Classification Using Imaging Data from the 2014 Sloan Digital Sky Survey

This analysis was performed on data found on Kaggle from the Sloan Digital Sky Survey in 2014. The Sloan Digital Sky Survey was founded in 2000 and is a multispectral-imaging and spectroscopic redshift survey that is performed in New Mexico using a 2.5 meter wide-angle optical telescope. The data from the 2014 instance of this survey includes the right ascension (ra) and declination (dec), which specify the direction of a point in the sky using the equatorial coordinate system, the response of the five bands of the telescope using the Thuan-Gunn astronomic magnitude system (u, g, r, i, z), and features describing a field within the image such as run, field number, scanline (camcol), or rerun, which describes how the image was processed. It also includes the class of the object (either star, galaxy, or quasar), the final redshift, which happens when light is increased in wavelength and shifted to the red end of the spectrum, the plate serial number, the Modified Julian Date (mjd), which indicates the date the image was taken, and the fiberID, which corresponds to the optical fibers directing light at the focal plane to the slithead. The recorded observations included mostly galaxies, followed closely by stars, and very few quasars, which are active galactic nuclei with a large black hole. This analysis will aim to compare different methods for modeling and classifying a given object observed by the Sloan Digital Sky Survey based on the 2014 data.

To begin my analysis, I first randomly separated out 3000 of the 10,000 observations to be my test set for evaluating model accuracy. I then checked for missing data and looked at a summary of the data and a correlation matrix to get an idea of the average values of each of the variables and which variables were strongly correlated with one another. I also visualized the total number of each class that were present and found that quasars seemed to be the least common, making up only about 500 of the remaining 7000 observations. Finally, before starting to build my model, I converted the classes to numerical levels with 0 being stars, 1 being galaxies, and 2 being quasars.

Next, I opted to fit a baseline-category logit model because my response variable had three nominal categories. I started with a full model including all available predictors in the data set except for rerun, because the rerun value was the same for every observation and therefore it would not make a difference to include it. Based on this initial model, redshift, z, u, and g all appeared to be potential significant predictors. Next, I checked the remaining predictor variables' significance when used as a sole predictor and found that only camcol and field were not significant at the level alpha = 0.05. I then created my initial main effects model using all of these predictor variables except for camcol and field, which appeared to be insignificant. After that, I conducted backward elimination on the main effects model by minimizing AIC and was left with a model including right ascension (ra), three of the telescope bands (u, g, z), redshift, and plate serial number. I then checked for plausible interactions among the terms in the model, including one three-way interaction among the telescope bands, and found that only the interaction between redshift and plate was very significant. Therefore, my final model included terms ra, u, g, z, redshift, plate, and redshift*plate. After deciding on my final model, I performed the Likelihood-Ratio test by comparing this model to a null model and found that it was in fact significant with a p-value of < 2.2e-16.

Next, I performed the exact same process to construct a binomial logistic regression model with the goal of predicting if an object was either a star or "other" (galaxy or quasar). This model included only one of the same predictors as the BCL model—redshift, it included declination and the interaction term was between declination and redshift rather than redshift and plate. This model also included the scanline or "camera column" (camcol) variable even though that was ruled out in the early stages of model selection for the BCL model. This model was also very significant based on the Likelihood Ratio test with a p-value of < 2.2e-16 as well.

As for accuracy, both models were extremely accurate on the test set and the BCL model had a rate of 98.37 percent accuracy while the binary logistic model had a rate of 99.77 percent. These results are not too surprising because the BCL model was addressing a more complicated prediction task with three classification categories while the binary logistic model was more simplified into just two categories of "star" or "other" and would therefore be easier to predict. Additionally, someone else who analyzed this data on Kaggle used a random forest model as their classifier and achieved 99.05 percent accuracy, which is slightly higher than my BCL model but still pretty close. This could potentially be because with a random forest you can usually just throw all predictor variables into the model and it is not necessary to go through the process of selecting them. Therefore, any accuracy that was lost through removing predictors from my BCL model was not lost in the random forest model.

