Final Project

Stat 158

Professor Peng Ding

Jooho Oh, Joowon Yang

Data : Bank Marketing Data Set

Link : https://archive.ics.uci.edu/ml/datasets/Bank+Marketing#

With heightened interest from the group in seeing what demographics are more susceptible to telemarketing from banks, the main question for this project arose: what are the impacts of characteristics of an individual, such as their age, marital status, education level, etc., that influence their likelihood of committing to a telemarketing call? More critically, the group hopes to find a few dominant characteristics that may hopefully explain most of the variance from our dependent variable through methods of linear modeling. One real-world insight that can potentially be learned is how to identify individuals who are likely to commit to fixed-income products out of a pool of potential candidates. And more importantly, this could give insight into the fiscal tendencies of various demographics through our modeling.

The primary focus of the model should be prediction focused, as the identification of individuals that are susceptible to bank telemarketing is not only useful for bankers but also for individuals to protect their own privacy. However, modeling this banking data could also be helpful in dimensionality reduction. There are many factors that an individual may have that are useful in financial decision-making, but if the model indicates that only a few of the features are useful, then perhaps it could increase the interpretability.

In this modeling section, we want to answer the following question: which feature should we include in our GLM model in order to minimize the mean square error of our prediction? We considered LASSO/Ridge Regression to apply penalties to models with redundant features (shrinkage methods), Forward/Backward selection to select a pseudo-optimal model that would hopefully minimize the mean

square error without going through the MSE of every single possible model, and ANOVA to determine which independent variable has the greatest impact on the binary response variable y.

