 -1.190e-01  0.90549
## APERSAUT    -5.829e-18  3.180e-17 -1.830e-01  0.85458
## ABESAUT      7.694e-17  2.790e-16  2.760e-01  0.78274
## AMOTSCO     -4.032e-16  1.614e-16 -2.497e+00  0.01259 *
## AVRAAUT     -3.170e-16  1.178e-15 -2.690e-01  0.78794
## AAANHANG    -8.786e-17  2.018e-16 -4.350e-01  0.66337
## ATRACTOR    -3.994e-18  7.249e-17 -5.500e-02  0.95607
## AWERKT      -3.480e-17  9.215e-16 -3.800e-02  0.96987
## ABROM        3.449e-17  1.225e-16  2.820e-01  0.77825
## ALEVEN      -1.175e-16  3.572e-17 -3.289e+00  0.00102 **
## APERSONG     4.888e-17  2.058e-16  2.370e-01  0.81233
## AGEZONG      3.161e-16  3.907e-16  8.090e-01  0.41859
## AWAOREG     -2.791e-17  3.623e-16 -7.700e-02  0.93861
## ABRAND       3.579e-17  2.754e-17  1.300e+00  0.19388
## AZEILPL             NA         NA         NA       NA
## APLEZIER    -5.893e-16  2.578e-16 -2.286e+00  0.02237 *
## AFIETS      -1.120e-16  8.556e-17 -1.309e+00  0.19071
## AINBOED      6.031e-17  1.847e-16  3.260e-01  0.74412
## ABYSTAND     3.479e-17  2.834e-16  1.230e-01  0.90229
## PurchaseYes  1.000e+00  3.038e-17  3.291e+16  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.391e-16 on 2243 degrees of freedom
## Multiple R-squared:      1,  Adjusted R-squared:      1
## F-statistic: 1.4e+31 on 85 and 2243 DF,  p-value: < 2.2e-16
```

#Therefore, the positive coefficients and variables with lower p value and number of '*'s against it have significance and show likelihood of interest of purchase of Caravan insurance policy.

#Now, we will develop a model using Forwards Selection, Backwards Selection, Lasso regression, and Ridge regression.

#Forward selection

```
library(leaps)
set.seed(123)
regfit.fwd <- regsubsets(Purch~., data = caravan_train, nbest = 1, nvmax = 85, method = "forward")
```

```
## Warning in leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax, force.in =
## force.in, : 1 linear dependencies found
```

```
## Reordering variables and trying again:
```

```
## Warning in rval$lopt[] <- rval$vorder[rval$lopt]: number of items to replace is
## not a multiple of replacement length
```

```
summary_fwd <- summary(regfit.fwd)

#identifying the optimal models
which(summary_fwd$cp == min(summary_fwd$cp))
```

```
## [1] 24
```

```
which(summary_fwd$bic == min(summary_fwd$bic))
```

```
## [1] 11
```

```
which(summary_fwd$rss == min(summary_fwd$rss))
```

```
## [1] 85
```

```
which(summary_fwd$adjr2 == max(summary_fwd$adjr2))
```

```
##  [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
## [26] 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
## [51] 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75
## [76] 76 77 78 79 80 81 82 83 84 85
```

```
#selecting 85 as the best subset forward. We chose 85 from the optimal model although we could have cho
sen fewer variables as there are a lot of fluctuations in the data so we had to keep all the variables
even though the complexity will increase. But we cannot afford to lose the variation.
coef(regfit.fwd, 85)
```

