Predicting University Faculty Wikipedia Use Behavior with Logistic Regression, LDA, and QDA

STAT 508 | Individual Midterm Project

David Chen

March 24th, 2020

Introduction

In this report, we will be attempting to predict high vs low use of Wikipedia among university faculty members by conducting a principal component analysis, and then applying a logistic regression, a linear discriminant analysis (LDA), and a quadratic discriminant analysis (QDA). The data comes a 2012-2013 survey of 913 faculty members from two Spanish universities, containing general demographic information (age, position, gender, etc.), as well as likert type data regarding their perceptions and thoughts regarding Wikipedia. From our analysis, we determine that the logistic regression model is the best moving forward.

Data

For this analysis, we will be analyzing the `wiki4HE` dataset. As listed from the online documentation, this dataset consists of a 2012-2013 online survey of 913 faculty members from two Spanish universities: Universitat Oberta de Catalunya (UOC) and Universitat Pompeu Fabra (UPF). This survey was conducted to analyze university faculty member use and perceptions of Wikipedia, predominately in the teaching context. While detailed descriptions of all the variables can be found in the documentation linked above, we will list a general overview of the data here.

General demographic attributes:

1. `AGE` - Age
2. `GENDER` - Gender
3. `DOMAIN` - Which domain (Arts & Humanities, Sciences, etc) do they work in
4. `PhD` - Do they posses a PhD
5. `YEARSEXP` - Years of university teaching experience
6. `UNIVERSITY` - UOC or UPF
7. `UOC_POSITION` - If they work at UOC, what is their position title (Professor, Associate, etc.)
8. `OTHER` - Do they have a main job at another university if they are part-time
9. `OTHER_POSITION` - Position if they work part-time at another university and are UPF members
10. `USERWIKI` - Are they a Wikipedia registered user

Likert scale (1-5) data categories:

Note that (#) indicates the total number of unique questions for each category.

1. Perceived Usefulness (3)
2. Perceived Ease of Use (3)
3. Perceived Enjoyment (2)
4. Quality (5)
5. Visibility (3)
6. Social Image (3)
7. Sharing attitude (3)
8. Use behavior (5)
9. Profile 2.0 (3)
10. Job relevance (2)
11. Behavioral intention (2)
12. Incentives (4)
13. Experience (5)

EDA

Overall Distribution

First, we check for the overall distribution of some of the variables. We first note that the age has a fairly wide range, from 23 to 69 years old. We also note that range of experience, as the mean is relatively small.

```
##       AGE           YEARSEXP   
##  Min.   :23.0   Min.   : 0.0  
##  1st Qu.:36.0   1st Qu.: 5.0  
##  Median :42.0   Median :10.0  
##  Mean   :42.2   Mean   :10.9  
##  3rd Qu.:47.0   3rd Qu.:15.0  
##  Max.   :69.0   Max.   :43.0  
##                 NA's   :23    
```

We also note the total number of participants from each university. It appears that UOC is represented far more than UPF, and so future sampling must take this into consideration.

UNIVERSITY	Count
UOC	800
UPF	113

The domains also tells a similar story. Given that the domains are not equally distributed, we will want to ensure that future training and testing sets account for this difference.

DOMAIN	Count
1	183
2	56
3	73
4	137
5	101
6	361
NA	2

Missing Values

Next, we check for the number of missing values for each variable. As observed from the table below, several of the variables have a large number of missing values. In particular, we note that OTHERSTATUS, OTHER_POSITION, UOC_POSITION have the most, as expected since not all faculty members work part-time at another university. We also note that some of the likert predictors, such as PEU3, VIS1, VIS2, IM3 have fairly large numbers of missing values as well.

Since UOC_POSITION and OTHERSTATUS are typically mutually exclusive from each other (unless a faculty member works in both), we will combine the two together into POSITION. If a faculty member does work at both universities, only their UOC position will be considered.