**Method 1**: Model Selection

In this section, we will randomly split the data into a training set and a testing set. Firstly, we will use the training data to find the pseudo-optimal model using forward selection with AIC, backward selection with AIC, forward selection with BIC, and backward selection with BIC. Then, we will compare the cross-validation mean square error of the four models' predictions. The model with the lowest MSE should be the best model out of the four. And finally, we will make sure that the model is the best by giving four different sets of predictions on the testing data using the four different models, and comparing their MSE of the four sets of predictions. The four different models are shown in the tables below.

```r
forwardAIC_model = step(intercept_model,
                        scope = list(lower=intercept_model, upper=train.glm),
                        direction="forward",
                        k=2,
                        trace=FALSE)
```

```r
summary(forwardAIC_model)
```

```
##
## Call:
## glm(formula = y_boolean ~ duration + nr.employed + month + poutcome +
##     emp.var.rate + cons.price.idx + contact + cons.conf.idx +
##     default + pdays + day_of_week + campaign + job + euribor3m +
##     previous, family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.9207  -0.3018  -0.1864  -0.1359   3.3312
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -2.575e+02  4.543e+01  -5.667 1.45e-08 ***
## duration             4.648e-03  8.839e-05  52.589  < 2e-16 ***
## nr.employed          7.180e-03  3.695e-03   1.943 0.052021 .
## monthaug             7.737e-01  1.432e-01   5.401 6.62e-08 ***
## monthdec             4.532e-01  2.425e-01   1.869 0.061646 .
## monthjul             5.634e-02  1.145e-01   0.492 0.622771
## monthjun            -5.723e-01  1.496e-01  -3.826 0.000130 ***
## monthmar             1.998e+00  1.728e-01  11.563  < 2e-16 ***
## monthmay            -4.840e-01  9.832e-02  -4.923 8.52e-07 ***
## monthnov            -5.165e-01  1.450e-01  -3.562 0.000367 ***
## monthoct             1.515e-01  1.826e-01   0.830 0.406558
## monthsep             4.089e-01  2.128e-01   1.922 0.054649 .
## poutcomenonexistent  3.582e-01  1.132e-01   3.165 0.001551 **
## poutcomesuccess      6.942e-01  2.486e-01   2.792 0.005235 **
## emp.var.rate        -1.774e+00  1.686e-01 -10.522  < 2e-16 ***
## cons.price.idx       2.330e+00  2.993e-01   7.785 6.96e-15 ***
## contacttelephone    -7.347e-01  9.322e-02  -7.882 3.23e-15 ***
## cons.conf.idx        3.032e-02  9.258e-03   3.275 0.001057 **
## defaultunknown      -3.360e-01  7.910e-02  -4.247 2.16e-05 ***
## defaultyes          -7.302e+00  1.385e+02  -0.053 0.957950
## pdays               -1.257e-03  2.552e-04  -4.925 8.44e-07 ***
## day_of_weekmon      -1.397e-01  7.871e-02  -1.774 0.075984 .
## day_of_weekthu       2.342e-02  7.650e-02   0.306 0.759517
## day_of_weektue       8.014e-02  7.812e-02   1.026 0.304912
## day_of_weekwed       1.994e-01  7.795e-02   2.558 0.010533 *
## campaign            -3.813e-02  1.366e-02  -2.792 0.005243 **
## jobblue-collar      -2.921e-01  7.839e-02  -3.727 0.000194 ***
## jobentrepreneur     -1.610e-01  1.469e-01  -1.096 0.273045
## jobhousemaid        -1.919e-01  1.738e-01  -1.105 0.269374
## jobmanagement       -2.767e-02  9.975e-02  -0.277 0.781465
## jobretired           1.211e-01  1.002e-01   1.208 0.226919
## jobself-employed    -1.216e-01  1.396e-01  -0.871 0.383641
```

```
## jobservices         -1.558e-01  9.571e-02  -1.628 0.103513
## jobstudent           1.394e-01  1.218e-01   1.144 0.252649
## jobtechnician       -2.158e-02  7.579e-02  -0.285 0.775834
## jobunemployed       -1.636e-01  1.553e-01  -1.054 0.291888
## jobunknown          -2.268e-01  3.105e-01  -0.730 0.465243
## euribor3m            2.582e-01  1.543e-01   1.673 0.094287 .
## previous            -1.136e-01  7.184e-02  -1.582 0.113703
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 20357  on 28831  degrees of freedom
## Residual deviance: 12016  on 28793  degrees of freedom
## AIC: 12094
##
## Number of Fisher Scoring iterations: 10
```

```
backwardAIC_model = step(train.glm,
                         scope = list(lower=intercept_model, upper=train.glm),
                         direction="backward",
                         k=2,
                         trace=FALSE)
```

```
summary(backwardAIC_model)
```

```
##
## Call:
## glm(formula = y_boolean ~ job + default + contact + month + day_of_week +
##     duration + campaign + pdays + previous + poutcome + emp.var.rate +
##     cons.price.idx + cons.conf.idx + euribor3m + nr.employed,
##     family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.9207  -0.3018  -0.1864  -0.1359   3.3312
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -2.575e+02  4.543e+01  -5.667 1.45e-08 ***
## jobblue-collar      -2.921e-01  7.839e-02  -3.727 0.000194 ***
## jobentrepreneur     -1.610e-01  1.469e-01  -1.096 0.273045
## jobhousemaid        -1.919e-01  1.738e-01  -1.105 0.269374
## jobmanagement       -2.767e-02  9.975e-02  -0.277 0.781465
## jobretired           1.211e-01  1.002e-01   1.208 0.226919
## jobself-employed    -1.216e-01  1.396e-01  -0.871 0.383641
## jobservices         -1.558e-01  9.571e-02  -1.628 0.103513
## jobstudent           1.394e-01  1.218e-01   1.144 0.252649
## jobtechnician       -2.158e-02  7.579e-02  -0.285 0.775834
## jobunemployed       -1.636e-01  1.553e-01  -1.054 0.291888
## jobunknown          -2.268e-01  3.105e-01  -0.730 0.465243
## defaultunknown      -3.360e-01  7.910e-02  -4.247 2.16e-05 ***
## defaultyes          -7.302e+00  1.385e+02  -0.053 0.957950
```

```
## contacttelephone    -7.347e-01  9.322e-02  -7.882 3.23e-15 ***
## monthaug             7.737e-01  1.432e-01   5.401 6.62e-08 ***
## monthdec             4.532e-01  2.425e-01   1.869 0.061646 .
## monthjul             5.634e-02  1.145e-01   0.492 0.622771
## monthjun            -5.723e-01  1.496e-01  -3.826 0.000130 ***
## monthmar             1.998e+00  1.728e-01  11.563  < 2e-16 ***
## monthmay            -4.840e-01  9.832e-02  -4.923 8.52e-07 ***
## monthnov            -5.165e-01  1.450e-01  -3.562 0.000367 ***
## monthoct             1.515e-01  1.826e-01   0.830 0.406558
## monthsep             4.089e-01  2.128e-01   1.922 0.054649 .
## day_of_weekmon      -1.397e-01  7.871e-02  -1.774 0.075984 .
## day_of_weekthu       2.342e-02  7.650e-02   0.306 0.759517
## day_of_weektue       8.014e-02  7.812e-02   1.026 0.304912
## day_of_weekwed       1.994e-01  7.795e-02   2.558 0.010533 *
## duration             4.648e-03  8.839e-05  52.589  < 2e-16 ***
## campaign            -3.813e-02  1.366e-02  -2.792 0.005243 **
## pdays               -1.257e-03  2.552e-04  -4.925 8.44e-07 ***
## previous            -1.136e-01  7.184e-02  -1.582 0.113703
## poutcomenonexistent  3.582e-01  1.132e-01   3.165 0.001551 **
## poutcomesuccess      6.942e-01  2.486e-01   2.792 0.005235 **
## emp.var.rate        -1.774e+00  1.686e-01 -10.522  < 2e-16 ***
## cons.price.idx       2.330e+00  2.993e-01   7.785 6.96e-15 ***
## cons.conf.idx        3.032e-02  9.258e-03   3.275 0.001057 **
## euribor3m            2.582e-01  1.543e-01   1.673 0.094287 .
## nr.employed          7.180e-03  3.695e-03   1.943 0.052021 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 20357  on 28831  degrees of freedom
## Residual deviance: 12016  on 28793  degrees of freedom
## AIC: 12094
##
## Number of Fisher Scoring iterations: 10
```

```
forwardBIC_model = step(intercept_model,
                        scope = list(lower=intercept_model, upper=train.glm),
                        direction="forward",
                        k=log(dim(train)[1]),
                        trace=FALSE)
```

```
summary(forwardBIC_model)
```

```
##
## Call:
## glm(formula = y_boolean ~ duration + nr.employed + month + poutcome +
##     emp.var.rate + cons.price.idx + contact + cons.conf.idx +
##     pdays + default, family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.9154  -0.3045  -0.1869  -0.1380   3.2247
```

```
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -3.011e+02  3.735e+01  -8.061 7.56e-16 ***
## duration              4.639e-03  8.798e-05  52.731  < 2e-16 ***
## nr.employed           1.181e-02  2.450e-03   4.820 1.44e-06 ***
## monthaug              8.602e-01  1.398e-01   6.152 7.66e-10 ***
## monthdec              5.962e-01  2.299e-01   2.593 0.009504 **
## monthjul              9.696e-02  1.121e-01   0.865 0.387031
## monthjun             -5.933e-01  1.480e-01  -4.008 6.13e-05 ***
## monthmar              2.105e+00  1.645e-01  12.799  < 2e-16 ***
## monthmay             -4.582e-01  9.504e-02  -4.821 1.43e-06 ***
## monthnov             -3.366e-01  1.149e-01  -2.929 0.003405 **
## monthoct              3.625e-01  1.496e-01   2.423 0.015411 *
## monthsep              6.097e-01  1.852e-01   3.292 0.000996 ***
## poutcomenonexistent   5.038e-01  7.627e-02   6.605 3.97e-11 ***
## poutcomesuccess       8.023e-01  2.381e-01   3.369 0.000754 ***
## emp.var.rate         -1.770e+00  1.681e-01 -10.528  < 2e-16 ***
## cons.price.idx        2.551e+00  2.681e-01   9.517  < 2e-16 ***
## contacttelephone     -7.436e-01  9.276e-02  -8.016 1.09e-15 ***
## cons.conf.idx         4.426e-02  6.544e-03   6.763 1.35e-11 ***
## pdays                -1.127e-03  2.373e-04  -4.749 2.05e-06 ***
## defaultunknown       -3.873e-01  7.779e-02  -4.979 6.39e-07 ***
## defaultyes           -7.220e+00  1.387e+02  -0.052 0.958487
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 20357  on 28831  degrees of freedom
## Residual deviance: 12077  on 28811  degrees of freedom
## AIC: 12119
##
## Number of Fisher Scoring iterations: 10
```