Based on these results and my two final models, it is clear that redshift is a very strong predictor of the class of an object observed by the Sloan Digital Sky Survey in any case. This is because the redshift tells you how an object is moving in space and how an object moves provides a good indicator for what class it might fall into. It is also known that all galaxies have a positive redshift (rather than a negative redshift, also known as blueshift) and so it makes sense that this would be a highly significant factor. However, it is unclear how the specific telescope bands are related to this classification task and where the distinctions lie among those five different bands. It is also interesting that only either right ascension or declination were significant in the model and never both and it may be worthwhile to investigate why this is the case. For future research, I recommend further investigating the distinctions and relationships between the five telescope bands and how their measurements correlate to the class of an object, how plate serial number and redshift might be related or just how plate serial number is relevant at all to this classification, or the relationships and interactions between right ascension and declination and why only one is needed in the model. It also may be beneficial to apply this research to survey data from other years and compare if there is a difference in observations among them.

Appendix

Summary of the data:

```
##      objid                ra              dec               u
## Min.   :1.238e+18   Min.   :  8.235   Min.   :-5.3826   Min.   :12.99
## 1st Qu.:1.238e+18   1st Qu.:157.808   1st Qu.:-0.5343   1st Qu.:18.16
## Median :1.238e+18   Median :180.239   Median : 0.4101   Median :18.86
## Mean   :1.238e+18   Mean   :175.967   Mean   :14.9678   Mean   :18.61
## 3rd Qu.:1.238e+18   3rd Qu.:201.607   3rd Qu.:37.0608   3rd Qu.:19.26
## Max.   :1.238e+18   Max.   :260.851   Max.   :68.5423   Max.   :19.60
##        g               r               i               z
## Min.   :12.80   Min.   :12.43   Min.   :11.95   Min.   :11.61
## 1st Qu.:16.82   1st Qu.:16.17   1st Qu.:15.85   1st Qu.:15.61
## Median :17.49   Median :16.86   Median :16.55   Median :16.39
## Mean   :17.37   Mean   :16.84   Mean   :16.58   Mean   :16.42
## 3rd Qu.:18.01   3rd Qu.:17.51   3rd Qu.:17.26   3rd Qu.:17.14
## Max.   :19.92   Max.   :24.80   Max.   :24.36   Max.   :20.80
##       run             rerun         camcol          field
## Min.   : 308.0   Min.   :301   Min.   :1.000   Min.   : 11.0
## 1st Qu.: 752.0   1st Qu.:301   1st Qu.:2.000   1st Qu.:186.0
## Median : 756.0   Median :301   Median :4.000   Median :299.0
## Mean   : 980.8   Mean   :301   Mean   :3.665   Mean   :303.3
## 3rd Qu.:1331.0   3rd Qu.:301   3rd Qu.:5.000   3rd Qu.:415.0
## Max.   :1412.0   Max.   :301   Max.   :6.000   Max.   :768.0
##    specobjid             class           redshift             plate
## Min.   :2.996e+17   Min.   :0.0000   Min.   :-0.004136   Min.   : 266
## 1st Qu.:3.390e+17   1st Qu.:0.0000   1st Qu.: 0.000086   1st Qu.: 301
## Median :4.988e+17   Median :1.0000   Median : 0.042591   Median : 443
## Mean   :1.638e+18   Mean   :0.6716   Mean   : 0.149503   Mean   :1455
## 3rd Qu.:2.881e+18   3rd Qu.:1.0000   3rd Qu.: 0.093221   3rd Qu.:2559
## Max.   :9.469e+18   Max.   :2.0000   Max.   : 5.353854   Max.   :8410
##       mjd            fiberid
## Min.   :51608   Min.   :   1.0
```

