Lab 2

James Beck, Colby Carter, Andrew Lam

10/23/2017

1. Introduction

In this report, our team seeks to predict the contribution dollar bucket in which our sample alumni preside based on potential explanatory variables such as gender, class year, area of degree and study, and recent university contributions and participation. In order to maximize this predictive power, we test two sets of variables in multinomial logistic regressions, as well as consider the usefulness of a similar ordered logistic regression. Because we are skeptical of there being proportional odds between the given contribution buckets, we see history of prior giving using the multinomial model to be most predictive, though with its own limitations.

2. Data Examination and EDA

The provided dataset contains 1,000 observations of sampled university graduates spanning five decades of classes, spanning from 1972 to 2012. Included in the dataset are an ID number X, gender, marital status, an indicator of whether the alumnus attended a recent university event, his or her major and advanced degree (if applicable), and the annual individual donation amounts from 2012 to 2016; there are no missing values in any of these cells, nor any contribution amounts that would suggest entry errors.

Univariate Analysis & Transformations

```
data <- read.csv("lab2data.csv")
str(data)
```

```
## 'data.frame':    1000 obs. of  12 variables:
##  $ X              : int  761 620 214 373 748 1080 1155 1069 1161 457 ...
##  $ Gender         : Factor w/ 2 levels "F","M": 1 2 1 1 2 1 1 1 1 1 ...
##  $ Class.Year     : int  2002 2002 1982 1992 2002 2012 2012 2012 2012 1992 ...
##  $ Marital.Status : Factor w/ 4 levels "D","M","S","W": 2 3 2 2 3 3 3 3 3 2 ...
##  $ Major          : Factor w/ 45 levels "American Studies",..: 39 25 25 2 30 2 3 26 39 15 ...
##  $ Next.Degree    : Factor w/ 47 levels "AA","BA","BAE",..: 37 39 39 35 39 15 39 35 39 18 ...
##  $ AttendenceEvent: int  1 0 1 1 0 1 0 1 0 0 ...
##  $ FY12Giving     : num  50 0 100 0 0 0 0 5 0 0 ...
##  $ FY13Giving     : num  51 0 0 0 0 0 0 10 0 75 ...
##  $ FY14Giving     : num  51 0 100 0 0 0 0 25 0 0 ...
##  $ FY15Giving     : num  0 0 100 0 0 0 0 25 0 0 ...
##  $ FY16Giving     : num  0 0 100 0 0 0 0 50 0 60 ...
```

```
head(data, 5)
```

```
##     X Gender Class.Year Marital.Status        Major Next.Degree
## 1 761      F       2002              M    Sociology         MSW
## 2 620      M       2002              S      History        NONE
## 3 214      F       1982              M      History        NONE
## 4 373      F       1992              M Anthropology          MS
```

```
## 5 748      M       2002              S   Philosophy        NONE
##   AttendenceEvent FY12Giving FY13Giving FY14Giving FY15Giving FY16Giving
## 1               1         50         51         51          0          0
## 2               0          0          0          0          0          0
## 3               1        100          0        100        100        100
## 4               1          0          0          0          0          0
## 5               0          0          0          0          0          0
```

```
data$Class.Year <- factor(data$Class.Year)
data$AttendenceEvent <- factor(data$AttendenceEvent, levels = c(0,
    1), labels = c("Not Attend", "Attend"))
data$Gender <- factor(data$Gender, levels = c("M", "F"), labels = c("Male",
    "Female"))
data$Marital.Status <- factor(data$Marital.Status, levels = c("M",
    "S", "D", "W"), labels = c("Married", "Single", "Divorced",
    "Widowed"))
data$FY16GivingCat <- cut(data$FY16Giving, breaks = c(0, 1, 100,
    250, 500, 20000), right = F, labels = c("[0,1)", "[1,100)",
    "[100,250)", "[250,500)", "[500,20000)"))
```

Beginning with the demographic and alumni type factor variables, we see a sample of 1,000 graduates who are approximately evenly split by gender and mostly married, with a class size that increases for each sample year (decade). We also see that a majority of the sample attended the recent alumni event, indicating support for the university post-graduation.

```
describe(data[c("Gender", "Class.Year", "Marital.Status", "AttendenceEvent")])
```

```
## data[c("Gender", "Class.Year", "Marital.Status", "AttendenceEvent")]
##
##  4  Variables      1000  Observations
## ---------------------------------------------------------------------------
## Gender
##        n  missing distinct
##     1000        0        2
##
## Value        Male Female
## Frequency     495    505
## Proportion  0.495  0.505
## ---------------------------------------------------------------------------
## Class.Year
##        n  missing distinct
##     1000        0        5
##
## Value       1972  1982  1992  2002  2012
## Frequency    105   176   203   223   293
## Proportion 0.105 0.176 0.203 0.223 0.293
## ---------------------------------------------------------------------------
## Marital.Status
##        n  missing distinct
##     1000        0        4
##
## Value       Married   Single Divorced  Widowed
## Frequency       584      344       61       11
## Proportion    0.584    0.344    0.061    0.011
## ---------------------------------------------------------------------------
```

```
## AttendenceEvent
##        n  missing distinct
##     1000        0        2
##
## Value      Not Attend     Attend
## Frequency         395        605
## Proportion      0.395      0.605
## ---------------------------------------------------------------------------
```

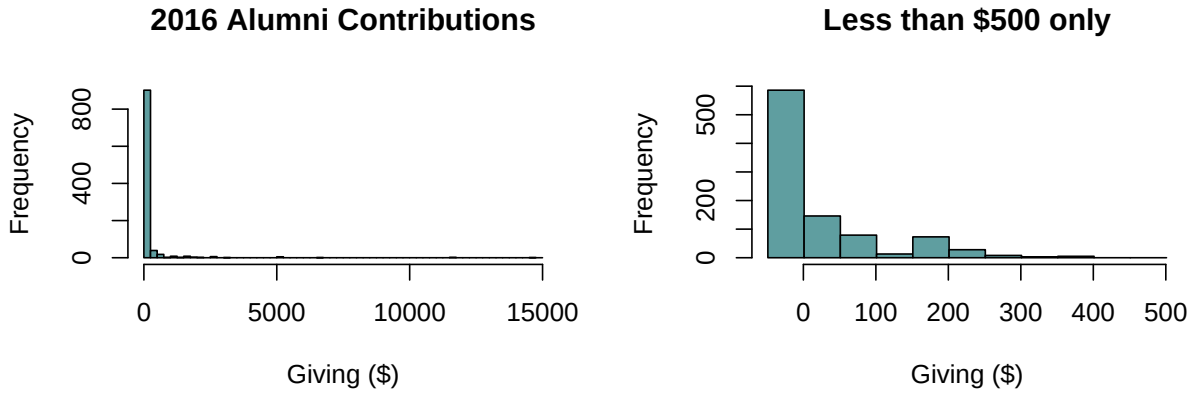
```
data$Marital.Status <- relevel(data$Marital.Status, ref = "Married")
```