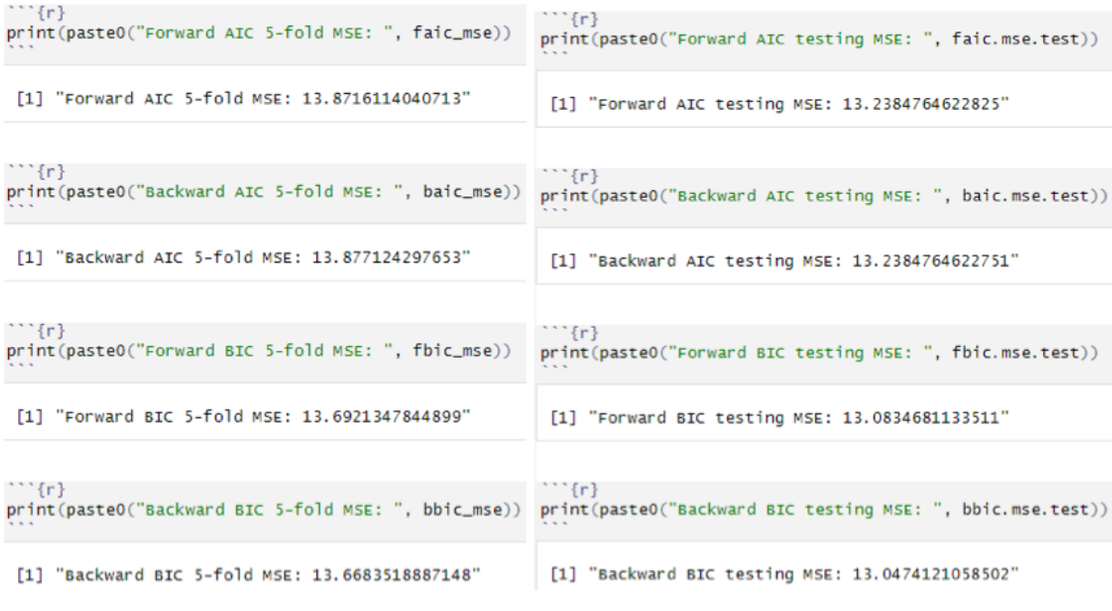
```
backwardBIC_model = step(train.glm,
                         scope = list(lower=intercept_model, upper=train.glm),
                         direction="backward",
                         k=log(dim(train)[1]),
                         trace=FALSE)
```

```
summary(backwardBIC_model)
```

```
##
## Call:
## glm(formula = y_boolean ~ default + contact + month + duration +
##     pdays + poutcome + emp.var.rate + cons.price.idx + euribor3m,
##     family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -5.9718  -0.3027  -0.1861  -0.1382   3.1460
##
```

```
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -1.839e+02  1.181e+01 -15.576  < 2e-16 ***
## defaultunknown      -3.800e-01  7.780e-02  -4.884 1.04e-06 ***
## defaultyes          -7.264e+00  1.388e+02  -0.052 0.958276
## contacttelephone    -6.291e-01  8.471e-02  -7.426 1.12e-13 ***
## monthaug             9.299e-01  1.016e-01   9.153  < 2e-16 ***
## monthdec             4.605e-01  2.194e-01   2.099 0.035835 *
## monthjul             1.442e-01  1.091e-01   1.322 0.186195
## monthjun            -4.422e-01  1.252e-01  -3.532 0.000412 ***
## monthmar             1.910e+00  1.351e-01  14.137  < 2e-16 ***
## monthmay            -5.134e-01  8.775e-02  -5.851 4.89e-09 ***
## monthnov            -5.832e-01  1.270e-01  -4.591 4.41e-06 ***
## monthoct             1.129e-01  1.391e-01   0.812 0.416906
## monthsep             3.641e-01  1.384e-01   2.630 0.008531 **
## duration             4.646e-03  8.799e-05  52.807  < 2e-16 ***
## pdays               -1.134e-03  2.376e-04  -4.772 1.83e-06 ***
## poutcomenonexistent  4.758e-01  7.613e-02   6.250 4.11e-10 ***
## poutcomesuccess      8.165e-01  2.386e-01   3.422 0.000621 ***
## emp.var.rate        -1.784e+00  1.228e-01 -14.532  < 2e-16 ***
## cons.price.idx       1.908e+00  1.232e-01  15.481  < 2e-16 ***
## euribor3m            6.280e-01  9.360e-02   6.709 1.96e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 20357  on 28831  degrees of freedom
## Residual deviance: 12085  on 28812  degrees of freedom
## AIC: 12125
##
## Number of Fisher Scoring iterations: 10
```