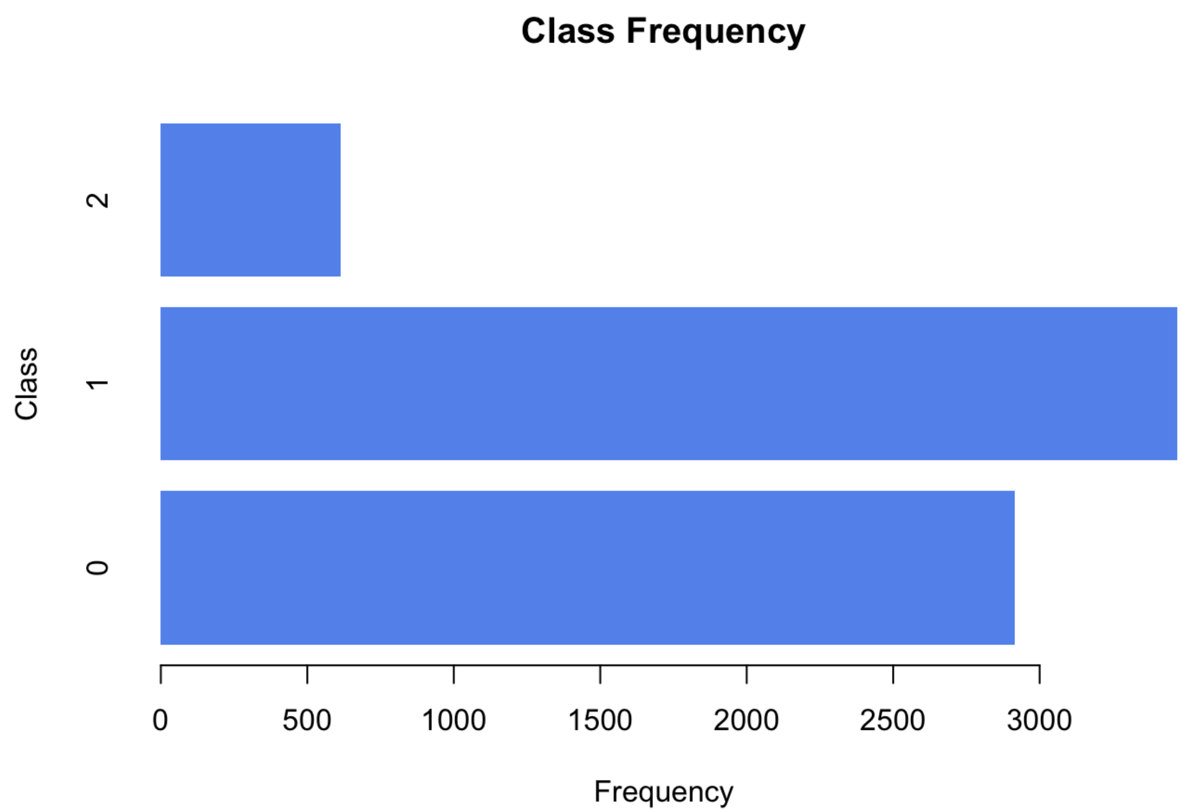
```
##  1st Qu.:51900   1st Qu.: 192.0
##  Median :51997   Median : 353.5
##  Mean   :52942   Mean   : 355.5
##  3rd Qu.:54468   3rd Qu.: 510.0
##  Max.   :57481   Max.   :1000.0
```