	NA Values
AGE	0
GENDER	0

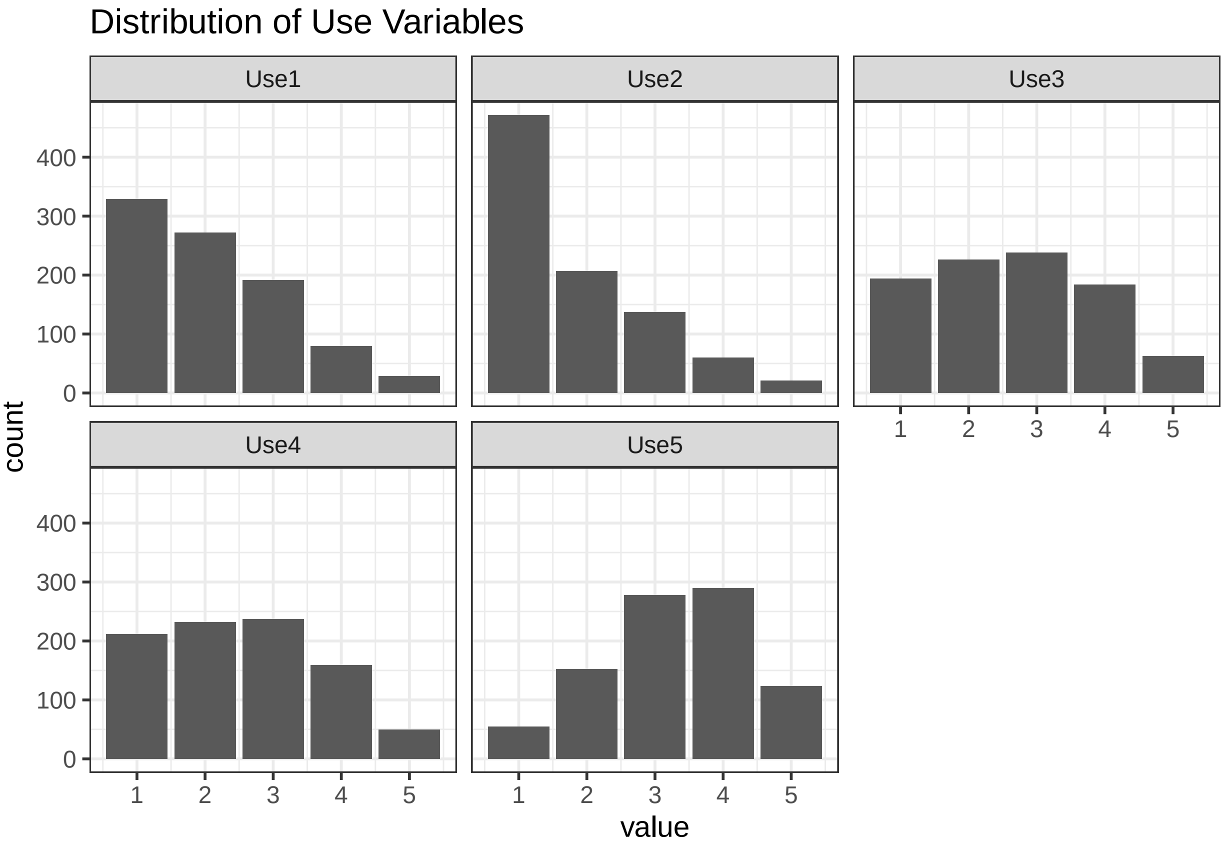
	NA Values
DOMAIN	2
PhD	0
YEARSEXP	23
UNIVERSITY	0
UOC_POSITION	113
OTHER_POSITION	261
OTHERSTATUS	540
USERWIKI	4
PU1	7
PU2	11
PU3	5
PEU1	4
PEU2	14
PEU3	97
ENJ1	7
ENJ2	17
Qu1	7
Qu2	10
Qu3	15
Qu4	22
Qu5	29
Vis1	72
Vis2	117
Vis3	8
Im1	22
Im2	20
Im3	57
SA1	11
SA2	12
SA3	11
Use1	14
Use2	17
Use3	9
Use4	23
Use5	15
Pf1	11
Pf2	6
Pf3	14
JR1	27
JR2	53
BI1	32
BI2	43
Inc1	35
Inc2	35
Inc3	37
Inc4	42
Exp1	13
Exp2	11
Exp3	13
Exp4	14
Exp5	13