In the end, the backward selection with BIC gives a ten-feature model with the lowest cross-validation MSE as well as the lowest testing MSE.



However, the nature of forward/backward selection decides that this model might not be the actual optimal model. Some of the remaining features of this model are macro-economic features, including euribor3m, cons.price.idx, emp.var.rate. This makes sense because the interest rate, inflation rate, and employment rate are highly correlated with the market demand for bank savings by nature. The demographic profile of the potential customer does not seem to matter much since age, job, marital status, and education level are not included in this pseudo-optimal model. The only important demographic status is whether or not the customer has defaulted in the past. The marketing effort does matter since the duration of the marketing call and the contact communication type can influence the success rate. However, we have to note that "duration" should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model. Customer loyalty also seems to play a role in the success rate of the marketing campaign since the outcome is in the model and thus, the success/failure of the last marketing campaign can influence the success rate of a future campaign. The most surprising result is that the month of the year is important in determining the success rate. However, this might be due to the fact that this data was collected during the 2008-2010 financial crisis, and there might be several months in those three years that had particularly bad macro-economics conditions. Maybe customers in those months with "bad economies" might resort to or stay away from bank savings. This is worth further investigation in future research.

**Method 2**: Shrinkage Methods

The two most used and useful shrinkage methods are the Lasso and Ridge regressions. These methods are used to reduce model complexity while preventing over-fitting which may happen when simple regressions are learned. The ride regression adds a penalty proportional to the sum of the squares of the coefficient. This helps us reduce complexity and multi-colinearity. On the other hand, Lasso regression penalizes the sum of absolute values of the coefficients. This helps reduce over-fitting and helps with feature selection.

In terms of assumptions, both Lasso and ridge are parametric functions, so the linearity assumption needs to be met. However, both Lasso and Ridge are able to handle multicollinearity, so this assumption does not need to be met for our shrinkage methods. However, the independence and homoscedasticity assumptions should be met in order for the shrinkage methods to take full advantage of the regularizing terms.