Correlation Matrix:

```
##           objid          ra         dec           u           g           r
## objid         1          NA          NA          NA          NA          NA
## ra           NA  1.000000000  0.002366913  0.02790285  0.03844940  0.041094077
## dec          NA  0.002366913  1.000000000  0.04257839  0.06438842  0.064637836
## u            NA  0.027902851  0.042578392  1.00000000  0.85002063  0.694625946
## g            NA  0.038449402  0.064388418  0.85002063  1.00000000  0.959186467
## r            NA  0.041094077  0.064637836  0.69462595  0.95918647  1.000000000
## i            NA  0.039858692  0.058857333  0.61418191  0.91529498  0.982653430
## z            NA  0.036244250  0.055865748  0.55149322  0.87962784  0.969230222
## run          NA -0.092340049  0.782613735  0.04256700  0.05843318  0.056225778
## rerun        NA          NA          NA          NA          NA          NA
## camcol       NA -0.014454846  0.115105412  0.01097718  0.01009961  0.012146259
## field        NA  0.591551510 -0.134848963  0.01095271  0.01695709  0.018964829
## specobjid    NA -0.104867195  0.074628949 -0.13059946 -0.05292960  0.023093301
## class        NA  0.038303770  0.091416682  0.28562879  0.35895318  0.280028726
## redshift     NA  0.027284459  0.061002336  0.16754699  0.41396907  0.449293372
## plate        NA -0.104869315  0.074625077 -0.13060081 -0.05293146  0.023091525
## mjd          NA -0.095649675  0.052108782 -0.17231286 -0.09238723 -0.007461857
## fiberid      NA  0.049604822  0.154115044  0.01504483  0.05162417  0.067202858
##                    i          z         run rerun       camcol        field
## objid             NA         NA          NA    NA           NA           NA
## ra        0.03985869 0.03624425 -0.09234005    NA -0.014454846  0.591551510
## dec       0.05885733 0.05586575  0.78261373    NA  0.115105412 -0.134848963
## u         0.61418191 0.55149322  0.04256700    NA  0.010977176  0.010952710
## g         0.91529498 0.87962784  0.05843318    NA  0.010099610  0.016957089
## r         0.98265343 0.96923022  0.05622578    NA  0.012146259  0.018964829
## i         1.00000000 0.98717984  0.04912128    NA  0.014103698  0.021986054
## z         0.98717984 1.00000000  0.04577971    NA  0.013050198  0.020838928
## run       0.04912128 0.04577971  1.00000000    NA  0.155485275 -0.463181471
## rerun             NA         NA          NA     1           NA           NA
## camcol    0.01410370 0.01305020  0.15548527    NA  1.000000000 -0.079417901
## field     0.02198605 0.02083893 -0.46318147    NA -0.079417901  1.000000000
## specobjid 0.07548865 0.11809508  0.08580900    NA  0.040439003 -0.095985248
## class     0.20880640 0.15236821  0.10280267    NA  0.005523042 -0.010022400
## redshift  0.44139911 0.43273723  0.06137717    NA  0.004560048  0.007823757
## plate     0.07548702 0.11809372  0.08580554    NA  0.040424227 -0.095984459
```

```
## mjd       0.05163136 0.10010955  0.06139328    NA  0.027726776 -0.089627684
## fiberid   0.07476838 0.07580808  0.14270689    NA  0.529060795 -0.051289785
##            specobjid       class     redshift       plate         mjd
## objid              NA          NA          NA          NA          NA
## ra        -0.10486719  0.038303770  0.027284459 -0.10486932 -0.095649675
## dec        0.07462895  0.091416682  0.061002336  0.07462508  0.052108782
## u         -0.13059946  0.285628788  0.167546992 -0.13060081 -0.172312862
## g         -0.05292960  0.358953176  0.413969073 -0.05293146 -0.092387231
## r          0.02309330  0.280028726  0.449293372  0.02309152 -0.007461857
## i          0.07548865  0.208806400  0.441399112  0.07548702  0.051631361
## z          0.11809508  0.152368211  0.432737232  0.11809372  0.100109547
## run        0.08580900  0.102802674  0.061377174  0.08580554  0.061393280
## rerun              NA          NA          NA          NA          NA
## camcol     0.04043900  0.005523042  0.004560048  0.04042423  0.027726776
## field     -0.09598525 -0.010022400  0.007823757 -0.09598446 -0.089627684
## specobjid  1.00000000 -0.469128464 -0.042926136  1.00000000  0.967251844
## class     -0.46912846  1.000000000  0.629063963 -0.46913117 -0.527565486
## redshift  -0.04292614  0.629063963  1.000000000 -0.04292773 -0.061462568
## plate      1.00000000 -0.469131170 -0.042927728  1.00000000  0.967252754
## mjd        0.96725184 -0.527565486 -0.061462568  0.96725275  1.000000000
## fiberid    0.24297848 -0.019023139  0.044813822  0.24295170  0.203109378
##               fiberid
## objid              NA
## ra         0.04960482
## dec        0.15411504
## u          0.01504483
## g          0.05162417
## r          0.06720286
## i          0.07476838
## z          0.07580808
## run        0.14270689
## rerun              NA
## camcol     0.52906080
## field     -0.05128978
## specobjid  0.24297848
## class     -0.01902314
## redshift   0.04481382
## plate      0.24295170
## mjd        0.20310938
## fiberid    1.00000000
```