Use Behavior Variables

As mentioned previously, the goal of this analysis is to predict the use behavior of faculty members. The relevant use survey questions are copied below (taken directly from the online documentation:

1. `USE1` - I use Wikipedia to develop my teaching materials
2. `USE2` - I use Wikipedia as a platform to develop educational activities with students
3. `USE3` - I recommend my students to use Wikipedia
4. `USE4` - I recommend my colleagues to use Wikipedia
5. `USE5` - I agree my students use Wikipedia in my courses

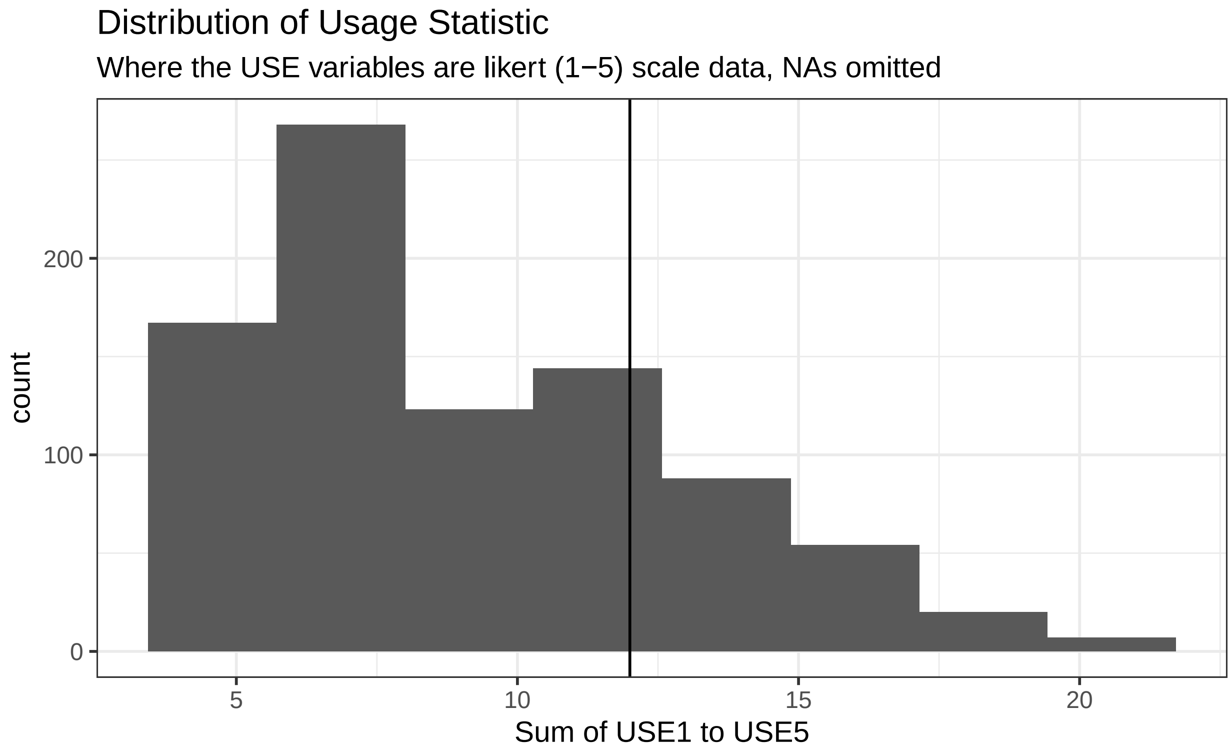
First, we observe the overall distribution of the `USE` variables (removing NAs). We note that most of the data is skewed to the right, especially for `USE1` and `USE2`, indicating that professors typically don't use Wikipedia for their own teaching materials.