```
##   (Intercept)       MOSTYPE      MAANTHUI       MGEMOMV      MGEMLEEF 
## -1.357990e-15 -6.663940e-18  2.250506e-17  1.174907e-17 -2.520324e-17 
##      MOSHOOFD        MGODRK        MGODPR        MGODOV        MGODGE 
##  2.999415e-17  5.377831e-18 -5.049559e-18 -6.892938e-18  3.079183e-18 
##        MRELGE        MRELSA        MRELOV      MFALLEEN      MFGEKIND 
## -2.323189e-17 -2.144323e-17 -2.373654e-17  2.593616e-17  2.430566e-17 
##      MFWEKIND      MOPLHOOG      MOPLMIDD      MOPLLAAG      MBERHOOG 
##  2.018239e-17  2.663821e-18  5.914913e-18  1.032406e-17  2.733103e-18 
##      MBERZELF      MBERBOER      MBERMIDD      MBERARBG      MBERARBO 
##  9.193413e-18  6.465524e-18 -9.195897e-18 -1.434499e-18 -3.835161e-18 
##          MSKA         MSKB1         MSKB2          MSKC          MSKD 
## -2.835442e-19  1.193448e-17  9.162417e-18  6.541751e-18  1.322550e-17 
##        MHHUUR        MHKOOP         MAUT1         MAUT2         MAUT0 
##  4.106925e-17  3.608565e-17 -6.402191e-18 -6.635797e-18 -3.563101e-18 
##       MZFONDS        MZPART       MINKM30      MINK3045      MINK4575 
##  1.157356e-16  1.177088e-16 -6.808477e-18 -1.096041e-17 -9.698290e-18 
##      MINK7512      MINK123M       MINKGEM      MKOOPKLA       PWAPART 
## -1.471294e-17  8.564862e-18  2.609361e-18 -7.749850e-18 -2.753531e-17 
##       PWABEDR       PWALAND      PPERSAUT       PBESAUT       PMOTSCO 
##  2.655706e-17  9.045934e-18 -8.705997e-18 -7.427704e-18  5.718082e-17 
##       PVRAAUT      PAANHANG      PTRACTOR        PWERKT         PBROM 
##  7.726107e-17  3.175987e-17  1.900937e-17  1.757003e-17 -8.616540e-19 
##        PLEVEN      PPERSONG       PGEZONG       PWAOREG        PBRAND 
##  3.832487e-17  1.198691e-18 -1.404349e-16  2.357841e-17 -1.382199e-17 
##       PZEILPL      PPLEZIER        PFIETS       PINBOED      PBYSTAND 
##  8.650263e-17 -1.414214e-16  6.426925e-17 -2.384157e-17 -8.087890e-18 
##       AWAPART       AWABEDR       AWALAND      APERSAUT       ABESAUT 
##  4.444005e-17 -1.093568e-16 -3.516693e-17 -3.843519e-18  4.981494e-17 
##       AMOTSCO       AVRAAUT      AAANHANG      ATRACTOR        AWERKT 
## -2.672700e-16 -2.150599e-16 -6.834828e-17 -2.912419e-18 -3.480355e-17 
##         ABROM        ALEVEN      APERSONG       AGEZONG       AWAOREG 
##  2.539695e-17 -8.412037e-17  3.424018e-17  2.061655e-16 -1.835380e-17 
##        ABRAND        AFIETS       AINBOED      ABYSTAND   PurchaseYes 
##  2.464654e-17 -6.953316e-17  3.459071e-17  2.351572e-17  1.000000e+00 
##       AZEILPL 
##  0.000000e+00
```

#Therefore, the positive coefficients show chances of purchase Caravan insurance policy.

#backward selection

```
library(leaps)
set.seed(123)
regfit.bwd <- regsubsets(Purch~., data = caravan_train, nbest = 1, nvmax = 85, method = "backward")
```

```
## Warning in leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax, force.in =
## force.in, : 1 linear dependencies found
```

```
## Reordering variables and trying again:
```

```
## Warning in rval$lopt[] <- rval$vorder[rval$lopt]: number of items to replace is
## not a multiple of replacement length
```

```
# examine the best "p" variables models


summary_bwd <- summary(regfit.bwd)


#identifying the optimal models
which(summary_bwd$cp == min(summary_bwd$cp))
```

```
## [1] 24
```

```
which(summary_bwd$bic == min(summary_bwd$bic))
```

```
## [1] 7
```

```
which(summary_bwd$rss == min(summary_bwd$rss))
```

```
## [1] 85
```

```
which(summary_bwd$adjr2 == max(summary_bwd$adjr2))
```

```
##  [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
## [26] 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50
## [51] 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75
## [76] 76 77 78 79 80 81 82 83 84 85
```

```
##selecting 85 as the best subset forward. We chose 85 from the optimal model although we could have ch
osen fewer variables as there are a lot of fluctuations in the data so we had to keep all the variables
even though the complexity will increase. But we cannot afford to lose the variation.
coef(regfit.bwd, 85)
```