We first benchmarked the generalized linear model without any regularizing terms to establish a baseline. Again, we are predicting the outcome of a banking marketing call using information about the individual and macroeconomic factors. Here are the results from the unregularized general linear model.

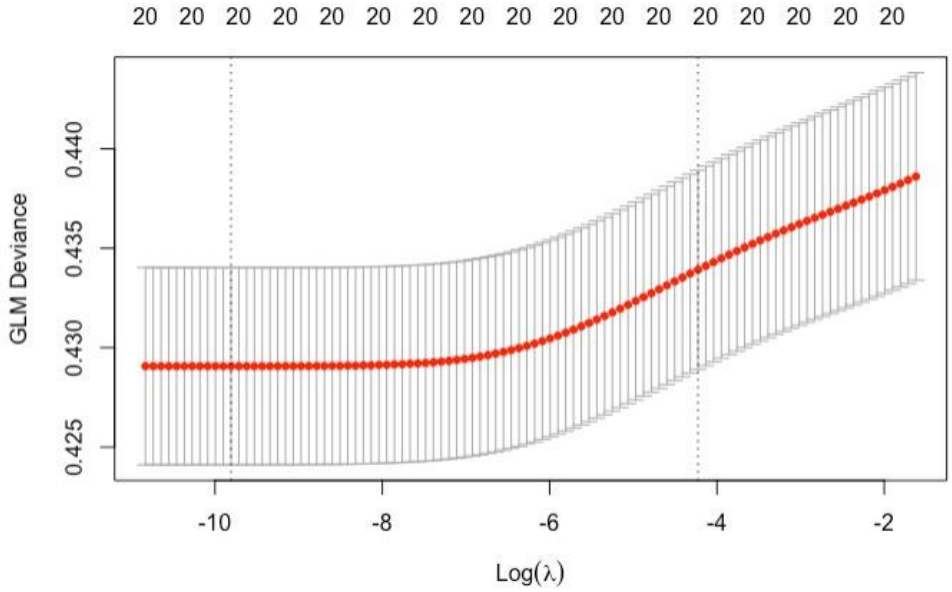
```
[1] "the unregularized MSE is: 0.0638723851269843"
[1] "the unregularized accuracy is: 0.910246034315312"
```

And the intercepts are shown here.

| (Intercept) | age | jobblue-collar | jobentrepreneur |
|---|---|---|---|
| -2.592084e+02 | -5.188314e-04 | -2.230757e-01 | -1.543479e-01 |
| jobhousemaid | jobmanagement | jobretired | jobself-employed |
| -1.330020e-01 | -3.910358e-02 | 1.839547e-01 | -1.270466e-01 |
| jobservices | jobstudent | jobtechnician | jobunemployed |
| -1.110349e-01 | 1.574635e-01 | -3.973814e-03 | -1.252125e-01 |
| jobunknown | maritalmarried | maritalsingle | maritalunknown |
| -2.089915e-01 | 3.440455e-02 | 4.207444e-02 | 1.514376e-01 |
| educationbasic.6y | educationbasic.9y | educationhigh.school | educationilliterate |
| -1.315362e-02 | -9.940191e-03 | 3.507052e-02 | 1.484119e+00 |
| educationprofessional.course | educationuniversity.degree | educationunknown | defaultunknown |
| 4.540590e-02 | 1.208174e-01 | 1.078687e-01 | -3.288056e-01 |
| defaultyes | housingunknown | housingyes | loanunknown |
| -7.295241e+00 | -6.706570e-02 | -1.776312e-02 | NA |
| loanyes | contacttelephone | monthaug | monthdec |
| -8.332179e-02 | -7.355614e-01 | 7.636090e-01 | 4.605033e-01 |
| monthjul | monthjun | monthmar | monthmay |
| 5.486555e-02 | -5.818085e-01 | 1.991649e+00 | -4.789203e-01 |
| monthnov | monthoct | monthsep | day_of_weekmon |
| -5.174196e-01 | 1.574934e-01 | 4.113721e-01 | -1.395910e-01 |
| day_of_weekthu | day_of_weektue | day_of_weekwed | duration |
| 2.131361e-02 | 8.364798e-02 | 1.997486e-01 | 4.651347e-03 |
| campaign | pdays | previous | poutcomenonexistent |
| -3.826762e-02 | -1.271723e-03 | -1.136690e-01 | 3.566130e-01 |
| poutcomesuccess | emp.var.rate | cons.price.idx | cons.conf.idx |
| 6.783750e-01 | -1.777293e+00 | 2.339730e+00 | 3.015945e-02 |
| euribor3m | nr.employed | | |
| 2.543907e-01 | 7.327428e-03 | | |

As we can see, all of the features are present in this regression, making inference somewhat difficult, and we do not have a clear understanding of what are the most important driving factors in predicting whether someone will be susceptible to marketing calls.

