Assignment 1

Matthew Adair

2/22/2018

Contents

Project Introduction

The intent of this project was to see how well I could predict whether an admittd student for a university would attend the school or not. The dataset was admitted student data and it originally contained 54 variables and 2,947 records. In the original dataset there were 29 qualitative variables and 25 quantitative variables. In my final cleaned dataset, I used only 24 variables, eight of which are quantitative and the rest qualitative while being able to retain all 2,947 records. The response varaible for this project was the "Decision" variable, which is whether or not the admitted student decided to attend the university or not.

Cleaning and Prep

For the cleaning process I began with deleting the variable "Entry Term (Application)" because all of the entries are for the year of 2017, so this variable will not help with analysis as well as deleting the variable "Admit Type" because all of the applicants are looking to be first years. Furthermore, I also deleted "Permanent Postal" and decided to use other location variables such as "PrmntCountry" and "PermntGeom," which I thought would provide the right amount of macroview data and more granular view. The "Permanent Postal" variable I thought would be too specific and cause problems with splitting test and training sets and with the logistic regression model as a whole. For the "PermntCountry" variable I boiled down the levels to either inside the U.S. or Out, since all of the countires outside the United States accounted for approximately 8% of applicants. For the variable "Ethnicity," I transformed it into a binary factored variable with 0 representing students who identified Non-Hispanic/Latino and 1 representing Hispanic/Latino. I also for this dataset filled in the 89 NA values with 0 since 80% of the students identified themselves as Non-Hispanic/Latino. I first tried to randomly fill in the NAs proportionally with a function but I could not get a function to work and therefore filled the NAs with the values of 0.

With the "Race" variable I refactored the levels to be "White", "African American", "Asian" and "Other." I did this because I wanted to retain as many levels as possible while combining the levels that had a small amount of observations and put them all under the name of "Other" so that they could be used for analysis. For the religion variable I created five levels instead of the ten plus levels originally found in the data. I did this in order to avoid some of the religions that had only a small handful of records and thus could skew data analysis by grossly misrepresnting the affect of those beliefs on the analysis. The levels are "Protestant", "Catholic", "Other", "None", and "UnId." I deleted the variable "First_Source Origin First Source Date" because I did not think that hearing about a school early on in high school would be as impactful as "DecisionPlan" or "Inquired" and be more or less not very beneficial overall. I changed the variable "Inquiry Date" to "Inquired" and also converted the dates into either "Yes," if they did inquire, or

"No," if they did not inquire. I did this because I the dates were not that important to me, it was more the fact that the student inquired about something that I wanted to retain. I also deleted the variable "Submitted" because the variable "DecisionPlan" essentially expresses the exact same idea and I believe that the decison plan carries more wieht because by choosing early action or early decision, the student is likely showing more interest than the student who applied on regular decision.

Furthermore, I deleted the variable "Application Source" because it does not affect the student's decision to come to the university or not. It may impact applying to the university but it likely does not impact the decision. I then converted the variable "Legacy" into a factor and created levels 0 (for not Legacy) and 1 (having legacy). I did this because I really only cared about how legacy overall would impact a student going to TU or not. Additionally, there are not enough observations for each type of legacy to be able to provide accurate information. Deleted "First_Source Origin First Source Summary" due to the fact that it just relates how the students found out about university. For the variable "Athlete" I releveled it to have "Athlete" (represented by 1), "Athlete, Opt Out" (Represented by 2), "Other" (represented by 3), and "Not Athlete" (represented by a 0). I did this so that I could retain some sort of granular view of this variable. With the variable "Sport1" I renamed it to "Sport" and refactored and relabeled the options to the sport, i.e. "Basketball", and got rid of "Basketball Men" and "Basketball Women." I did this because I wanted to retain as many levels as possible of this variable while combining the levels that had only of handful of records to make the variable "Sport" useable. I deleted the variables "Sport 2" through "Sport 3 Rating" because they are redundant, and most likely, if the student is going to play a sport, they are going to play their first sport.