```
##   (Intercept)       MOSTYPE      MAANTHUI       MGEMOMV      MGEMLEEF 
## -9.915682e-16 -4.702376e-18  1.366956e-17  8.611830e-18 -1.633434e-17 
##      MOSHOOFD        MGODRK        MGODPR        MGODOV        MGODGE 
##  2.076419e-17  3.363751e-18 -2.743637e-18 -2.809722e-18  2.511727e-18 
##        MRELGE        MRELSA        MRELOV      MFALLEEN      MFGEKIND 
## -1.517771e-17 -1.342924e-17 -1.542254e-17  1.553239e-17  1.415910e-17 
##      MFWEKIND      MOPLHOOG      MOPLMIDD      MOPLLAAG      MBERHOOG 
##  1.197531e-17  3.183768e-18  4.805199e-18  7.177187e-18  2.949428e-18 
##      MBERZELF      MBERBOER      MBERMIDD      MBERARBG      MBERARBO 
##  6.901808e-18  4.997538e-18 -5.061682e-18  5.096014e-21 -2.352177e-18 
##          MSKA         MSKB1         MSKB2          MSKC          MSKD 
## -2.033979e-18  6.106360e-18  4.353903e-18  3.046682e-18  7.443211e-18 
##        MHHUUR        MHKOOP         MAUT1         MAUT2         MAUT0 
##  2.899289e-17  2.564104e-17 -5.149871e-18 -5.175774e-18 -3.123846e-18 
##       MZFONDS        MZPART       MINKM30      MINK3045      MINK4575 
##  8.881025e-17  9.034171e-17 -4.512273e-18 -7.716008e-18 -7.019417e-18 
##      MINK7512      MINK123M       MINKGEM      MKOOPKLA       PWAPART 
## -1.040673e-17  6.687067e-18  2.221934e-18 -6.824231e-18 -2.862275e-17 
##       PWABEDR       PWALAND      PPERSAUT       PBESAUT       PMOTSCO 
##  2.776990e-17  9.842152e-18 -8.218440e-18 -8.514065e-18  6.081317e-17 
##       PVRAAUT      PAANHANG      PTRACTOR        PWERKT         PBROM 
##  7.725806e-17  3.875753e-17  1.489090e-17  1.487666e-17 -2.265862e-18 
##        PLEVEN      PPERSONG       PGEZONG       PWAOREG        PBRAND 
##  3.665415e-17 -4.145610e-18 -1.366743e-16  1.882602e-17 -1.134612e-17 
##       PZEILPL      PPLEZIER        PFIETS       PINBOED      PBYSTAND 
##  7.137415e-17 -1.644337e-16  7.020814e-17 -2.378398e-17 -7.625487e-18 
##       AWAPART       AWABEDR       AWALAND      APERSAUT       ABESAUT 
##  4.839109e-17 -1.098040e-16 -3.962391e-17 -3.716853e-18  5.057254e-17 
##       AMOTSCO       AVRAAUT      AAANHANG      ATRACTOR        AWERKT 
## -2.827081e-16 -2.172882e-16 -6.627519e-17 -3.011514e-18 -3.614042e-17 
##         ABROM        ALEVEN      APERSONG       AGEZONG       AWAOREG 
##  2.616901e-17 -8.020101e-17  3.221969e-17  2.164520e-16 -1.753621e-17 
##        ABRAND        AFIETS       AINBOED      ABYSTAND   PurchaseYes 
##  2.246802e-17 -7.359418e-17  3.527755e-17  2.302733e-17  1.000000e+00 
##       AZEILPL 
##  0.000000e+00
```

#Therefore, the positive coefficients show chances of purchase Caravan insurance policy.

#Ridge regression

```
library(glmnet)
```

```
## Loading required package: Matrix
```

```
## Warning: package 'Matrix' was built under R version 4.3.3
```

```
## Loaded glmnet 4.1-8
```

```r
set.seed(123)
X_train = model.matrix(Purch~., data = caravan_train)
X_test = model.matrix(Purch~., data = caravan_test)
#Choosing lambda using cross-validation
cv.out = cv.glmnet(X_train, caravan_train$Purch, alpha=0)
sel = cv.out$lambda.min
sel
```

```
## [1] 0.02424007
```

```r
#fitting ridge model
ridge_mod = glmnet(X_train, caravan_train$Purch, alpha = 0, lambda=sel)
#Make predictions
ridge_pred = predict(ridge_mod, s=sel, newx = X_test, type = "response")
#Calculate test error
coef(ridge_mod)
```

