USF Class of 2018

By: Adam Villarreal and Zach Dougherty

Question

How do various attributes of USF Freshmen affect their final cumulative GPA? How does it affect whether or not they graduate?

Data

- 1,495 students
- 9 predictor variables:
  - Gender
  - Pell grant status
  - unmet tuition percentage
  - in/ out of state
  - Major
  - science class
  - lab classes
  - Residence Hall
  - Ethnicity

- 3 Predictors:
  - GPA
  - GPA Credits (redundant)
  - Credits

Goals

Explanatory Modeling

- Impact of attributes
- Variable selection
- Fixing assumptions/ diagnostics

Predictive Modeling

- Which model is best?
- Best guess of GPA and likelihood to graduate at USF

Methods

Explanatory Modeling

- LASSO -> OLS
- Diagnostic Plots

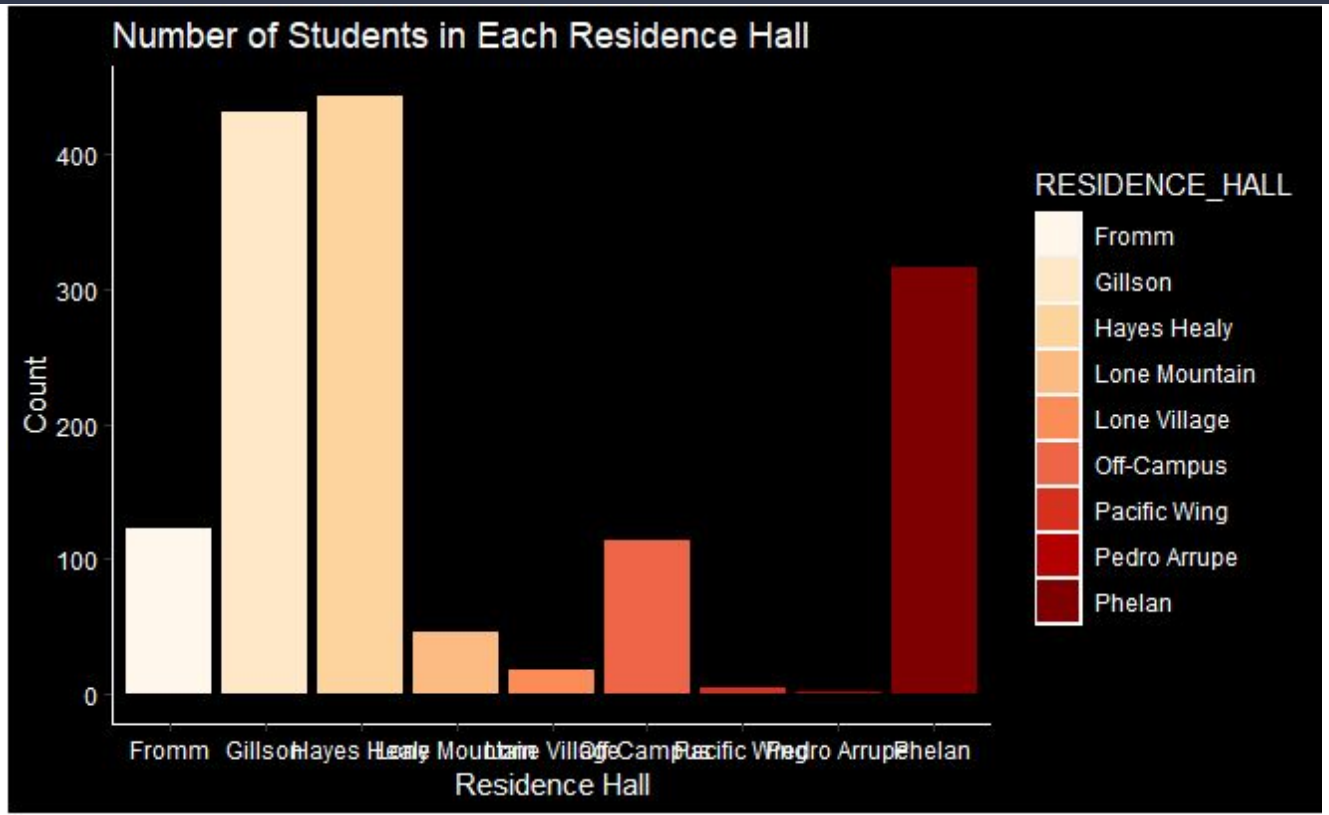
Predictive Modeling

- Regularization
  - LASSO
  - Ridge Regression
- Compare metrics

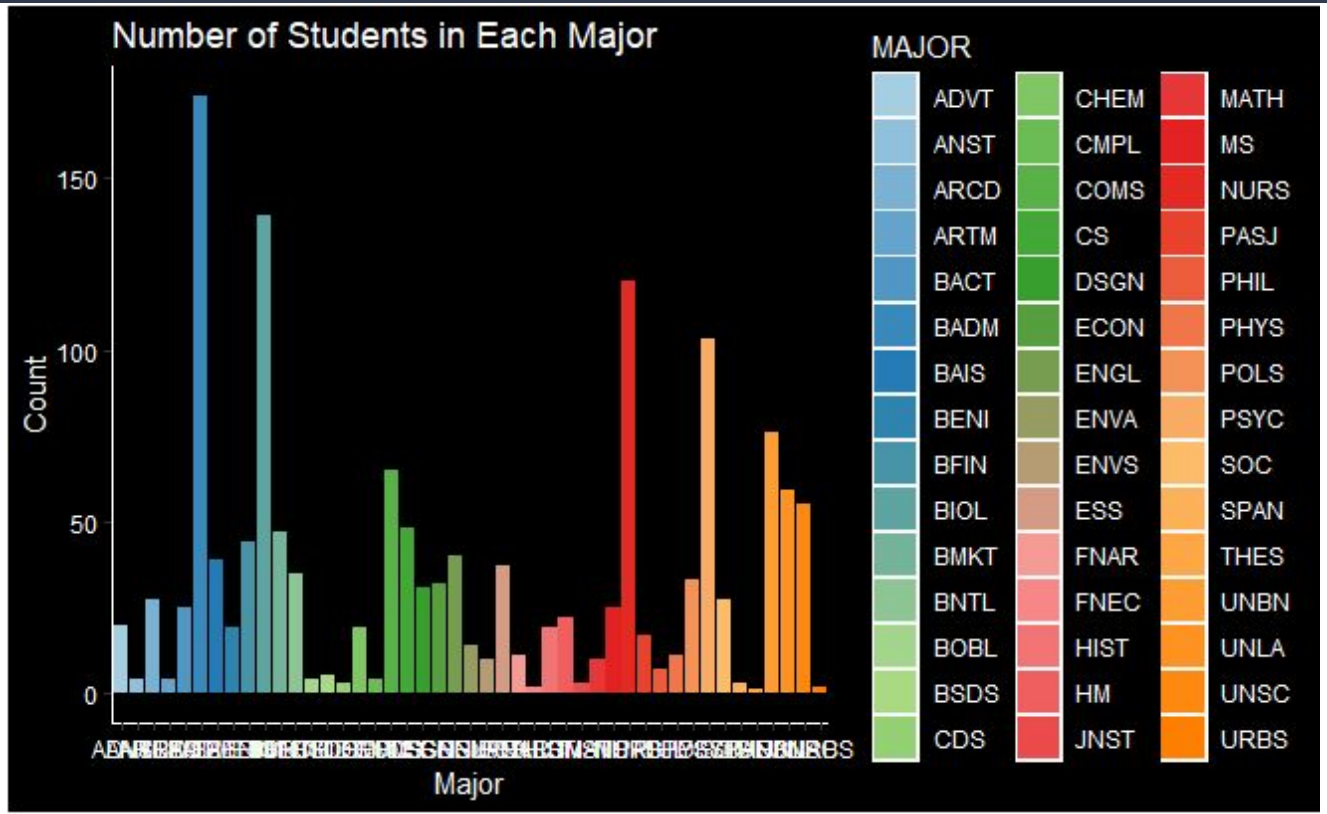
Data Challenges

- Interpretation is loose
  - Hard to say what certain credit values mean
  - Ex: 68 credits, drop-out? Transfer?
  - Max credits was 198, grad student? Change of majors?
- Students with 0 GPA
  - Box Cox requires positive response
  - Did they drop out?