Class Frequency (0: STAR, 1: GALAXY, 2: QSO)



Initial Main Effects Model (BCL):

```
## Call:
## vglm(formula = class ~ ra + dec + u + g + r + i + z + run + redshift +
##     plate + mjd + fiberid, family = multinomial, data = SDSS)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept):1 -8.550e+01  3.051e+01  -2.803  0.00507 **
## (Intercept):2 -4.276e+01  2.799e+01  -1.527  0.12664
## ra:1           6.222e-03  3.046e-03   2.043  0.04105 *
## ra:2           4.544e-03  2.813e-03   1.616  0.10619
## dec:1         -1.324e-03  9.861e-03  -0.134  0.89315
## dec:2          1.321e-03  8.949e-03   0.148  0.88263
## u:1            5.968e+00  5.064e-01  11.785  < 2e-16 ***
## u:2            5.170e+00  4.428e-01  11.674  < 2e-16 ***
## g:1           -9.203e+00  1.087e+00  -8.468  < 2e-16 ***
## g:2           -6.431e+00  9.241e-01  -6.959 3.41e-12 ***
## r:1           -4.206e-01  9.967e-01  -0.422  0.67302
## r:2           -9.665e-01  8.781e-01  -1.101  0.27102
## i:1            5.338e-01  3.740e-01      NA       NA
```

```
## i:2           2.808e-01  4.014e-01   0.700  0.48416
## z:1           3.714e+00  4.945e-01   7.511 5.86e-14 ***
## z:2           2.142e+00  4.510e-01   4.750 2.03e-06 ***
## run:1        -4.130e-04  9.432e-04  -0.438  0.66149
## run:2         3.415e-04  8.627e-04   0.396  0.69221
## redshift:1   -7.906e+01  3.297e+00 -23.977  < 2e-16 ***
## redshift:2   -7.309e+00  7.881e-01  -9.275  < 2e-16 ***
## plate:1      -4.519e-04  4.980e-04  -0.907  0.36424
## plate:2      -7.051e-04  4.413e-04  -1.598  0.11007
## mjd:1         1.508e-03  5.858e-04      NA       NA
## mjd:2         7.691e-04  5.379e-04   1.430  0.15279
## fiberid:1    -1.684e-03  7.917e-04  -2.127  0.03340 *
## fiberid:2    -1.393e-03  7.223e-04  -1.928  0.05385 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Names of linear predictors: log(mu[,1]/mu[,3]), log(mu[,2]/mu[,3])
##
## Residual deviance: 1397.116 on 13974 degrees of freedom
##
## Log-likelihood: -698.5582 on 13974 degrees of freedom
##
## Number of Fisher scoring iterations: 18
##
## Warning: Hauck-Donner effect detected in the following estimate(s):
## 'u:1', 'u:2', 'g:2', 'i:1', 'z:1', 'redshift:1', 'mjd:1'
##
##
## Reference group is level  3  of the response
```

AIC of initial model:

```
## [1] 1449.116
```

AIC after Backward Elimination:

```
fit6 <- vglm(class ~ ra + u + g + z + run + redshift + plate,
             family = multinomial, data = SDSS)
```

```
## [1] 1378.933
```