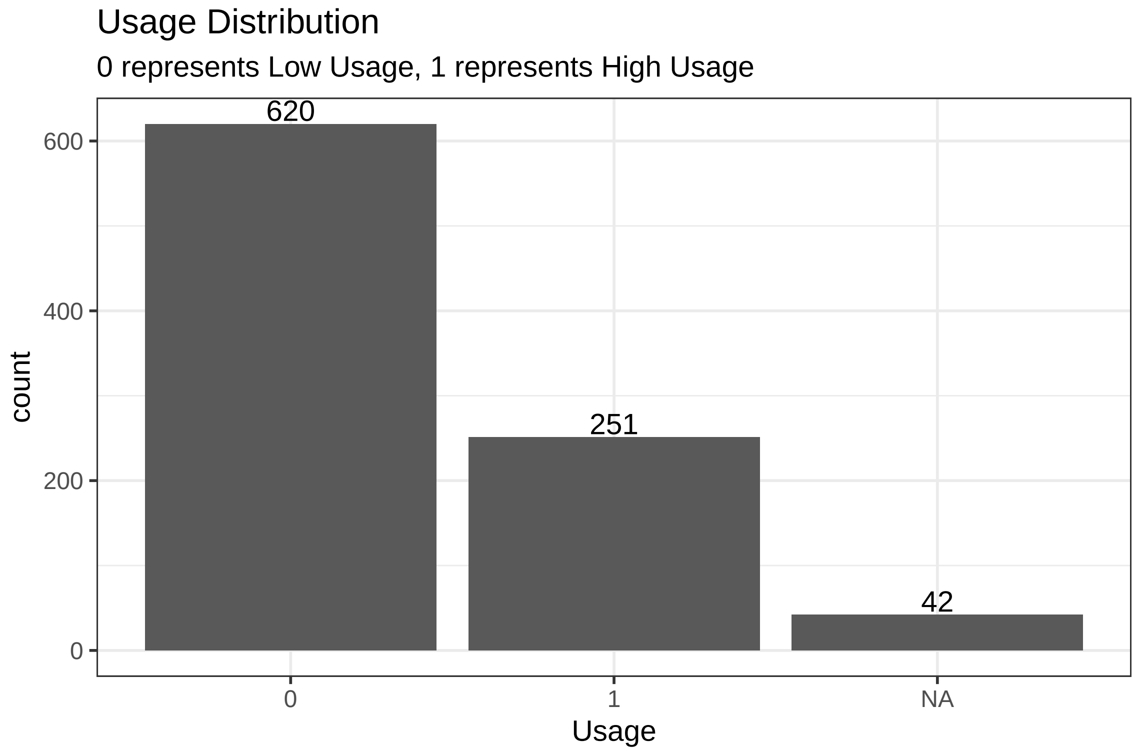
Since we want to use classification methods to predict the general use behavior of faculty members, we will create a new binary `Usage` statistic. We will sum `USE1` to `USE4`, then split the results based on the middle value of 12. Given that a value of 12 is equivalent to answering 3 (neutral) on all the questions, all values lower are considered "low usage". Values of 12 and above will be considered "high usage". The boundary point of 12 is included in "high usage" since it reflects a generally non-negative viewpoint on Wikipedia. `USE5` is not included since it does not reflect the instructor's personal use of Wikipedia, instead referencing just their beliefs for their students' behavior.

When considering just the 4 `USE` variables, we note that 42 faculty members left at least one of the responses NA. For now, we will omit those responses.

Examining the distribution of the four `USE` variables summed together, we can see that the data is still skewed to the right. From the plot, note that the line at x = 12 represents the boundary where all values to the left are considered "low", and all values to the right are considered "high".