By generating some key descriptive statistics and a histogram of the contributions, we can garner some important insights into the distribution of the variable that we will be attempting to model, *FY16Giving*. Most notably, the shape of the histogram indicates a distribution that is heavily skewed toward a contribution of 0. From this we can see the benefit of modeling the contribution variable as a set of contribution tiers rather than as a continuous variable.

```
describe(data$FY16Giving)
```

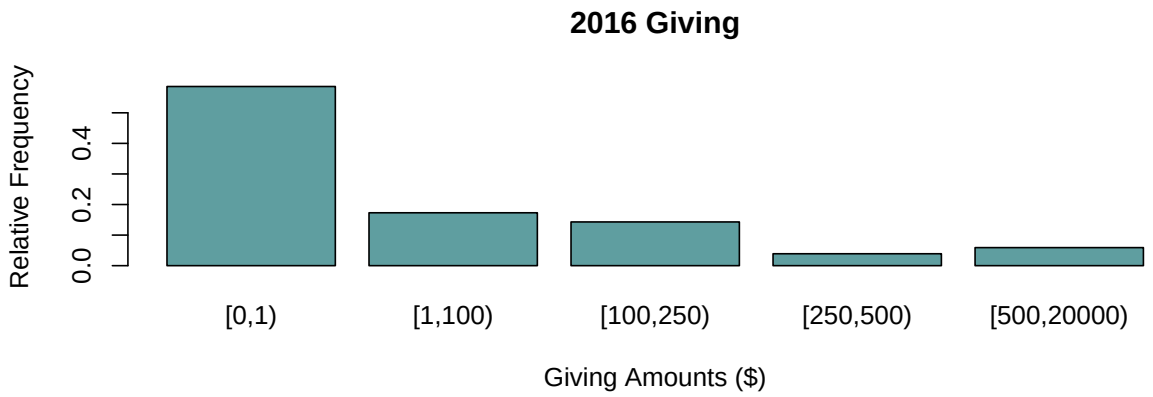
```
## data$FY16Giving
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1000        0       71    0.798      170    308.2        0        0
##      .25      .50      .75      .90      .95
##        0        0       75      216      500
##
## lowest :     0.00     5.00    10.00    15.00    18.00
## highest:  5000.00  6500.00 11500.00 11505.84 14655.25
```

```
par(mfrow = c(1, 2))
hist(data$FY16Giving, breaks = c(250 * 0:60 - 1), main = "2016 Alumni Contributions",
    col = "cadetblue", xlab = "Giving ($)")
hist(data$FY16Giving[data$FY16Giving < 500], breaks = c(50 *
    0:11 - 49), main = "Less than $500 only", col = "cadetblue",
    xlab = "Giving ($)")
```



By bucketing the data into the categories shown in the barplot below, it can be seen that the data is still very much skewed towards the non-contributor category, but if our ultimate goal is to be able to build a model that can help us differentiate between alumni that will and will not donate, then this categorization of the dependent variable will serve that purpose better.

```
barplot(prop.table(table(data$FY16GivingCat)), main = "2016 Giving",
    xlab = "Giving Amounts ($)", ylab = "Relative Frequency",
    col = "cadetblue")
```

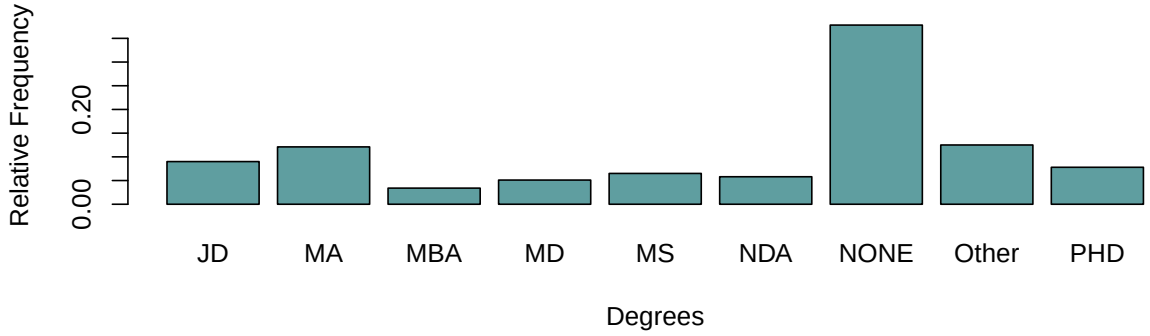


We opt to consolidate our *degree* buckets categorization somewhat to collapse the data into a smaller number of manageable categories. Most notably we recategorize degrees with a very small representation in the data as 'Other' and maintain distinction between the more majorly represented degrees (e.g. MA, MS, MBA, PHD). In doing so we can see from the barplot below that the 'None' category actually represents a large chunk of the total data, but otherwise there is fairly even distribution across the categories.

```
degrees <- data.frame(table(data$Next.Degree))
colnames(degrees) <- c("degree", "freq")
degrees[which(degrees$freq > 30), ]
```

```
##    degree freq
## 15     JD   90
## 18     MA  108
## 23    MBA   34
## 25     MD   42
## 35     MS   53
## 38    NDA   58
## 39   NONE  378
## 40    PHD   78
```

```
degree_buckets <- function(deg) {
    if (deg %in% c("JD", "PHD", "NONE", "MBA", "NDA")) {
        deg
    } else if (substr(deg, 1, 2) == "MA") {
        "MA"
    } else if (substr(deg, 1, 2) == "MS") {
        "MS"
    } else if (substr(deg, 1, 2) == "MD") {
        "MD"
    } else {
        "Other"
    }
}
data$degree_bucket <- factor(apply(data["Next.Degree"], 1, degree_buckets))
barplot(prop.table(table(data$degree_bucket)), xlab = "Degrees",
    ylab = "Relative Frequency", col = "cadetblue")
```




```
data$degree_bucket <- relevel(data$degree_bucket, ref = "NONE")
```