```
## 88 x 1 sparse Matrix of class "dgCMatrix"
##                         s0
## (Intercept) -4.549485e-03
## (Intercept)  .
## MOSTYPE      2.119687e-05
## MAANTHUI    -1.320276e-03
## MGEMOMV     -6.434751e-04
## MGEMLEEF     1.430209e-03
## MOSHOOFD    -8.570549e-05
## MGODRK      -2.143193e-04
## MGODPR       3.436368e-04
## MGODOV       4.306365e-04
## MGODGE      -1.451947e-04
## MRELGE       3.620773e-04
## MRELSA       3.971872e-04
## MRELOV       3.672592e-04
## MFALLEEN    -3.688613e-04
## MFGEKIND    -3.212371e-04
## MFWEKIND    -1.213933e-04
## MOPLHOOG     2.186417e-04
## MOPLMIDD    -2.918179e-05
## MOPLLAAG    -2.528715e-04
## MBERHOOG    -2.218276e-04
## MBERZELF    -4.188276e-04
## MBERBOER    -5.656863e-04
## MBERMIDD     4.554514e-04
## MBERARBG    -3.005769e-05
## MBERARBO     1.933767e-04
## MSKA         1.796890e-04
## MSKB1       -3.355508e-04
## MSKB2       -2.446937e-04
## MSKC        -4.825290e-05
## MSKD        -5.252179e-04
## MHHUUR      -1.916784e-04
## MHKOOP       1.565859e-04
## MAUT1        1.801673e-04
## MAUT2        2.504645e-04
## MAUT0        9.087718e-05
## MZFONDS      6.371200e-07
## MZPART      -1.362759e-04
## MINKM30     -5.445171e-05
## MINK3045     2.934219e-04
## MINK4575     2.146168e-04
## MINK7512     7.108674e-04
## MINK123M    -1.084183e-03
## MINKGEM     -1.530490e-04
## MKOOPKLA     5.171060e-04
## PWAPART      5.853392e-04
## PWABEDR     -1.189489e-03
## PWALAND     -1.959988e-04
## PPERSAUT     5.507647e-04
## PBESAUT     -3.495711e-07
## PMOTSCO     -1.292095e-03
## PVRAAUT     -8.112521e-04
## PAANHANG    -7.750952e-04
## PTRACTOR    -1.119091e-03
## PWERKT      -5.489106e-04
```

```
## PBROM       -7.876887e-05
## PLEVEN      -2.322435e-03
## PPERSONG     2.600440e-04
## PGEZONG      3.747301e-03
## PWAOREG     -8.098295e-04
## PBRAND       7.613389e-04
## PZEILPL     -2.537312e-03
## PPLEZIER     6.137893e-03
## PFIETS      -1.501150e-03
## PINBOED      4.606388e-04
## PBYSTAND     3.411442e-04
## AWAPART     -1.334235e-04
## AWABEDR      6.334616e-03
## AWALAND      5.748752e-04
## APERSAUT     1.470086e-03
## ABESAUT     -1.996066e-03
## AMOTSCO      6.873816e-03
## AVRAAUT      2.813825e-04
## AAANHANG     2.579875e-03
## ATRACTOR    -8.521492e-04
## AWERKT       4.245022e-04
## ABROM       -1.570358e-03
## ALEVEN       5.510425e-03
## APERSONG    -4.072764e-03
## AGEZONG      1.501283e-03
## AWAOREG     -2.599437e-03
## ABRAND      -9.807489e-04
## AZEILPL     -3.052556e-03
## APLEZIER     3.664493e-02
## AFIETS       3.900760e-03
## AINBOED      1.381277e-04
## ABYSTAND    -7.211506e-04
## PurchaseYes  9.024521e-01
```

#Therefore, the positive coefficients show likelihood of purchase of Caravan insurance policy.

#LASSO regression

```
set.seed(123)
X_train = model.matrix(Purch~., data = caravan_train)
X_test = model.matrix(Purch~., data = caravan_test)
cv.out2 = cv.glmnet(X_train, caravan_train$Purch, alpha=1)
sel2 = cv.out2$lambda.min
sel2
```

```
## [1] 0.007066108
```

```
#Fitting lasso model
lasso_mod = glmnet(X_train, caravan_train$Purch, alpha=1, lambda=sel2)
#Make predictions
lasso_pred = predict(lasso_mod, s=sel2, newx=X_test)
coef(lasso_mod)
```