Final Model with Interaction:

```
## Call:
## vglm(formula = class ~ ra + u + g + z + run + redshift + plate +
##     redshift * plate, family = multinomial, data = SDSS)
##
## Coefficients:
```

```
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept):1    -1.009e+01  4.662e+00  -2.164 0.030479 *
## (Intercept):2    -4.950e+00  4.365e+00  -1.134 0.256726
## ra:1              5.661e-03  3.711e-03   1.525 0.127143
## ra:2              4.058e-03  3.312e-03   1.226 0.220372
## u:1               6.336e+00  5.404e-01  11.724  < 2e-16 ***
## u:2               6.089e+00  4.958e-01  12.280  < 2e-16 ***
## g:1              -8.478e+00  7.861e-01 -10.785  < 2e-16 ***
## g:2              -8.198e+00  7.328e-01 -11.187  < 2e-16 ***
## z:1               2.900e+00  4.193e-01   6.915 4.67e-12 ***
## z:2               2.389e+00  3.934e-01   6.073 1.25e-09 ***
## run:1            -1.064e-04  6.826e-04  -0.156 0.876153
## run:2             6.609e-04  5.911e-04   1.118 0.263531
## redshift:1       -1.876e+02  9.325e+00 -20.117  < 2e-16 ***
## redshift:2       -1.018e+01  1.126e+00  -9.035  < 2e-16 ***
## plate:1           1.700e-04  1.788e-04   0.950 0.341882
## plate:2          -4.437e-04  1.591e-04  -2.788 0.005303 **
## redshift:plate:1  1.231e-02  2.701e-03   4.556 5.21e-06 ***
## redshift:plate:2  1.014e-03  2.954e-04   3.432 0.000599 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Names of linear predictors: log(mu[,1]/mu[,3]), log(mu[,2]/mu[,3])
##
## Residual deviance: 891.9332 on 13982 degrees of freedom
##
## Log-likelihood: NA on 13982 degrees of freedom
##
## Number of Fisher scoring iterations: 17
##
## Warning: Hauck-Donner effect detected in the following estimate(s):
## 'u:1', 'u:2', 'z:1', 'z:2', 'redshift:1', 'redshift:2'
##
##
## Reference group is level  3  of the response
```

Likelihood Ratio Test:

```
## Likelihood ratio test
##
## Model 1: class ~ ra + u + g + z + run + redshift + plate
## Model 2: class ~ 1
##     #Df  LogLik Df Chisq Pr(>Chisq)
## 1 13984  -673.5
## 2 13998 -6484.3 14 11622  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Likelihood ratio test
```

```
##
## Model 1: class ~ ra + u + g + z + run + redshift + plate + redshift *
##     plate
## Model 2: class ~ ra + u + g + z + run + redshift + plate
##     #Df  LogLik Df Chisq Pr(>Chisq)
## 1 13982
## 2 13984 -673.47  2
```

Binomial Logistic Regression Full Model:

```
## Call:
## glm(formula = class ~ ra + dec + u + g + r + i + z + run + camcol +
##     field + redshift + plate + mjd + fiberid, family = binomial,
##     data = SDSS)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4738  -0.0883   0.0000   0.0000   3.8380
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.356e+01  3.651e+01  -1.193    0.233
## ra          -3.739e-03  6.163e-03  -0.607    0.544
## dec          2.141e-02  1.910e-02   1.121    0.262
## u           -1.995e-01  9.188e-01  -0.217    0.828
## g            1.606e+00  2.226e+00   0.721    0.471
## r           -3.214e+00  2.831e+00  -1.135    0.256
## i            3.059e-01  1.173e+00   0.261    0.794
## z            1.237e+00  1.684e+00   0.735    0.462
## run         -2.083e-04  1.910e-03  -0.109    0.913
## camcol      -1.876e-01  1.388e-01  -1.351    0.177
## field        3.699e-04  2.437e-03   0.152    0.879
## redshift     1.476e+03  2.073e+02   7.120 1.08e-12 ***
## plate       -7.679e-04  6.395e-04  -1.201    0.230
## mjd          8.468e-04  6.883e-04   1.230    0.219
## fiberid     -3.860e-04  1.302e-03  -0.296    0.767
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9506.91  on 6999  degrees of freedom
## Residual deviance:  238.45  on 6985  degrees of freedom
## AIC: 268.45
##
## Number of Fisher Scoring iterations: 18
```

Backward Elimination:

```
##
## Call:  glm(formula = class ~ dec + r + camcol + redshift, family = binomial,
##     data = SDSS)
##
## Coefficients:
## (Intercept)          dec            r       camcol     redshift
##     0.82026      0.02058     -0.33698     -0.22281   1452.90200
##
## Degrees of Freedom: 6999 Total (i.e. Null);  6995 Residual
## Null Deviance:       9507
## Residual Deviance: 241.7     AIC: 251.7
```