With the variables "Academic Interest 1" I relabeled the levels to departments that the university has and I renamed the variable "AcdmcInt1". NAs for "AcdmcInt1" I made Undecided because I thought that by doing so I wouldn't affect the analysis negatively much becuase it is plausible that they do not know what their interests are and since Undecided was the level that appeared the least amount of times I thought that by adding those to NA would not hurt analysis. For the variable "School 1 Class Rank" I will fill all the NA's with "0" (meaning not provided). I am doing this because I think that since it is a strategy for applying for schools, much like providing class rank and GPA is if it makes the student stand out from the crowd, it is essentially just naming another method for applying (while retaining the rest of the record). I will do the same for the variable "School 1 Class Size." I renamed both variables, one being "ClassRank" and the other being "ClassSize" to shorten the name. I Deleted the variable "School 1 GPA" because the variable "School 1 GPA Recalculated" is more insightful and it does not seem to make too much sense to keep relatively redundant variables. I renamed the variable "School 1 GPA Recalculated" to "GPARecalc." I also deleted the variable "School 1 GPA Scale" because knowing the scale is not necessary when there is a variable that has the GPA score recalculated. I deleted the variables "School #2 Organization Category" and "School #3 Organization Category" because both of those variables only contained null values. Also, I deleted the variables "ACT English", "ACT Reading", "ACT Math", "ACT Science Reasoning", "ACT Writing", "SAT I CR+M", "SAT I Critical Reading", "SAT I Math", "SAT I Writing", "SAT R Evidence-Based Reading and Writing Section + Math Section", "SAT R Evidence-Based Reading and Writing Section", "SAT R Math Section", and "SAT Concordance" because I thought they were redundant since I was going to be useing ACTComp for my variable of testing how standardized college entry exams affect students deciding on what school they are going to go to.

With the variable "Decision", I converted everything into 0, meaning not accepted, and 1, meaning accepted. I did this because all we care about is whether a student decided to attend the university or not and not what specific type of decision a student made. For the variable "Intend to Apply for Financial Aid" I filled 2/3 of NA's with the 1 and 1/3 0 to retain the ratio of 1 to 0. Additionally, I renamed the variable to "FinancialAid." Furthermore, I deleted variable "Staff Assigned Name" because this variable will provide more noise than weight due to the fact that the staff is assigned over a region and thus can't really measure the effect of each staff member because the staff do not rotate regions as far as I know. For the "Permanent Geomarket" I split his variable into two columns, one with state or international and the other containing the region code. I deleted the variable that had just the region # because I saw that some of the regions had just a handful of data of data points and thus not provide a lot of accuracy. I then renamed it to "PrmntGeom" and created 6 levels which are INT, W, MW, SW, SE, NE, to avoid having to delete rows that only have one observation (which will create problems in making the test and training set). Made US into SouthWest. Also, with the

"MeritAward" variable, I removed all of the letters just to retain the numbers because I found out that the numbers represent the amount of scholarship money and that particular component of the data is all that I need. Especially when the specific naming system is causing difficulties with splitting test and training set. Lastly, I changed the the variable "AcmcInd"" to an ordered factor to accurately reflect what data conveys and to be able to use it appropriately.

Modeling

For the modeling process I first created testing and training sets that are reproducible to use for creating and testing my baseline logistic model. I used for sample size 7/10 and I did this because I was having trouble with new levels coming up in the testing set that were not in the training set and so I decided to try to include all the levels possible for the training set so that I would not have this issue. The baseline logistic model with every variable being regressed against the y variable "Decision", created an accuracy of .8393 with an AIC of 1594.6.

I then ran feature selection based off of AIC and found the optimal variables to use.

```
# Splitting into training and test dataset
smpl <- floor(.7*nrow(AdmDat))
set.seed(123)
train_ind <- sample(seq_len(nrow(AdmDat)), size = smpl)
trn <- AdmDat[train_ind, ]
tst <- AdmDat[-train_ind, ]

# Feature Selection
modl <- glm(formula = Decision ~ .-ID, data=AdmDat)
summary(modl)
step <- stepAIC(modl, direction="backward", na.action=na.remove)
step$anova
```