Examining the number of Low and High usage counts, we observe that ~29% of the respondents (ignoring NAs) were categorized as High.



Analysis

We begin by dividing the data evenly (50/50) into training and test sets. This is done so that we can conduct the PCA and model building on the training data, and then determine the accuracy on the test data. This ensures that our model does not attempt to make "new" predictions on data that was used to create it.

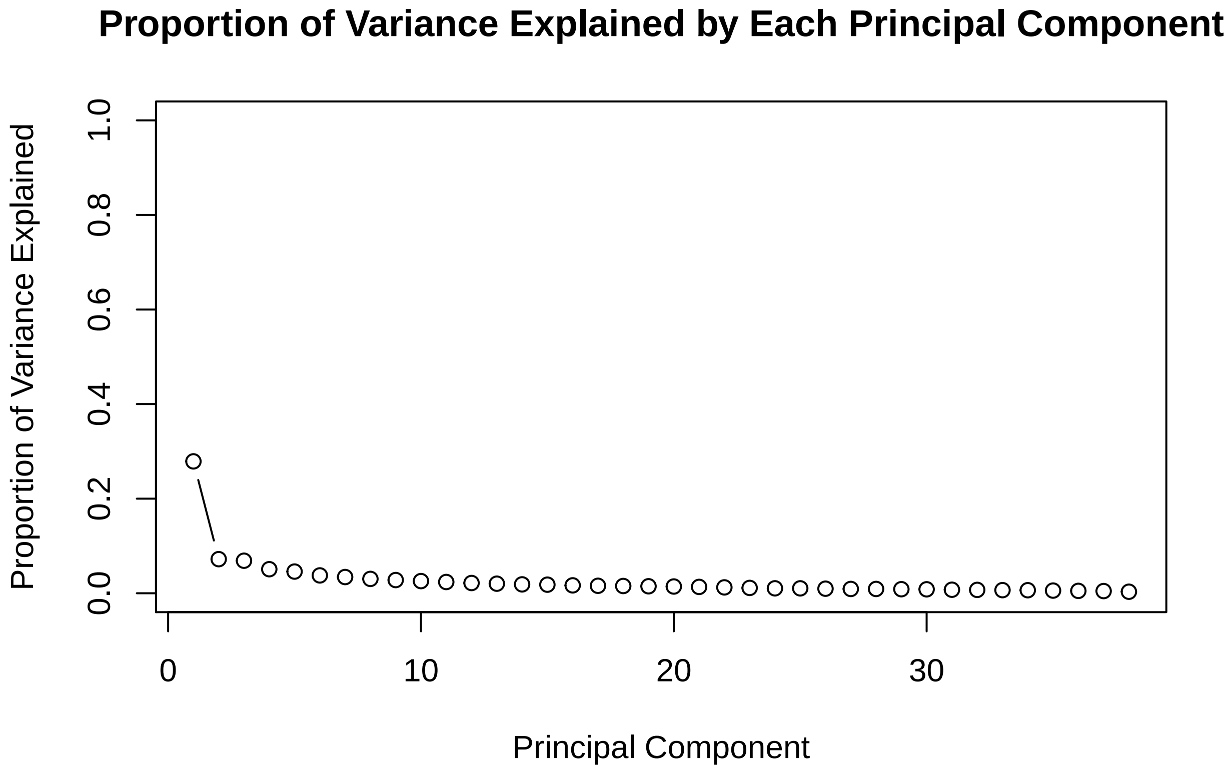
While splitting the data, we will stratify the random splitting by university and domain. Since the overall distribution of these results are not uniform, and we want to ensure that they are represented in each set, we split accordingly.