A similar technique is applied to the *major* variable in the data set to consolidate categories.

```
majors <- data.frame(table(data$Major))
colnames(majors) <- c("major", "freq")
majors[which(majors$freq > 30), ]
```

```
##                major freq
## 2       Anthropology   35
## 3                Art   32
## 4            Biology   88
## 5          Chemistry   53
## 10         Economics   78
## 15           English  101
## 25           History  102
## 27       Mathematics   32
## 35 Political Science   61
## 36        Psychology   65
## 39         Sociology   49
```

```
major_buckets <- function(maj) {
    maj <- tolower(maj)
    if (grepl("economics", maj)) {
        "Economics"
    } else if (maj %in% c("chinese", "french", "german", "russian",
        "spanish")) {
        "Foreign Language"
    } else if (grepl("history", maj) | maj == "american studies") {
        "History"
    } else if (maj == "english") {
        "English"
    } else if (maj %in% c("biology", "chemistry", "physics", "math") |
        grepl("general science", maj)) {
        "Math & Science"
    } else if (grepl("sociolog", maj) | grepl("psych", maj) | grepl("anthro",
        maj)) {
        "Social Science"
```

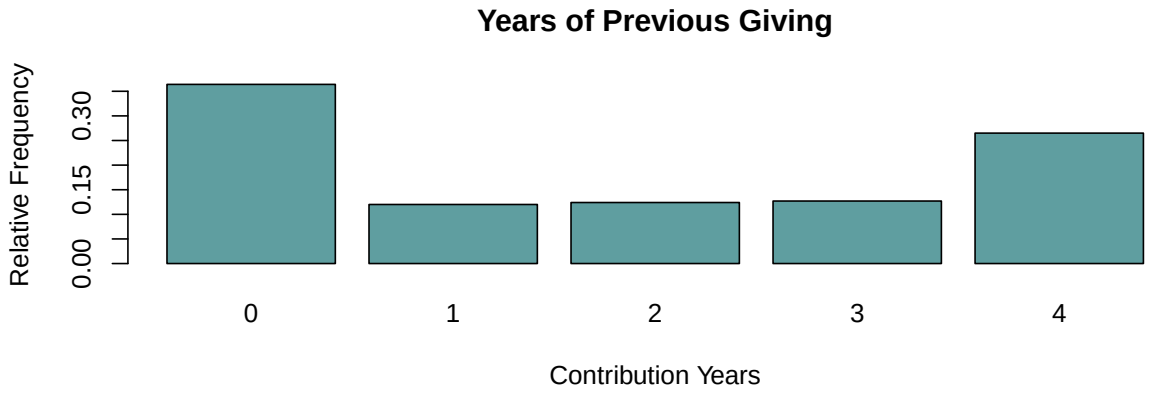
```
    } else if (grepl("political sci", maj) | grepl("pol. sci",
        maj)) {
        "Political Science"
    } else if (maj %in% c("art", "music", "theatre")) {
        "Arts & Music"
    } else if (grepl("phil", maj) | grepl("relig", maj)) {
        "Philosophy & Religion"
    } else {
        "Other"
    }
}
data$major_bucket <- factor(apply(data["Major"], 1, major_buckets))
table(data$major_bucket)
```

```
##
##          Arts & Music             Economics               English
##                    76                   102                   101
##      Foreign Language               History       Math & Science
##                    68                   125                   173
##                 Other Philosophy & Religion     Political Science
##                    86                    46                    62
##        Social Science
##                   161
```

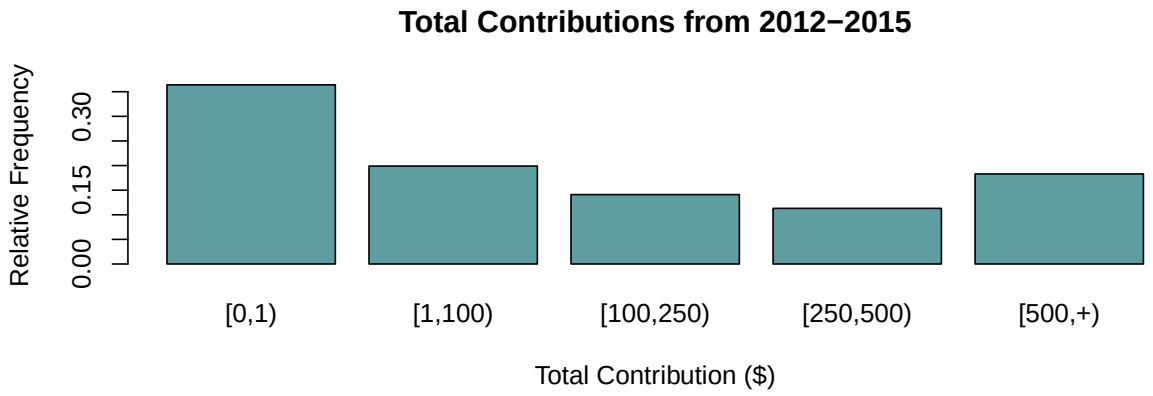
```
data$major_bucket <- relevel(data$major_bucket, ref = "Math & Science")
```

In an effort to summarize information from the previous years, we construct a new variable that describes how many years a particular contributor has donated. The major bucket is 0 years - this is not seen as a surprise given our previous observation that the majority of potential contributors did not donate any money in 2016. What is interesting to note is that the second largest category is the 4 years category. This implies that those who give tend to be consistent in contributing.

```
pre16_giving <- c("FY12Giving", "FY13Giving", "FY14Giving", "FY15Giving")
data$pre16_give_cnt <- factor(apply(data[pre16_giving] != 0,
    1, sum))
barplot(prop.table(table(data$pre16_give_cnt)), xlab = "Contribution Years",
    ylab = "Relative Frequency", main = "Years of Previous Giving",
    col = "cadetblue")
```



```
data$pre16_giving <- rowSums(data[, pre16_giving])
data$pre16_bin <- cut(data$pre16_giving, breaks = c(0, 1, 100,
    250, 500, 2e+06), right = F, labels = c("[0,1)", "[1,100)",
    "[100,250)", "[250,500)", "[500,+)"))
barplot(prop.table(table(data$pre16_bin)), xlab = "Total Contribution ($)",
    col = "cadetblue", ylab = "Relative Frequency", main = "Total Contributions from 2012-2015")
```