Final Model with Interaction:

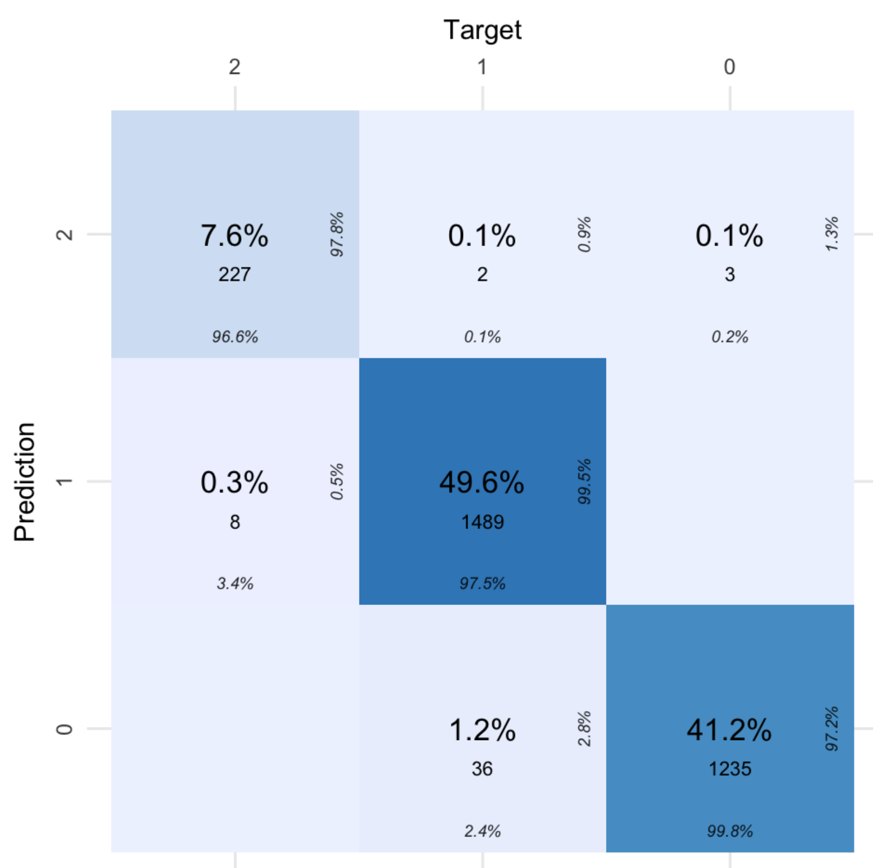
```
## Call:
## glm(formula = class ~ dec + z + camcol + redshift + plate + dec *
##     redshift, family = binomial, data = SDSS)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5109  -0.0883   0.0000   0.0000   3.5783
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.043e+00  2.864e+00  -0.364   0.7158
## dec           1.645e-02  9.944e-03   1.654   0.0981 .
## z            -2.042e-01  1.755e-01  -1.163   0.2446
## camcol       -2.595e-01  1.218e-01  -2.131   0.0331 *
## redshift      1.341e+03  1.712e+02   7.831 4.82e-15 ***
## plate        -7.709e-05  1.323e-04  -0.583   0.5600
## dec:redshift  3.360e+01  1.444e+01   2.327   0.0200 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9506.91  on 6999  degrees of freedom
## Residual deviance:  235.12  on 6993  degrees of freedom
## AIC: 249.12
##
## Number of Fisher Scoring iterations: 19
```