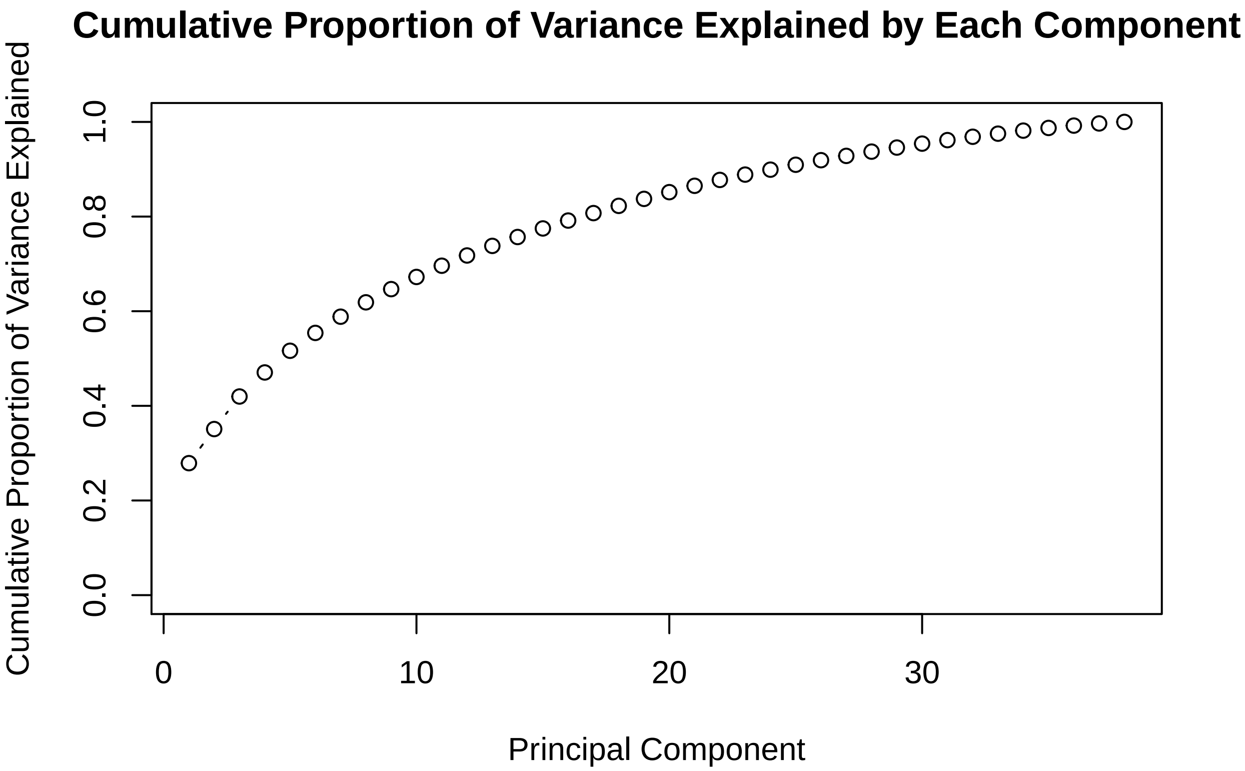
Part 1 - PCA

Given the large amount of questions, we first attempt to conduct a principal component analysis in order to reduce the number. For this, we first take the training data and isolate the ordinal (likert scale) data, removing the `USE` behaviors since that is our response later on.

While typically PCA should be reserved for continuous data only since it assumes a normal distribution and calculates a Pearson correlation, for this analysis we will continue forward. We do note however, that using polychoric correlations may have lead to a more accurate result. In this situation, we will implicitly assume that the likert type data can be treated as continuous. These predictors will be scaled to ensure that they are all weighted equally.

From the two plots below, we observe that the proportion of variance explained by each additional component begins to level off after just 2 components. Given that only 35.762% of the original variance is explained by those two components combined, we extend the number to 16 components (80.594%). This value was chosen since it covers a fairly large proportion of the variance explained, while reaching for higher (i.e. 90%) would have taken substantially more components.





Part 2 - Predict ’Use behavior

As explained in the first section, we will be estimating `Usage`, a binary statistic constructed by summing `USE1` to `USE4`, and then splitting the value based on the threshold of 12. Those with ratings 12 and above are considered ‘High Usage’, while ratings below 12 are considered ‘Low Usage’.