```
## 88 x 1 sparse Matrix of class "dgCMatrix"
##                      s0
## (Intercept) 0.001827384
## (Intercept) .
## MOSTYPE     .
## MAANTHUI    .
## MGEMOMV     .
## MGEMLEEF    .
## MOSHOOFD    .
## MGODRK      .
## MGODPR      .
## MGODOV      .
## MGODGE      .
## MRELGE      .
## MRELSA      .
## MRELOV      .
## MFALLEEN    .
## MFGEKIND    .
## MFWEKIND    .
## MOPLHOOG    .
## MOPLMIDD    .
## MOPLLAAG    .
## MBERHOOG    .
## MBERZELF    .
## MBERBOER    .
## MBERMIDD    .
## MBERARBG    .
## MBERARBO    .
## MSKA        .
## MSKB1       .
## MSKB2       .
## MSKC        .
## MSKD        .
## MHHUUR      .
## MHKOOP      .
## MAUT1       .
## MAUT2       .
## MAUT0       .
## MZFONDS     .
## MZPART      .
## MINKM30     .
## MINK3045    .
## MINK4575    .
## MINK7512    .
## MINK123M    .
## MINKGEM     .
## MKOOPKLA    .
## PWAPART     .
## PWABEDR     .
## PWALAND     .
## PPERSAUT    .
## PBESAUT     .
## PMOTSCO     .
## PVRAAUT     .
## PAANHANG    .
## PTRACTOR    .
## PWERKT      .
```

```
## PBROM       .
## PLEVEN      .
## PPERSONG    .
## PGEZONG     .
## PWAOREG     .
## PBRAND      .
## PZEILPL     .
## PPLEZIER    .
## PFIETS      .
## PINBOED     .
## PBYSTAND    .
## AWAPART     .
## AWABEDR     .
## AWALAND     .
## APERSAUT    .
## ABESAUT     .
## AMOTSCO     .
## AVRAAUT     .
## AAANHANG    .
## ATRACTOR    .
## AWERKT      .
## ABROM       .
## ALEVEN      .
## APERSONG    .
## AGEZONG     .
## AWAOREG     .
## ABRAND      .
## AZEILPL     .
## APLEZIER    .
## AFIETS      .
## AINBOED     .
## ABYSTAND    .
## PurchaseYes 0.970849469
```

#Therefore, the positive coefficients show likelihood of purchase of Caravan insurance policy.

#C)Develop a model using logistic regression

```
library(leaps)
logistic_reg <- glm(Purch ~., data = caravan_train, family = "binomial")
```

```
## Warning: glm.fit: algorithm did not converge
```

```
pred_logistic <- predict(logistic_reg, newdata = caravan_test, type = "response")
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from rank-deficient fit; attr(*, "non-estim") has doubtful cases
```