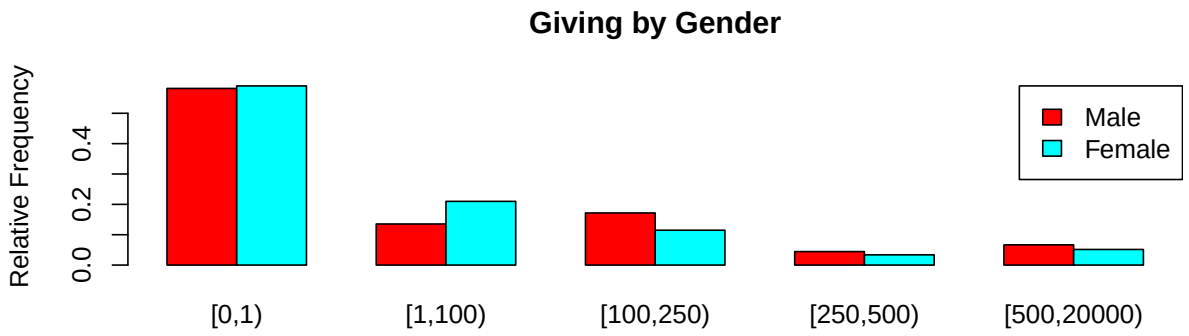
Bivariate Relationships with Dependent Variable

Examining the relationship between gender and contribution category we can see that the breakdown of representation in the different contribution categories is fairly similar across genders. There is a somewhat notable difference between the representation in the [1,100) and the [100,250) categories in that a greater percentage of females are present in the [1,100) category while there are a greater percentage of males in the [100,250) category.

```
table(data$Gender, data$FY16GivingCat)
```

```
##
##          [0,1) [1,100) [100,250) [250,500) [500,20000)
##   Male     288      67        85        22          33
##   Female   298     106        58        17          26
```

```
barplot(prop.table(table(data$Gender, data$FY16GivingCat), 1),
    col = rainbow(length(levels(data$Gender))), beside = TRUE,
    main = "Giving by Gender", ylab = "Relative Frequency")
legend("topright", legend = levels(data$Gender), fill = rainbow(length(levels(data$Gender))))
```

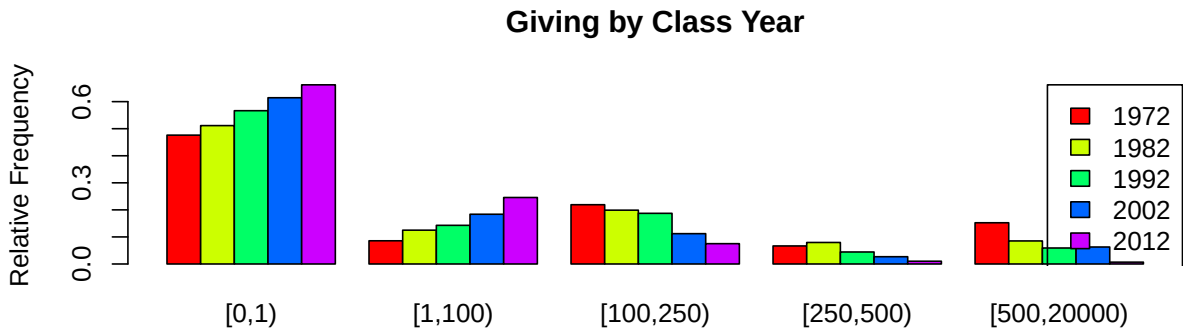


Examining the relationship between class year and our categorical contribution variable offers some interesting insights as well. Namely the older the class year seemingly the more likely the person is to donate. This is not terribly surprising given that more recent graduates would not be expected to have the sort of discretionary income that would allow them to donate to their alma mater. If the 2012 class does donate, they tend to stick to the [1,100) category. Again, this is likely due to having less to donate. The other class years instead tend to contribute in the higher tiers.

```
table(data$Class.Year, data$FY16GivingCat, dnn = c("Class Year",
    "2016 Giving Bin"))
```

```
##            2016 Giving Bin
## Class Year [0,1) [1,100) [100,250) [250,500) [500,20000)
##       1972    50       9        23         7          16
##       1982    90      22        35        14          15
##       1992   115      29        38         9          12
##       2002   137      41        25         6          14
##       2012   194      72        22         3           2
```

```
barplot(prop.table(table(data$Class.Year, data$FY16GivingCat),
    1), col = rainbow(length(levels(data$Class.Year))), beside = TRUE,
    main = "Giving by Class Year", ylab = "Relative Frequency")
legend("topright", legend = levels(data$Class.Year), fill = rainbow(length(levels(data$Class.Year))))
```



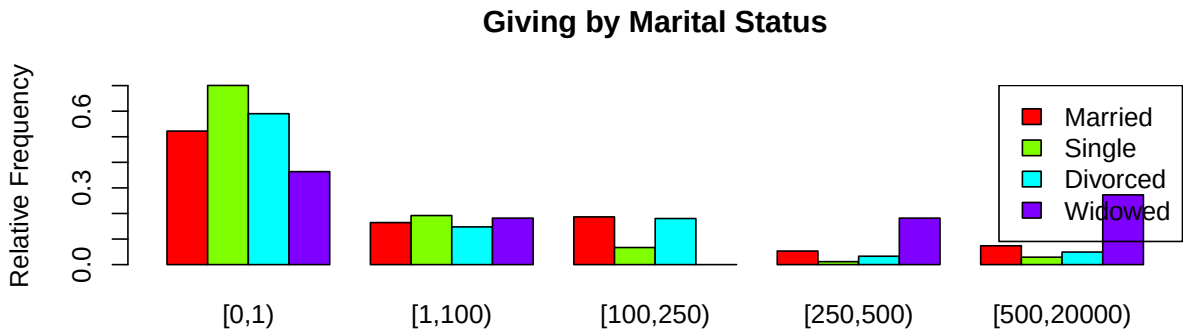
Marital status has a very interesting breakdown as shown in the count table and barplot below. Married and Single alumni heavily dominate the population as shown in the table, but the bar plot gives a little more

insight into how the marital status might help indicate whether an alum has donated or not. One thing we note is that within the Single category there are many more alum in the non-contributor category relative to the other marital statuses. We imagine that this could be a result of a factor discussed in our exploration of class year. Younger alum are also more likely to be single and would potentially not contribute for the same reason: lack of discretionary income. Widowed alum exhibit a very interesting breakdown in this visualization as well. Their overall representation with the alum population is quite small, but widowed alum are far more likely to contribute significantly than the other marital categories. We expect that, again, this could simply be related to age. Widows are more likely to be older and in turn more likely to have discretionary income to donate.

```
table(data$Marital.Status, data$FY16GivingCat)
```

```
##
##            [0,1) [1,100) [100,250) [250,500) [500,20000)
##   Married    305      96        109        31          43
##   Single     241      66         23         4          10
##   Divorced    36       9         11         2           3
##   Widowed      4       2          0         2           3
```

```
barplot(prop.table(table(data$Marital.Status, data$FY16GivingCat),
    1), col = rainbow(length(levels(data$Marital.Status))), beside = TRUE,
    main = "Giving by Marital Status", ylab = "Relative Frequency")
legend("topright", legend = levels(data$Marital.Status), fill = rainbow(length(levels(data$Marital.Statu
```