Note that since missing values are omitted for this model building, 157 observations are lost for the training data (from 459), and 147 from the testing data (from 454).

Teacher Attributes Selected

Besides the principal components determined previously, we proceed forward by deciding which general demographic attributes we wish to keep. We will maintain `AGE`, `GENDER`, `DOMAIN`, `PhD`, `YEARSEXP`, `UNIVERSITY`, `POSITION`, and `USERWIKI`.

```r
# Restrict to just Use1:Use4
na_count <-
  full_model_vars %>%
  na.omit() %>%
  nrow()

# We observe that there are 42 (out of 913) survey responses with an NA
nrow(full_model_vars) - na_count
```

Logistic

To begin, we conduct a logistic regression for `Usage`. By including the attributes specified previously along with the 16 principal components, we are able to create the model below.

Of the 28 original variables, we note that 5 of the principal components are significant for the full model, yet none of the teacher attributes are. While this interpretation may change given various variable selection techniques, for this report we will continue with the full model. Examining the confusion matrix, we observe that there is an accuracy of 84.7%, a sensitivity (true positive rate) of 73.2%, and a specificity (true negative rate) of 88.9%. Thus, it seems like the model over-predicts keep in mind.

```
##
## glm_pred   0   1
##        0 200  22
##        1  25  60
```

```
##
## Call:
## glm(formula = Usage_binary ~ ., family = binomial, data = omitted_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7534  -0.3460  -0.0835   0.2155   2.5573
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -4.00431    2.05017   -1.95   0.0508 .
## AGE          0.00836    0.03552    0.24   0.8139
## GENDER      -0.38577    0.48278   -0.80   0.4243
## DOMAIN      -0.06241    0.11500   -0.54   0.5873
## PhD          0.21028    0.53795    0.39   0.6959
## YEARSEXP    -0.00232    0.04624   -0.05   0.9599
## UNIVERSITY   0.79251    0.81026    0.98   0.3280
## USERWIKI     0.28506    0.58816    0.48   0.6279
## Position     0.22423    0.16545    1.36   0.1753
## PC1          0.90260    0.12757    7.08  1.5e-12 ***
## PC2         -0.96332    0.16802   -5.73  9.8e-09 ***
## PC3         -0.40869    0.13856   -2.95   0.0032 **
## PC4         -0.34697    0.17861   -1.94   0.0521 .
## PC5         -0.23932    0.17253   -1.39   0.1654
## PC6          0.17555    0.18856    0.93   0.3519
## PC7         -0.17197    0.22226   -0.77   0.4391
## PC8         -0.04031    0.21561   -0.19   0.8517
## PC9          0.28707    0.23230    1.24   0.2165
## PC10         0.29920    0.25968    1.15   0.2492
## PC11         0.06376    0.24455    0.26   0.7943
## PC12         0.43513    0.25151    1.73   0.0836 .
## PC13         0.12498    0.26762    0.47   0.6405
## PC14        -0.23426    0.26235   -0.89   0.3719
## PC15         0.71937    0.29275    2.46   0.0140 *
## PC16         0.55361    0.29300    1.89   0.0588 .
## PC17        -0.07510    0.29093   -0.26   0.7963
## PC18        -0.22559    0.29476   -0.77   0.4441
## PC19        -0.69334    0.29189   -2.38   0.0175 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 380.65  on 301  degrees of freedom
## Residual deviance: 158.77  on 274  degrees of freedom
## AIC: 214.8
##
## Number of Fisher Scoring iterations: 7
```