```
summary(logistic_reg)
```

```
##
## Call:
## glm(formula = Purch ~ ., family = "binomial", data = caravan_train)
##
## Coefficients: (1 not defined because of singularities)
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.657e+01  1.071e+06   0.000    1.000
## MOSTYPE     -3.903e-14  5.688e+03   0.000    1.000
## MAANTHUI     9.097e-14  2.007e+04   0.000    1.000
## MGEMOMV     -5.474e-13  1.777e+04   0.000    1.000
## MGEMLEEF    -3.089e-13  1.198e+04   0.000    1.000
## MOSHOOFD     2.363e-13  2.552e+04   0.000    1.000
## MGODRK       5.681e-14  1.375e+04   0.000    1.000
## MGODPR      -1.174e-13  1.469e+04   0.000    1.000
## MGODOV      -1.962e-13  1.327e+04   0.000    1.000
## MGODGE      -1.376e-13  1.415e+04   0.000    1.000
## MRELGE      -4.081e-13  1.906e+04   0.000    1.000
## MRELSA      -3.205e-13  1.798e+04   0.000    1.000
## MRELOV      -5.775e-13  1.922e+04   0.000    1.000
## MFALLEEN     2.321e-13  1.636e+04   0.000    1.000
## MFGEKIND     1.458e-13  1.684e+04   0.000    1.000
## MFWEKIND     2.860e-13  1.754e+04   0.000    1.000
## MOPLHOOG     1.488e-14  1.703e+04   0.000    1.000
## MOPLMIDD    -4.784e-14  1.781e+04   0.000    1.000
## MOPLLAAG     1.428e-13  1.827e+04   0.000    1.000
## MBERHOOG     2.595e-13  1.121e+04   0.000    1.000
## MBERZELF     2.631e-14  1.297e+04   0.000    1.000
## MBERBOER     2.227e-13  1.263e+04   0.000    1.000
## MBERMIDD     4.005e-13  1.131e+04   0.000    1.000
## MBERARBG     2.412e-13  1.096e+04   0.000    1.000
## MBERARBO     2.429e-13  1.115e+04   0.000    1.000
## MSKA         1.876e-13  1.276e+04   0.000    1.000
## MSKB1       -2.479e-13  1.263e+04   0.000    1.000
## MSKB2       -2.828e-14  1.134e+04   0.000    1.000
## MSKC        -2.249e-13  1.235e+04   0.000    1.000
## MSKD        -2.504e-14  1.164e+04   0.000    1.000
## MHHUUR      -1.836e-12  1.006e+05   0.000    1.000
## MHKOOP      -1.825e-12  1.005e+05   0.000    1.000
## MAUT1       -1.644e-13  1.848e+04   0.000    1.000
## MAUT2       -1.760e-13  1.686e+04   0.000    1.000
## MAUT0       -5.245e-14  1.771e+04   0.000    1.000
## MZFONDS      1.338e-12  1.170e+05   0.000    1.000
## MZPART       1.246e-12  1.169e+05   0.000    1.000
## MINKM30      6.161e-14  1.273e+04   0.000    1.000
## MINK3045     7.461e-14  1.217e+04   0.000    1.000
## MINK4575    -7.392e-14  1.228e+04   0.000    1.000
## MINK7512    -4.746e-14  1.275e+04   0.000    1.000
## MINK123M    -1.001e-13  1.672e+04   0.000    1.000
## MINKGEM      1.312e-13  1.097e+04   0.000    1.000
## MKOOPKLA     2.743e-14  5.606e+03   0.000    1.000
## PWAPART     -2.172e-13  4.036e+04   0.000    1.000
## PWABEDR      4.308e-13  6.633e+04   0.000    1.000
## PWALAND     -4.404e-13  1.217e+05   0.000    1.000
## PPERSAUT     1.134e-13  6.803e+03   0.000    1.000
## PBESAUT      8.766e-13  5.825e+04   0.000    1.000
## PMOTSCO      1.307e-12  3.789e+04   0.000    1.000
## PVRAAUT     -3.985e-12  3.980e+05   0.000    1.000
```