The first shrinkage method that was used is Ridge regression which penalized the squared sum of the coefficients. We created a custom sequence of lambdas to feed into the cv.glmnet function since there has been literature stating that the default values are not optimal. The graph below demonstrates the optimal log values of lambdas from the cross-validation that the package performs. Since we want the optimal MSE scores, we selected lambda.min value when building our optimal ridge regression model.




The accuracy measurements are shown here.

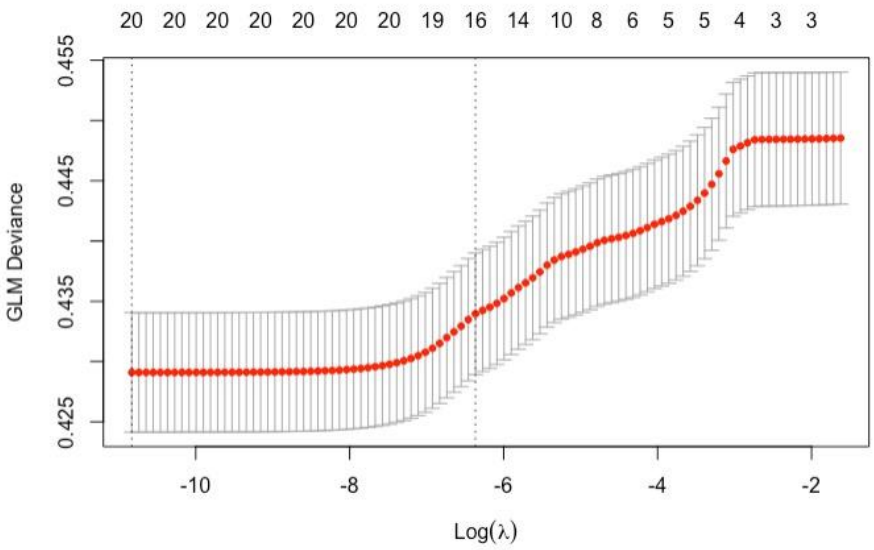
```
[1] "the ridge regression MSE is: 0.0633577130946119"
[1] "the ridge regression accuracy is: 0.911379087083198"
```

It has been previously shown that ridge regression should perform strictly not worse than the regularized version. And luckily our results confirm this. The ridge MSE is smaller than the 0.0638 value from the unregularized general linear model, and the ridge regression accuracy is also higher than the unregularized accuracy. This is perhaps because the penalty added in ridge regression helps reduce overfitting and gives us better testing accuracy.

| | s1 |
|---|---|
| (Intercept) | -16.110831784 |
| age | 0.003996037 |
| job | 0.005872729 |
| marital | 0.081757932 |
| education | 0.042261375 |
| default | -0.397795178 |
| housing | -0.015310409 |
| loan | -0.047068866 |
| contact | -0.747654334 |
| month | -0.111748702 |
| day_of_week | 0.057321509 |
| duration | 0.004527507 |
| campaign | -0.030454284 |
| pdays | -0.001253130 |
| previous | -0.131816215 |
| poutcome | 0.359423661 |
| emp.var.rate | -0.873639881 |
| cons.price.idx | 0.789834499 |
| cons.conf.idx | 0.028211614 |
| euribor3m | 0.536144337 |
| nr.employed | -0.011748028 |

Here are the coefficients that are from the ridge regression outputs. We see that compared to the unregularized version, the coefficients are significantly smaller (on the magnitude of 10x on average). However, none of the coefficients are zero, since the squared sum is penalized not the absolute value. One coefficient to highlight here is the poutcome coefficient, which suggests that if an individual has committed before, they are more likely to commit again. Another coefficient to highlight is the contact variable's coefficient. Combined with poutcomes, we see that if an individual has been contacted before, but didn't commit, it's very unlikely that they will commit on the next call.

The last shrinkage method that was performed is the Lasso regression. The motivation here is similar to why the Ridge regression was used. We want to limit overfitting while attempting to understand the most important features of the predictions. The results are as follows:




In the graph above, we ran the same lambda schedule as for the ridge regression. Again, this is different than the default lambda sequence from cv.glmnet. We see that for larger log values of lambda, the deviance is higher, perhaps because the penalty is too large, and many important features are being truncated, causing a deterioration in performance from the model. Here, we select the right dotted vertical line or the lambda.1se value to be used, since we want to promote sparsity in the data.

```
[1] "the lasso regression MSE is: 0.0633616798599029"
[1] "the lasso regression accuracy is: 0.911136290061509"
```

By invoking the Lasso regression, the regression MSE also improves from the baseline no-shrinkage model. Perhaps this is from benefitting from less overfitting. We also see that the regression accuracy improves too, giving us more predictive power for our model.