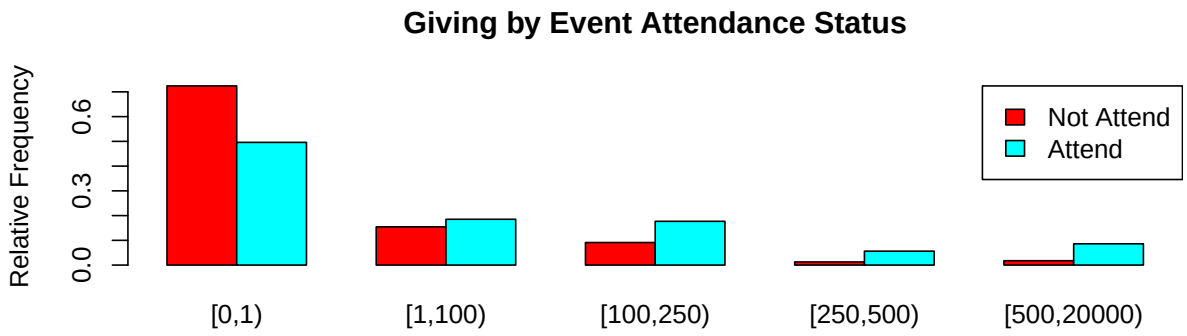


Event attendance is a fairly unsurprising categorical variable. As seen in the chart below, the proportion of alumni that donate if they have previously attended an event is notably greater than the proportion of alumni that dontate if they have not previously attended an event. This observation points to the inclusion of the event attendance variable in our eventual modeling.

```
table(data$AttendenceEvent, data$FY16GivingCat)
```

```
##
##              [0,1) [1,100) [100,250) [250,500) [500,20000)
##   Not Attend   286      61        36         5           7
##   Attend       300     112       107        34          52
```

```
barplot(prop.table(table(data$AttendenceEvent, data$FY16GivingCat),
    1), col = rainbow(length(levels(data$AttendenceEvent))),
    beside = TRUE, main = "Giving by Event Attendance Status",
    ylab = "Relative Frequency")
legend("topright", legend = levels(data$AttendenceEvent), fill = rainbow(length(levels(data$AttendenceE
```



Considering our created major buckets, we might expect different areas of study–corresponding to different income or philanthropic tendencies–to affect amount of giving back to the university. We see some possible evidence that the arts, english and economics correspond to less giving than other majors, but these differences do not appear to be large.

```
round(prop.table(table(data$major_bucket, data$FY16GivingCat),
    1), 3)
```

```
##
##                        [0,1) [1,100) [100,250) [250,500) [500,20000)
##   Math & Science        0.590   0.191     0.145     0.017       0.058
##   Arts & Music          0.684   0.132     0.158     0.013       0.013
##   Economics             0.647   0.088     0.147     0.029       0.088
##   English               0.644   0.198     0.099     0.010       0.050
##   Foreign Language      0.574   0.176     0.132     0.103       0.015
##   History               0.512   0.152     0.184     0.048       0.104
##   Other                 0.500   0.186     0.140     0.116       0.058
##   Philosophy & Religion 0.696   0.130     0.174     0.000       0.000
##   Political Science     0.548   0.177     0.145     0.065       0.065
##   Social Science        0.553   0.230     0.124     0.025       0.068
```

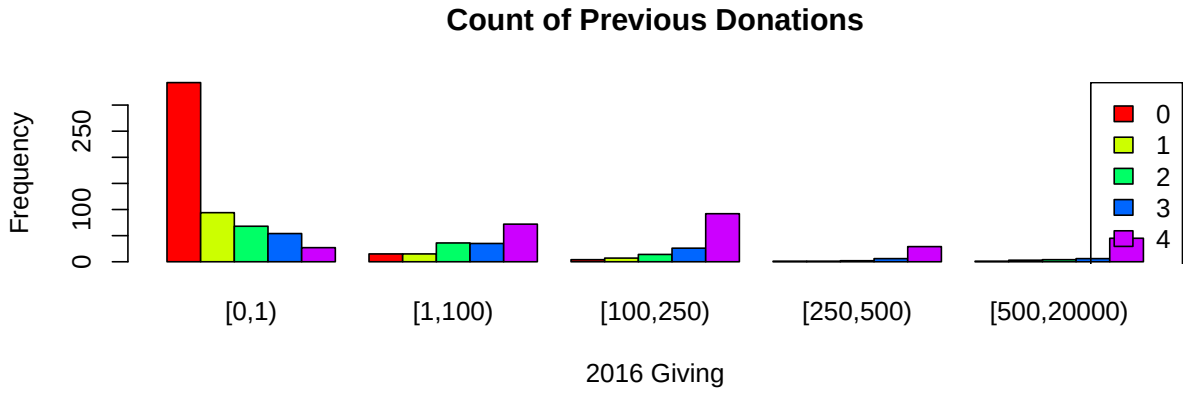
When looking at advanced degree, on the other hand, we see larger differences in giving versus non-giving bins. In particular, those with "NONE" appear least likely to contribute while those alumni with masters of science ("MS"), in law ("JD"), or with PhDs to be most likely contributors, and correspondingly, likely the wealthiest graduates.

```
round(prop.table(table(data$degree_bucket, data$FY16GivingCat),
    1), 3)
```

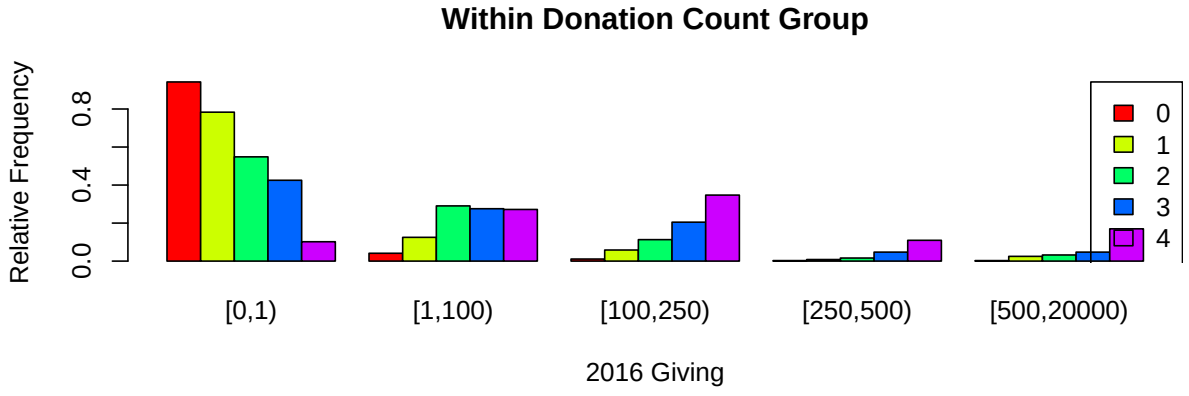
```
##
##         [0,1) [1,100) [100,250) [250,500) [500,20000)
##   NONE  0.720   0.108     0.108     0.040       0.024
##   JD    0.478   0.111     0.189     0.067       0.156
##   MA    0.504   0.273     0.140     0.033       0.050
##   MBA   0.529   0.088     0.206     0.059       0.118
##   MD    0.588   0.176     0.137     0.000       0.098
##   MS    0.385   0.277     0.215     0.046       0.077
##   NDA   0.603   0.276     0.086     0.017       0.017
##   Other 0.512   0.232     0.152     0.032       0.072
##   PHD   0.487   0.179     0.205     0.051       0.077
```