LDA

Next, we attempt to perform a linear discriminant analysis. Examining the confusion matrix, we observe that it has an accuracy of 82.736%, but has a sensitivity of 68.29% and a specificity of 88%. We again note that the true positive rate is relatively weak.

```
##
## lda_class   0   1
##         0 198  26
##         1  27  56
```

```
## Call:
## lda(Usage_binary ~ ., data = omitted_train)
##
## Prior probabilities of groups:
##      0      1
## 0.6755 0.3245
##
## Group means:
##      AGE  GENDER DOMAIN     PhD YEARSEXP UNIVERSITY USERWIKI
## 0 41.657 0.43627  4.299 0.44608   10.343     1.0980  0.10784
## 1 41.102 0.26531  4.000 0.41837   10.204     1.1531  0.30612
##   Position     PC1      PC2      PC3       PC4       PC5      PC6
## 0   5.1471 -1.2334  0.41248  0.13617  0.049650  0.045582 -0.01366
## 1   5.1837  2.6400 -0.89415 -0.33729 -0.050745 -0.092551  0.13428
##         PC7        PC8       PC9      PC10      PC11      PC12
## 0  0.035271 -0.0069985 -0.032443 -0.081768 -0.050741 -0.038382
## 1 -0.038626  0.0307182  0.045514  0.093038  0.104060  0.120131
##         PC13       PC14      PC15      PC16       PC17      PC18
## 0 0.00062599  0.0050886 -0.066728 -0.043739 -0.0053955  0.019611
## 1 0.03627344 -0.0030016  0.126493  0.046176 -0.0191137 -0.049720
##        PC19
## 0  0.049555
## 1 -0.125612
##
## Coefficients of linear discriminants:
##                   LD1
## AGE         0.0021497
## GENDER     -0.2096615
## DOMAIN     -0.0373724
## PhD         0.0535483
## YEARSEXP   -0.0040827
## UNIVERSITY  0.2936806
## USERWIKI   -0.1661251
## Position    0.0616517
## PC1         0.3608604
## PC2        -0.4308848
## PC3        -0.1726253
## PC4        -0.0685799
## PC5        -0.0737702
## PC6         0.0822688
## PC7        -0.0151996
## PC8         0.0573895
## PC9         0.0820837
## PC10        0.1422495
## PC11        0.0968800
## PC12        0.2239844
## PC13        0.0648377
## PC14       -0.0102898
## PC15        0.2930062
## PC16        0.1239507
## PC17        0.0067162
## PC18       -0.1177592
## PC19       -0.3249007
```

Quadratic Discriminant Analysis (QDA)

Finally, we move onto the QDA. Recall that while extremely similar to the LDA, QDA allows for unique covariance matrices and creates a quadratic relationship between the predictor and response.

Examining the confusion matrix, we observe that this model predicts more negative (Low Usage) than the previous models. While the accuracy is relatively close with a value of 82.085%, the sensitivity drops to 52.439%. While the specificity does go up to 92.889%, we note that this exchange is not a desired one, as the number of false negatives has drastically increased.

```
##
## qda_class   0   1
##         0 209  39
##         1  16  43
```

Comparison

Comparing the logistic regression, LDA, and QDA, we observed the following accuracy, sensitivity, and specificity rates:

From the chart below, we can observe that while all the model accuracies and specificities are relatively close, the logistic regression model maintained a relatively high sensitivity. Thus, given the choice between selecting one of these models, the logistic regression model seems to be the overall best.

	accuracy	sens	spec
LR	0.847	0.732	0.889
LDA	0.827	0.683	0.880
QDA	0.821	0.524	0.929

Conclusion

In conclusion, we observe that the logistic regression model provided the best predictive ability, having the overall highest accuracy and sensitivity, as well as a relatively high specificity. However, it still suffers from a relatively low true positive rate, overpredicting high usage among faculty members.

Regarding future progress, missing values must be accounted for much more effectively. While simply omitting observations with NAs does allow for models to be built on full data sets, there is a significant risk of introducing a large amount of bias. Given that the observations are not randomly missing, one cannot simply pretend that they don't exist. Thus, whether it's computing the missing values with a proportional odds model or some other methodology, one must consider the patterns behind the missing values.