Likelihood Ratio Test:

```
# check the significance of the final model against a null model
binom.fit0 <- glm(class ~ 1, family = binomial, data = SDSS)
anova(binom.final, binom.fit0, test = "LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: class ~ dec + z + camcol + redshift + plate + dec * redshift
## Model 2: class ~ 1
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      6993      235.1
## 2      6999     9506.9 -6  -9271.8 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

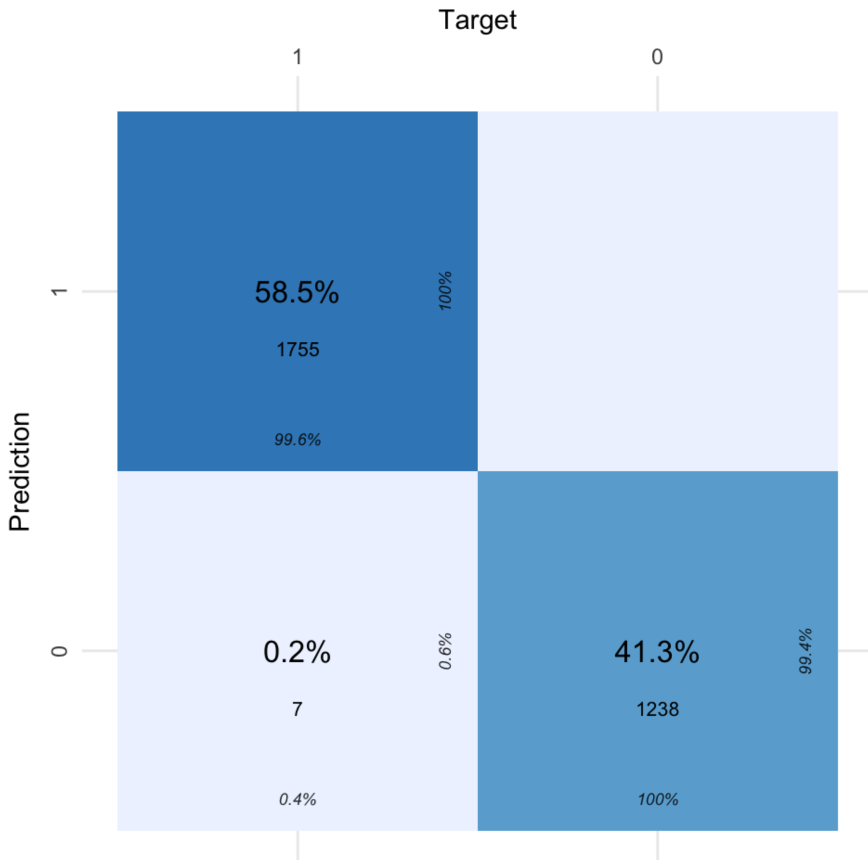
BCL Model Accuracy:




```
## Confusion Matrix and Statistics
##
##           Reference
```

```
## Prediction    0    1    2
##          0 1235    0    3
##          1   36 1489    2
##          2    0    8  227
##
## Overall Statistics
##
##                Accuracy : 0.9837
##                  95% CI : (0.9785, 0.9879)
##     No Information Rate : 0.499
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9711
##
##  Mcnemar's Test P-Value : 2.992e-09
##
## Statistics by Class:
##
##                      Class: 0 Class: 1 Class: 2
## Sensitivity            0.9717   0.9947  0.97845
## Specificity            0.9983   0.9747  0.99711
## Pos Pred Value         0.9976   0.9751  0.96596
## Neg Pred Value         0.9796   0.9946  0.99819
## Prevalence             0.4237   0.4990  0.07733
## Detection Rate         0.4117   0.4963  0.07567
## Detection Prevalence   0.4127   0.5090  0.07833
## Balanced Accuracy      0.9850   0.9847  0.98778
```

Binary Logistic Regression Model Accuracy:




```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1238    0
##          1    7 1755
##
##                Accuracy : 0.9977
##                  95% CI : (0.9952, 0.9991)
##     No Information Rate : 0.585
##     P-Value [Acc > NIR] : < 2e-16
##
##                   Kappa : 0.9952
##
##  Mcnemar's Test P-Value : 0.02334
##
##             Sensitivity : 0.9944
##             Specificity : 1.0000
##          Pos Pred Value : 1.0000
##          Neg Pred Value : 0.9960
##              Prevalence : 0.4150
##          Detection Rate : 0.4127
```

```
##    Detection Prevalence : 0.4127
##       Balanced Accuracy : 0.9972
##
##        'Positive' Class : 0
##
```