The result of the feature selection lead me to the model with the best AIC of 1539 and the highest accuracy of .8467.

```
smpl <- floor(.7*nrow(AdmDat)) # Making a sample with half of the data
set.seed(123) # set.seed to make reproducible work
train_ind <- sample(seq_len(nrow(AdmDat)), size = smpl) # random sample half amount of rows in AdmDat
trn <- AdmDat[train_ind, ] # Create training dataset
tst <- AdmDat[-train_ind, ] # Create testing dataset
lgm5 <- glm(Decision ~ PermntCountry + Ethnicity + Religion + Inquired +
              DecisionPlan + Legacy + Athlete + SportRating + EventParticipation +
              CampusVisits + ClassRank + ACTComp + AcdmcIndex + MeritAward,
            family = binomial(link = 'logit'), data = trn) # AIC = 1539
summary(lgm5)
```

```
##
## Call:
## glm(formula = Decision ~ PermntCountry + Ethnicity + Religion +
##     Inquired + DecisionPlan + Legacy + Athlete + SportRating +
##     EventParticipation + CampusVisits + ClassRank + ACTComp +
##     AcdmcIndex + MeritAward, family = binomial(link = "logit"),
##     data = trn)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4264  -0.5321  -0.3468  -0.1382   3.1735
```

```
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     6.093e-01  1.076e+00   0.566  0.57108
## PermntCountryUnited States     -6.730e-01  2.846e-01  -2.365  0.01804 *
## Ethnicity                       2.686e-01  1.673e-01   1.606  0.10832
## ReligionNone                   -1.672e+00  5.527e-01  -3.025  0.00249 **
## ReligionOther                  -1.329e-01  2.930e-01  -0.454  0.65017
## ReligionPrtsnt                  2.480e-01  2.002e-01   1.239  0.21550
## ReligionUnId                    4.781e-01  1.897e-01   2.521  0.01171 *
## InquiredYes                     3.192e-01  1.509e-01   2.116  0.03436 *
## DecisionPlanEarly Action II     2.040e-01  1.543e-01   1.322  0.18603
## DecisionPlanEarly Decision I    4.047e+00  8.227e-01   4.919 8.69e-07 ***
## DecisionPlanEarly Decision II   1.448e+01  3.044e+02   0.048  0.96206
## DecisionPlanRegular Decision    3.492e-01  1.977e-01   1.767  0.07730 .
## Legacy                          7.346e-01  2.123e-01   3.460  0.00054 ***
## Athlete                        -9.007e-01  2.281e-01  -3.949 7.85e-05 ***
## SportRatingFranchise           -4.879e-01  4.274e-01  -1.142  0.25357
## SportRatingNone                -2.087e+00  4.127e-01  -5.058 4.23e-07 ***
## SportRatingVarsity             -5.324e-01  3.269e-01  -1.629  0.10337
## EventParticipation              2.114e+00  1.301e-01  16.247  < 2e-16 ***
## CampusVisits                    5.576e-01  5.193e-02  10.737  < 2e-16 ***
## ClassRank                       1.353e-03  9.260e-04   1.461  0.14400
## ACTComp                        -7.716e-02  3.054e-02  -2.527  0.01152 *
## AcdmcIndex                      3.805e-01  7.108e-02   5.352 8.68e-08 ***
## MeritAward                      5.361e-06  1.100e-05   0.487  0.62604
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2218.5  on 2061  degrees of freedom
## Residual deviance: 1506.6  on 2039  degrees of freedom
## AIC: 1552.6
##
## Number of Fisher Scoring iterations: 13
```

```
tic()
fit.results5 <- predict(lgm5,newdata=tst,type='response')
fit.results5 <- ifelse(fit.results5 > .5, 1, 0)
accuracy5 <- mean(as.matrix(fit.results5) == as.matrix(tst$Decision),na.rm=TRUE)
print(paste('Accuracy =',accuracy5))
```

```
## [1] "Accuracy = 0.846153846153846"
```

```
toc()
```

```
## 0.012 sec elapsed
```

I then tested the model with a confusion matrix and found the results to be ok. With a strong sensitivity but a weak specificity. This model can of course be greatly improved but overall is an ok model.

```
confusionMatrix(as.factor(fit.results5), as.factor(tst$Decision), positive = NULL)
```

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

```
## Prediction   0   1
##          0 664 101
##          1  35  84
##
##                Accuracy : 0.8462
##                  95% CI : (0.8207, 0.8693)
##     No Information Rate : 0.7907
##     P-Value [Acc > NIR] : 1.667e-05
##
##                   Kappa : 0.465
##  Mcnemar's Test P-Value : 2.494e-08
##
##             Sensitivity : 0.9499
##             Specificity : 0.4541
##          Pos Pred Value : 0.8680
##          Neg Pred Value : 0.7059
##              Prevalence : 0.7907
##          Detection Rate : 0.7511
##    Detection Prevalence : 0.8654
##       Balanced Accuracy : 0.7020
##
##        'Positive' Class : 0
##
```