| | s1 |
|---|---|
| (Intercept) | 66.480254006 |
| age | 0.001714778 |
| job | 0.003878356 |
| marital | . |
| education | 0.048865180 |
| default | -0.163906546 |
| housing | . |
| loan | . |
| contact | -0.306069402 |
| month | -0.098119270 |
| day_of_week | 0.044599896 |
| duration | 0.004495338 |
| campaign | -0.028404772 |
| pdays | -0.001453730 |
| previous | -0.108295274 |
| poutcome | 0.179014172 |
| emp.var.rate | -0.086476972 |
| cons.price.idx | . |
| cons.conf.idx | 0.032732861 |
| euribor3m | 0.007578947 |
| nr.employed | -0.013072061 |

The coefficients from the model are shown above, where the dots indicate that the coefficient it zero, or in other words, the features are not included in the model. This behavior is expected from the Lasso shrinkage method. Furthermore, we see that marital status, housing, loan, and cons.price.idx are the features that are removed. It makes intuitive sense that martial status isn't very helpful in predicting whether someone will commit to a banking call, and we can understand the removal of cons.price.idx by the multicollinearity of the macroeconomic factors shown in the previous EDA report.

From the modeling portion, we see that we are able to predict whether an individual will commit to a banking marketing call with more than 91.1% accuracy, as that is the main objective of the model. Furthermore, we are able to see that marital status, housing, loan, and cons.price.idx are less significant when it comes to predictive power.

**Method 3**: Categorical variables / ANOVA

The main question we want to answer from this method is which factors are most important in predicting whether a client will subscribe to the term deposit or not. To answer this question, we will use a generalized linear model (GLM) with a binary response variable indicating whether a client subscribed to the term deposit or not. Before we build the GLM model, it is important to identify which features to include in the model in order to minimize the mean square error of our prediction. To do this, we will perform an analysis of variance (ANOVA) on the dataset to identify which features have a significant

effect on the response variable. By including only the most significant features in our model, we can reduce the complexity of the model and improve its predictive power.

In this context, ANOVA provides a useful tool for identifying the features that have the greatest impact on the response variable. We will use the results of the ANOVA to guide our selection of features for the GLM model, allowing us to build a more accurate and interpretable model for predicting the likelihood of a client subscribing to the term deposit.

There are several assumptions associated with ANOVA that need to be checked to ensure the validity of the analysis.

- Independence: The observations should be independent of each other. In this dataset, each observation corresponds to a unique client, so this assumption is met.
- Normality: The residuals should be normally distributed. We can check this assumption by plotting a histogram of the residuals or by using a normal probability plot, but it was difficult to check this assumption directly on the bank-additional-full dataset because we don't have the actual predicted values to calculate the residuals.
- Homogeneity of variance: The variances of the different groups should be approximately equal. We can check this assumption by plotting a box plot of the residuals for each group.
- Independence of groups: The groups being compared should be independent of each other. In this dataset, the groups are independent because each client is assigned to only one group.

Overall, the bank-additional-full dataset seems to meet the assumptions required for ANOVA. However, it is important to note that there may be other unmeasured factors that could impact the validity of the analysis.

In this analysis, we will be using the bank marketing dataset to perform an analysis of variance (ANOVA) to determine the significance of the relationship between various independent variables and the binary response variable 'y-binary'. THe bank marketing dataset contains information on the marketing campaign of a Portuguese banking institution, including attributes of the clients, the marketing campaign, and the final outcome of the campaign.