```
## PAANHANG     2.413e-12  1.191e+05   0.000    1.000
## PTRACTOR    -4.140e-13  3.352e+04   0.000    1.000
## PWERKT       2.545e-12  4.216e+05   0.000    1.000
## PBROM       -1.708e-13  4.244e+04   0.000    1.000
## PLEVEN       1.101e-12  1.778e+04   0.000    1.000
## PPERSONG     3.757e-13  7.435e+04   0.000    1.000
## PGEZONG     -1.144e-11  1.713e+05   0.000    1.000
## PWAOREG     -8.416e-14  7.820e+04   0.000    1.000
## PBRAND      -1.733e-13  9.086e+03   0.000    1.000
## PZEILPL     -6.143e-13  3.590e+05   0.000    1.000
## PPLEZIER     6.018e-11  8.129e+04   0.000    1.000
## PFIETS      -1.452e-11  1.296e+05   0.000    1.000
## PINBOED     -7.621e-13  9.628e+04   0.000    1.000
## PBYSTAND     6.139e-13  8.443e+04   0.000    1.000
## AWAPART      1.699e-13  7.896e+04   0.000    1.000
## AWABEDR     -3.212e-13  2.000e+05   0.000    1.000
## AWALAND      2.360e-12  4.352e+05   0.000    1.000
## APERSAUT    -5.782e-13  3.340e+04   0.000    1.000
## ABESAUT     -3.471e-12  2.930e+05   0.000    1.000
## AMOTSCO     -6.670e-12  1.696e+05   0.000    1.000
## AVRAAUT      1.244e-11  1.237e+06   0.000    1.000
## AAANHANG    -9.449e-13  2.120e+05   0.000    1.000
## ATRACTOR     7.906e-14  7.613e+04   0.000    1.000
## AWERKT      -5.727e-12  9.678e+05   0.000    1.000
## ABROM        2.580e-13  1.286e+05   0.000    1.000
## ALEVEN      -3.632e-12  3.751e+04   0.000    1.000
## APERSONG    -1.818e-12  2.162e+05   0.000    1.000
## AGEZONG      2.969e-11  4.104e+05   0.000    1.000
## AWAOREG      5.253e-13  3.806e+05   0.000    1.000
## ABRAND       3.613e-13  2.892e+04   0.000    1.000
## AZEILPL             NA         NA      NA       NA
## APLEZIER    -1.287e-10  2.708e+05   0.000    1.000
## AFIETS       1.114e-11  8.985e+04   0.000    1.000
## AINBOED     -6.758e-13  1.940e+05   0.000    1.000
## ABYSTAND    -2.448e-12  2.976e+05   0.000    1.000
## PurchaseYes  5.313e+01  3.191e+04   0.002    0.999
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1.0914e+03  on 2328  degrees of freedom
## Residual deviance: 1.3512e-08  on 2243  degrees of freedom
## AIC: 172
##
## Number of Fisher Scoring iterations: 25
```

#We know that, the coefficients of predictor variables reveal how changes in these factors affect the likelihood of someone being interested in buying caravan insurance. Predictor with a zero or close to zero coefficient has almost no impact on the likelihood of purchase. #Variables with positive coefficients indicate that an increase in their values increases the chances of interest in caravan insurance. #Negative coefficients of predictors indicates that an increase in their values is means less chances of interest in caravan insurance. #We see that the "PurchaseYes" variable has a coefficient of 5.313e+01, which means if a customer has already purchased caravan insurance (PurchaseYes = 1), then, there is a higher chance of the customer being interested in buying it again. This shows a strong positive connection between past purchases and current interest. #So, we can say that, customers who have previously purchased caravan insurance (PurchaseYes = 1) are more likely to be interested in buying it again.

Data Preparation and Linear Regression Model

Data Preparation

First, I loaded the Caravan dataset from the ISLR2 package. I converted the purchase indicator to a numeric variable, where 1 indicates a customer purchased caravan insurance and 0 indicates they did not. To ensure reproducibility, I set a seed for random number generation. Then, I split the data into a training set (40% of the data) and a test set (60%).

Linear Regression Model

I started by developing a linear regression model to predict the likelihood of purchasing caravan insurance. After fitting the model to the training data, I used it to make predictions on the test set. By analyzing the model summary, I identified which variables had significant coefficients—those with lower p-values and more asterisks ('*')—indicating they were likely to influence the purchase decision.

Model Development with Selection Methods

Forward Selection Next, I applied forward selection using the regsubsets function from the leaps package. This method iteratively added variables to the model to find the best subset that minimizes the criteria like Cp, BIC, RSS, and maximizes adjusted R^2. Although forward selection identified several models, I chose to include all 85 variables in the final model because of the data's complexity and variability.

Backward Selection

Similarly, I used backward selection, which starts with all variables and removes the least significant ones. Again, I analyzed the models based on Cp, BIC, RSS, and adjusted R^2. Despite the ability to select fewer variables, I decided to retain all 85 due to the fluctuations in the data.

Ridge and Lasso Regression

Ridge Regression

For the Ridge regression, I used the glmnet package. I created model matrices for the training and test sets and selected the optimal regularization parameter (lambda) through cross-validation. After fitting the Ridge model with this lambda, I predicted the test data and reviewed the coefficients. The positive coefficients indicated a higher likelihood of purchasing caravan insurance.

Lasso Regression

Lasso regression was applied similarly, but with an alpha value of 1, which enforces more sparsity in the model coefficients. After selecting the optimal lambda through cross-validation, I fitted the Lasso model and made predictions. Again, the coefficients helped me understand which variables were most influential in predicting insurance purchase likelihood.

###Logistic Regression Model

Developing the Logistic Regression Model Finally, I developed a logistic regression model to predict the binary outcome of purchasing caravan insurance. After fitting the model to the training data, I made predictions on the test set and examined the summary of the logistic regression.

Interpretation of Coefficients

In the logistic model, I focused on the coefficients of the predictor variables to understand their impact. Positive coefficients indicated that an increase in those variables increases the likelihood of purchasing insurance. Negative coefficients indicated a decrease in likelihood. Notably, the "PurchaseYes" variable had a high positive coefficient, suggesting that customers who previously purchased caravan insurance are more likely to be interested in buying it again. This strong positive connection between past purchases and current interest highlights a significant predictor of future sales.