Meanwhile, it stands to reason that previous giving can be a strong predictor of future giving, and looking at the frequency of the count of annual donations between 2012 and 2015 confirms this: those who did not give at all previously also make up the largest segment of those who did not contribute in 2016. Similarly, those who previously gave all four years appear more likely to fall in one of the giving buckets.

```
barplot(table(data$pre16_give_cnt, data$FY16GivingCat), xlab = "2016 Giving",
    col = rainbow(length(levels(data$pre16_give_cnt))), beside = TRUE,
    main = "Count of Previous Donations", ylab = "Frequency")
legend("topright", legend = levels(data$pre16_give_cnt), fill = rainbow(length(levels(data$pre16_give_cn
```




```
barplot(prop.table(table(data$pre16_give_cnt, data$FY16GivingCat),
    1), col = rainbow(length(levels(data$pre16_give_cnt))), beside = TRUE,
    main = "Within Donation Count Group", ylab = "Relative Frequency",
    xlab = "2016 Giving")
legend("topright", legend = levels(data$pre16_give_cnt), fill = rainbow(length(levels(data$pre16_give_cn
```



Bivariate Relationships between Independent Variables

We consider the distribution of genders amongst different degrees and majors in an effort to better understand both variables. In particular we would like to look for anything that may inadvertently create a collinearity concern in our modelling.

As an example, we can see that the proportion of married frequency decreases with increasing class year.

This will be something to keep in mind as we consider both variables in our models as finding significance in marital status as a predictor could be intertwined with the effect of class year.

```
round(prop.table(table(data$Class.Year, data$Marital.Status),
    1), 3)
```

```
##
##        Married Single Divorced Widowed
##   1972   0.771  0.057    0.114   0.057
##   1982   0.693  0.136    0.153   0.017
##   1992   0.685  0.227    0.084   0.005
##   2002   0.619  0.368    0.009   0.004
##   2012   0.355  0.635    0.010   0.000
```

Similarly, choice of major and degree tends to vary widely by gender, with males dominating econometrics degrees and MBAs, and females having majorities in English, language, social science and non-business masters degrees.

```
round(prop.table(table(data$major_bucket, data$Gender), 1), 3)
```

```
##
##                          Male Female
##   Math & Science        0.578  0.422
##   Arts & Music          0.461  0.539
##   Economics             0.824  0.176
##   English               0.386  0.614
##   Foreign Language      0.221  0.779
##   History               0.528  0.472
##   Other                 0.535  0.465
##   Philosophy & Religion 0.565  0.435
##   Political Science     0.516  0.484
##   Social Science        0.323  0.677
```

```
round(prop.table(table(data$degree_bucket, data$Gender), 1),
    3)
```

```
##
##          Male Female
##   NONE  0.497  0.503
##   JD    0.567  0.433
##   MA    0.405  0.595
##   MBA   0.706  0.294
##   MD    0.490  0.510
##   MS    0.385  0.615
##   NDA   0.552  0.448
##   Other 0.408  0.592
##   PHD   0.641  0.359
```

Interaction Effects

We consider several potential interaction effects, as well as the possibility that certain factor variables are not independent of each other, such as major and gender. We might expect event attendence to wane over time with contribution likelihood to be higher for earlier class years; however, actual differences appear to be mixed, for example with 2002 alumni more likely to contribute when they *didn't* attend.

```
round(prop.table(table(data$Class.Year, data$FY16GivingCat, data$AttendenceEvent),
    1), 3)
```

```
## , ,  = Not Attend
##
##
##        [0,1) [1,100) [100,250) [250,500) [500,20000)
##   1972 0.267   0.048     0.057     0.000       0.019
##   1982 0.290   0.068     0.085     0.017       0.011
##   1992 0.286   0.069     0.059     0.005       0.005
##   2002 0.242   0.036     0.000     0.004       0.009
##   2012 0.324   0.075     0.010     0.000       0.000
##
## , ,  = Attend
##
##
##        [0,1) [1,100) [100,250) [250,500) [500,20000)
##   1972 0.210   0.038     0.162     0.067       0.133
##   1982 0.222   0.057     0.114     0.062       0.074
##   1992 0.281   0.074     0.128     0.039       0.054
##   2002 0.372   0.148     0.112     0.022       0.054
##   2012 0.338   0.171     0.065     0.010       0.007
```

It's also possible that the genders contribute in different proportions depending on their marital status. But looking at the contingency table by contribution bin, there do not appear to be large diferences by gender.

```
round(prop.table(table(data$Marital.Status, data$FY16GivingCat,
    data$Gender), 1), 3)
```

```
## , ,  = Male
##
##
##             [0,1) [1,100) [100,250) [250,500) [500,20000)
##   Married   0.269   0.062     0.113     0.027       0.046
##   Single    0.334   0.084     0.044     0.009       0.012
##   Divorced  0.246   0.033     0.066     0.033       0.016
##   Widowed   0.091   0.000     0.000     0.091       0.091
##
## , ,  = Female
##
##
##             [0,1) [1,100) [100,250) [250,500) [500,20000)
##   Married   0.253   0.103     0.074     0.026       0.027
##   Single    0.366   0.108     0.023     0.003       0.017
##   Divorced  0.344   0.115     0.115     0.000       0.033
##   Widowed   0.273   0.182     0.000     0.091       0.182
```

3. Statistical Modeling

Based on the exploratory data analysis, we conjecture that class year, marital status, event attendence, factors representing prior contributions, and to a lesser extent degree and major will all reflect different likelihoods to contribute. Namely, each reflects some combination of earnings potential, university enthusiasm, and prior levels of generosity. We do not expect gender to directly have an effect, given that it shows little difference in

its contingency with 2016 and is likely reflected with highly correlated variables such as degree and major. We expect earlier classes to be more likely to contribute, along with majors and degrees aligned with high earnings, e.g. MBA and PhDs. We similarly expect more in contributions from married individuals, and especially from those who attend alumni events and have given to the university previously.