First, we load the bank marketing dataset into R and convert the response variable ‘y’ to a binary numeric variable ‘y-binary’, with 1 representing a successful marketing campaign and 0 representing an unsuccessful campaign. Then, we use the ‘aov’ function in R to perform the ANOVA on ‘y_binary’ and all other independent variables in the dataset.

```r
# Load the bank marketing dataset
bank_data <- read.csv("~/Desktop/151 final proj/bank-additional-full.csv", sep=";")

# Convert the response variable y to a binary numeric variable
bank_data$y_binary <- ifelse(bank_data$y == "yes", 1, 0)

#anova model
anova_model <- aov(y_binary ~ ., data = bank_data)

summary(anova_model)
```

```
##                   Df Sum Sq Mean Sq   F value Pr(>F)
## age                1    3.8     3.8 4.246e+31 <2e-16 ***
## job               11   92.8     8.4 9.415e+31 <2e-16 ***
## marital            3    8.1     2.7 2.999e+31 <2e-16 ***
## education          7    5.8     0.8 9.185e+30 <2e-16 ***
## default            2   31.2    15.6 1.739e+32 <2e-16 ***
## housing            2    0.3     0.1 1.604e+30 <2e-16 ***
## loan               1    0.2     0.2 1.748e+30 <2e-16 ***
## contact            1   57.7    57.7 6.437e+32 <2e-16 ***
## month              9  221.8    24.6 2.750e+32 <2e-16 ***
## day_of_week        4    4.1     1.0 1.131e+31 <2e-16 ***
## duration           1  639.7   639.7 7.139e+33 <2e-16 ***
## campaign           1    0.7     0.7 7.380e+30 <2e-16 ***
## pdays              1  225.7   225.7 2.518e+33 <2e-16 ***
## previous           1    0.5     0.5 5.086e+30 <2e-16 ***
## poutcome           2    4.2     2.1 2.325e+31 <2e-16 ***
## emp.var.rate       1  102.2   102.2 1.141e+33 <2e-16 ***
## cons.price.idx     1   53.7    53.7 5.994e+32 <2e-16 ***
## cons.conf.idx      1   17.2    17.2 1.917e+32 <2e-16 ***
## euribor3m          1    4.9     4.9 5.496e+31 <2e-16 ***
## nr.employed        1    0.0     0.0 2.655e+29 <2e-16 ***
## y                  1 2642.9  2642.9 2.949e+34 <2e-16 ***
## Residuals      41134    0.0     0.0
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

The ANOVA results are shown in the table above. The ANOVA table shows the results of testing the significance of each variable in the model, including age, job, marital status, education, default, housing, loan, contact, month, day of week, duration, campaign, pdays, previous, poutcome, emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, nr.employed, and the response variable y. Each row represents an independent variable, with columns showing the degrees of freedom (Df), the sum of squares (Sum Sq), the mean sum of squares (Mean Sq), the F-value, and the p-value (Pr(>F)).

We can see that all independent variables, except for ‘nr.employed’ and ‘loan’, have extremely low p-values (less than 2e-16), indicating that they have a significant relationship with the response

variable ‘y_binary’. This suggests that these variables may be important predictors of the success of a marketing campaign.

Furthermore, the F-values for each variable are extremely large, indicating a large amount of variance in the response variable that can be explained by each independent variable. For example, the variable 'duration' has an F-value of 7.139e+33, indicating that it explains a large amount of the variance in 'y_binary'. It should be noted that including the feature 'duration' in the model may only serve as a benchmark for comparison, and it may not be appropriate to include it in a realistic predictive model.

Overall, this ANOVA analysis suggests that there are several independent variables in the bank marketing dataset that are significantly related to the success of a marketing campaign. These variables include age, job, marital status, education, default, housing, contact, month, day of week, duration, campaign, pdays, previous, poutcome, emp.var.rate, cons.price.idx, cons.conf.idx, and euribor3m.

It is important to note that these variables may be correlated with each other, which could impact the results of any predictive model. Further analysis, such as correlation analysis or regression analysis, could be performed to investigate these relationships in more detail.

In conclusion, this ANOVA analysis of the bank marketing dataset has identified several independent variables that are significantly related to the success of a marketing campaign. These variables should be considered when developing a predictive model for marketing campaign success. However, it is important to be mindful of potential correlations between these variables and to perform further analysis to investigate these relationships.

**Conclusion**

From using ANOVA, we were able to conclude that the age, job, marital status, education, default, housing, contact, month, day of week, duration, campaign, pdays, previous, poutcome, emp.var.rate, cons.price.idx, cons.conf.idx, and euribor3m variables are all significant predictors for our outcome variable. This suggests that the outcome of banking call success cannot be simply reduced to one or two variables that capture most of the variance. As a next step, we used shrinkage methods to attempt

to truncate the number of features. The ridge regression and lasso regressions were both helpful in reducing overfit and helped increase testing accuracy while reducing MSE. From the outputs of the cross-validated Lasso output, we see that the marital status, housing, loan, and cons.price.idx are not as helpful to achieve higher accuracy and lower MSE.

Lastly, and more aggressively, we used model selection from AIC/BIC scores to build modeling in a stepwise manner. The best model that was created from the model selection process was from Backwards BIC, and some remaining features are poutcome, month, pday, default, contact, euribor3m, cons.price.idx, and emp.var.rate. This method may not have produced the most optimal combination of features, but helps us understand the relative importance of the features regardless.

By utilizing these modeling tools, we understand that the prediction of banking marketing call success is embedded in a relatively high dimension, and cannot be simply reduced to a few principle features that help with prediction. However, through various trials and errors, we were able to achieve 91% accuracy with the modeling tools.

A rather large limitation that we encountered is that the outcome variable is binary in nature. Even though we used a binomial family with a logit link function, we are still not perfectly capturing the nature of the underlying data. We would like to explore other techniques such as decision trees in order to make the models more interpretable while being able to handle classification tasks more efficiently. Our results from the model selection part yielded the conclusion that maybe customers in those months with "bad economies" might resort to or stay away from bank savings. An interesting future study that could be inspired by this is to create a calling schedule that takes the individuals' information into account while factoring into the macroeconomic conditions. Perhaps banks can leverage these calling schedules to encourage more people to commit to banking calls.