Because our team is tasked with predicting which 2016 giving bin in which the alumni most likely reside, we estimate a multinomial logistic regression to maximize the accuracy. We see each giving bin as distinct from each other, and particularly from the non-giving event, so estimating coefficients for each sub-model should produce the highest accuracy with the given data. We also estimate an ordered logistic regression in order to compare consistency of the statistical significance of our selected variables.

Initial Multinomial Logistic Regression Model

Beginning with our full hypothesized model, we split our data into training and test datasets, with 800 observations (80%) used for model estimation and the remaining 200 for testing. This model's estimates converge after 90 iterations, with an initial AIC of 1,286.7 on the training sample. Checking the covariates for statistical significance, the log likelihood ratio test produces test statistics which are extremely significant for pre-2016 contribution count and amount (binned), with major showing borderline significance for this training sample. The above confusion matrix shows the model's predicted 2016 giving compared to the actual contributions made by our test sample. We note this initial error rate is approximately 34%, with 17 of those sample observations predicting no contribution when they actually gave between 1 and 100.

```
# split train and test
set.seed(437)
sample <- sample.int(n = nrow(data), size = floor(0.8 * nrow(data)),
    replace = F)
train <- data[sample, ]
test <- data[-sample, ]

# full trained model
ml_full <- multinom(formula = FY16GivingCat ~ Class.Year + Marital.Status +
    degree_bucket + major_bucket + AttendenceEvent + pre16_give_cnt +
    pre16_bin, data = train)
```

```
## # weights:  175 (136 variable)
## initial  value 1287.550330
## iter  10 value 610.315773
## iter  20 value 518.859177
## iter  30 value 511.876810
## iter  40 value 511.447471
## iter  50 value 511.424634
## iter  60 value 511.394989
## iter  70 value 511.371362
## iter  80 value 511.364281
## iter  90 value 511.363298
## final  value 511.363241
## converged
```

```
ml_full$call
```

```
## multinom(formula = FY16GivingCat ~ Class.Year + Marital.Status +
##     degree_bucket + major_bucket + AttendenceEvent + pre16_give_cnt +
##     pre16_bin, data = train)
```

```r
print("Full model AIC:", quote = F)
```

```
## [1] Full model AIC:
```

```r
ml_full$AIC
```

```
## [1] 1286.726
```

```r
Anova(ml_full)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: FY16GivingCat
##                  LR Chisq Df Pr(>Chisq)
## Class.Year         24.438 16  0.0803558 .
## Marital.Status     19.131 12  0.0854273 .
## degree_bucket      33.336 32  0.4021025
## major_bucket       46.120 36  0.1203525
## AttendenceEvent    21.527  4  0.0002489 ***
## pre16_give_cnt    115.929 16  < 2.2e-16 ***
## pre16_bin         204.790 16  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
preds_full <- predict(ml_full, newdata = test)
probs_full <- round(predict(ml_full, newdata = test, type = "probs"),
    4)
test_full <- data.frame(test, preds_full, probs_full)
print("Confusion Matrix:", quote = F)
```

```
## [1] Confusion Matrix:
```

```r
table(test_full$preds_full, test_full$FY16GivingCat, dnn = c("Predicted",
    "Actual"))
```

```
##               Actual
## Predicted      [0,1) [1,100) [100,250) [250,500) [500,20000)
##   [0,1)          101      17         6         1           1
##   [1,100)          4      17         4         0           0
##   [100,250)        4       4         9         2           5
##   [250,500)        2       0         4         3           3
##   [500,20000)      2       0         6         3           2
```

```r
print("Error Rate:", quote = F)
```

```
## [1] Error Rate:
```

```r
mean(as.character(test_full$preds_full) != as.character(test_full$FY16GivingCat),
    na.rm = T)
```

```
## [1] 0.34
```

Final Multinomial Model

To improve the model's predictive power, we test whether the in-sample fit is improved when removing class year, marital status, and degree, and then performing a LRT test on class year to see whether it improves the model.

```
ml_restr <- multinom(formula = FY16GivingCat ~ AttendenceEvent +
    pre16_give_cnt + pre16_bin, data = train)
```

```
## # weights:  55 (40 variable)
## initial  value 1287.550330
## iter  10 value 630.467239
## iter  20 value 567.310074
## iter  30 value 566.211020
## iter  40 value 566.164459
## iter  50 value 566.161742
## iter  50 value 566.161739
## iter  50 value 566.161739
## final  value 566.161739
## converged
```

```
ml_restr$call
```

```
## multinom(formula = FY16GivingCat ~ AttendenceEvent + pre16_give_cnt +
##     pre16_bin, data = train)
```

```
print("Restricted model AIC:", quote = F)
```

```
## [1] Restricted model AIC:
```

```
ml_restr$AIC
```

```
## [1] 1204.323
```

```
Anova(ml_restr)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: FY16GivingCat
##                 LR Chisq Df Pr(>Chisq)
## AttendenceEvent   19.147  4  0.0007354 ***
## pre16_give_cnt   129.879 16  < 2.2e-16 ***
## pre16_bin        259.410 16  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Under the restricted model, the AIC falls to 1,204.3, with event attendance still significant by the LRT. We now compare this model with the inclusion of class year, but again find a non-significant Chi test statistic (p=0.44).

```
# confirm class year does not help in-sample fit
ml_class <- multinom(formula = FY16GivingCat ~ AttendenceEvent +
    pre16_give_cnt + pre16_bin + Class.Year, data = train)
```

```
## # weights:  75 (56 variable)
## initial  value 1287.550330
## iter  10 value 639.624386
## iter  20 value 563.049188
## iter  30 value 558.489180
## iter  40 value 558.107778
## final  value 558.102678
## converged
```

```
print("Restricted model AIC with major:", quote = F)
```

```
## [1] Restricted model AIC with major:
```

```
ml_class$AIC
```

```
## [1] 1220.205
```

```
anova(ml_class, ml_restr)
```

```
## Likelihood ratio tests of Multinomial Models
##
## Response: FY16GivingCat
##                                                          Model Resid. df
## 1              AttendenceEvent + pre16_give_cnt + pre16_bin      3164
## 2 AttendenceEvent + pre16_give_cnt + pre16_bin + Class.Year      3148
##   Resid. Dev   Test    Df LR stat.   Pr(Chi)
## 1   1132.323
## 2   1116.205 1 vs 2    16 16.11812 0.4447476
```

We again test the model's predictive power on the test set, for which we see a very marginal improvement to 32.5%. This final model is both the simplest along with the smallest residual deviance and the lowest test error.

```
# predictive power
preds_restr <- predict(ml_restr, newdata = test)
probs_restr <- round(predict(ml_restr, newdata = test, type = "probs"),
    4)
test_restr <- data.frame(test, preds_restr, probs_restr)
table(test_restr$preds_restr, test_restr$FY16GivingCat, dnn = c("Predicted",
    "Actual"))
```

```
##              Actual
## Predicted     [0,1) [1,100) [100,250) [250,500) [500,20000)
##   [0,1)         103      18         8         2           1
##   [1,100)         2      14         1         0           0
##   [100,250)       3       5        10         2           2
##   [250,500)       0       0         0         0           0
##   [500,20000)     5       1        10         5           8
```

```
mean(as.character(test_restr$preds_restr) != as.character(test_restr$FY16GivingCat),
    na.rm = T)
```

```
## [1] 0.325
```

Looking at the coefficients to each sub-model, we see mostly an intuitive story with a couple of unexpected relationship. In particular, the odds of contributing in the [250,500) or $500+ buckets go up by roughly 6 and 15 times, respectively, when someone attends the alumni event; those odds do not change drastically for those lower amounts under $250. Somewhat inversely, when someone has given for the previous four years, their odds of contributing a non-zero amount less than 500 increases by roughly 25 times, whereas the odds go up by roughly 1.9 and 2.4 for the larger contribution amount buckets. In terms of the total amount an individual has given previously, people appear to tend toward their historical amounts. For example, if that amount totals less than 100, then that person is roughly 12 times more likely to give again in that lower range. Similarly, those who have given more than $500 total are extremely likely to do so again, according to the model. However, we remain cautious of these odds considering the model still has a high error rate on the test data.

```
round(exp(coefficients(ml_restr)), 3)
```

```
##             (Intercept) AttendenceEventAttend pre16_give_cnt1
## [1,100)           0.047                 0.886           0.229
## [100,250)         0.012                 1.524           1.441
## [250,500)         0.001                14.829           0.357
## [500,20000)       0.001                 5.976           0.808
##             pre16_give_cnt2 pre16_give_cnt3 pre16_give_cnt4
## [1,100)               1.440           2.766          25.585
## [100,250)             3.075           4.882          27.104
## [250,500)             0.248           0.410           1.945
## [500,20000)           0.636           0.475           2.362
##             pre16_bin[1,100) pre16_bin[100,250) pre16_bin[250,500)
## [1,100)               11.902              6.991              1.391
## [100,250)              1.498              4.400              7.788
## [250,500)              0.004              0.003             29.674
## [500,20000)            2.882              0.000              5.787
##             pre16_bin[500,+)
## [1,100)               0.202
## [100,250)            11.423
## [250,500)           238.912
## [500,20000)         357.814
```

Using Proportional Odds Regression Model (Ordered Logistic)

Lastly, for an additional test of robustness, we run our full model as an ordered logistic, which assumes proportional changes in odds as we compare giving bins. To our surprise, we see class year with strong statistical significance under the LRT, where we did not see significance before. We do, however, similarly get strong significance for total prior giving (binned); we lose strong significance in the previous giving count. We approach this model with caution as the choice of bins are different enough in range that they may not satisfy the assumption of proportional odds in this model, making our estimates unreliable.

```
# proportional odds logistic model on full dataset
ordered_lm <- clm(formula = FY16GivingCat ~ Class.Year + Marital.Status +
    degree_bucket + major_bucket + AttendenceEvent + pre16_give_cnt +
    pre16_bin, data = train, link = "logit")
ordered_lm$call
```

```
## clm(formula = FY16GivingCat ~ Class.Year + Marital.Status + degree_bucket +
##     major_bucket + AttendenceEvent + pre16_give_cnt + pre16_bin,
##     data = train, link = "logit")
```

```
ordered_lm$info
```

```
##    link threshold nobs  logLik     AIC niter max.grad  cond.H
## 1 logit  flexible  800 -644.31 1360.61  6(0) 5.53e-11 4.2e+02
```

```
Anova(ordered_lm)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: FY16GivingCat
##                 Df    Chisq Pr(>Chisq)
## Class.Year       4 373.8288     <2e-16 ***
## Marital.Status   3   0.6555     0.8836
```

```
## degree_bucket    8   6.6554     0.5742
## major_bucket     9   8.8295     0.4532
## AttendenceEvent  1   0.2839     0.5942
## pre16_give_cnt   4   7.6261     0.1063
## pre16_bin        4 296.1082     <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4. Final Remarks

While the model produces a fairly high error rate when predicting the five giving categories, it does have a number of notable strengths. Namely, the model is fairly simple and easy to interpret. It relies solely on information regarding prior donations and event attendance. The significance of these factors should be unsurprising to the administration - these findings are logically consistent with the notion that those who have given or have been previously engaged with the university (via event attendance) in the past are likely to contribute in the future. One inherent weakness in this conclusion is that it doesn't offer much insight into what turns a non-contributor into one that does. Because the model relies significantly on whether a person has donated in the past, it doesn't offer much perspective on which people to engage that have not donated in the past.

This model is far from perfect. As seen in the model's corresponding confusion matrix, the model incorrectly predicts the contribution category 32.5% of the time. This is significant and warrants further investigation. Importantly, however, the administration's primary concern should be with differentiation between potential contributors and non-contributors. Looking again to the confusion matrix we can see that the error rate in that regard is closer to 20%.

The most mistaken contributor vs. non-contributor category within this confusion matrix are alumni who are predicted to be non-contributors and are actually contributors in the [1,100) category. It is unfortunate to miss out on engaging these potential donors, but fortunately the dollar amount that would go unrealized is a little less significant given that these donors were not major contributors.

The two major dimensions to our model are previous donations (amount and frequency) and event attendance. As a result, it may be useful to further explore the nature of this event to understand why it has such importance in the model. Does attendance indicate a propensity to engage with the university and be more apt to be involved with the university's future via donations? Or is there something about the event itself that convinces participants to donate or be more engaged in the university? If the truth is in the latter explanation, there could be additional information worth collecting to better understand what contributes to getting potential donors to attend the event (e.g. proximity to the event location). Further, we would want to see what events may have taken place on campus or initiatives announced over the previous four years that may have spurred greater giving from certain individuals, e.g. the announcement of a new planned science building or athletic arena. These future events could then be used in a targeted communications effort to reach those most likely to contribute based on that particular cause (e.g. former science students or season ticket holders, respectively).