Some statistics and Charts, Res Halls



Majors


Number of Students in Each Major
Count
150
100
50
0
Major
MAJOR
ADVT
ANST
ARCD
ARTM
BACT
BADM
BAIS
BENI
BFIN
BIOL
BMKT
BNTL
BOBL
BSDS
CDS
CHEM
CMPL
COMS
CS
DSGN
ECON
ENGL
ENVA
ENVS
ESS
FNAR
FNEC
HIST
HM
JNST
MATH
MS
NURS
PASJ
PHIL
PHYS
POLS
PSYC
SOC
SPAN
THES
UNBN
UNLA
UNSC
URBS

Statistics

- 65% of students are female and 35% are male
- 75% were from California while the other 25% are from elsewhere
- The avg student had 44.54% Unmet Need, so they had to pay 44.54% of tuition
- Only 25% of students qualified for Pell Grant while the other 75% did not
- The most popular majors are Business Administration, Biology, Nursing, and Psychology.

More Statistics

- There were 5 Data Science students!
- The average student took 7 science courses
    - One student took 40 science courses(Mistake?)
- The average student had 3 courses with required labs
- 7.5% of students lived Off-Campus their Freshman Year

Ethnicity Statistics

```
##   African American              Asian Hispanic or Latino
##         0.034782609        0.223411371        0.208026756
##      International         Multi Race     Native American
##         0.171906355        0.058862876        0.004013378
##   Pacific Islander            Unknown              White
##         0.008026756        0.011371237        0.279598662
```

These are the students ethnicities for the class of 2019 by percentage. The majority of the class identify as white at 28%. The smallest ethnicity group are those who idenitify as Native American which is less that 1% or 0.4% exactly. What is interesting is that 17% of the student population are international students which is the 4th largest ethnicity group.

Majority of this class identified as white, while African American, Pacific Islander, Unknown and Native American were the less represented.

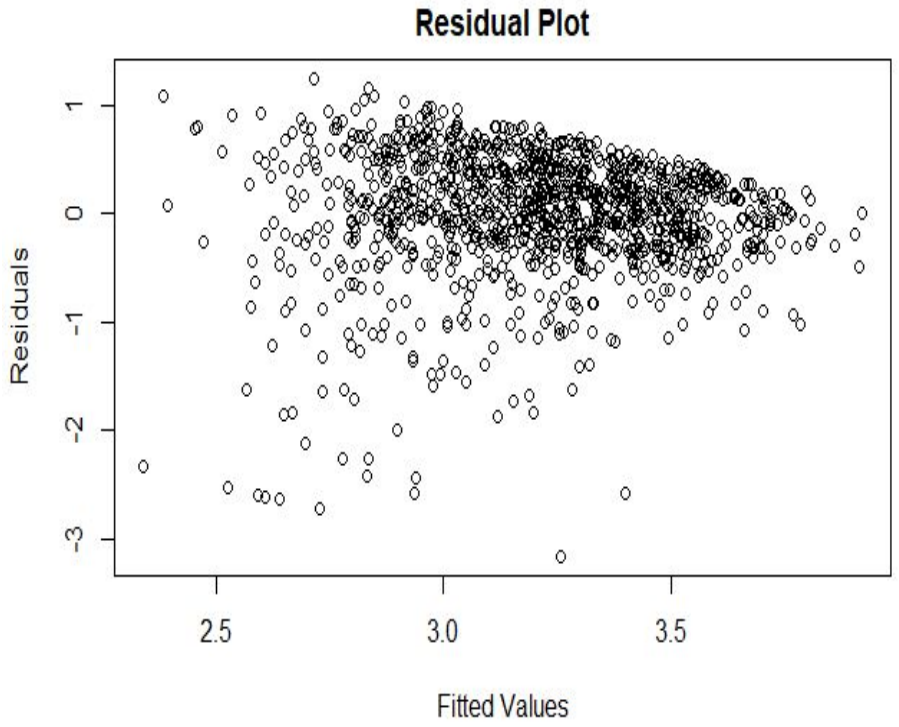
More Statistics

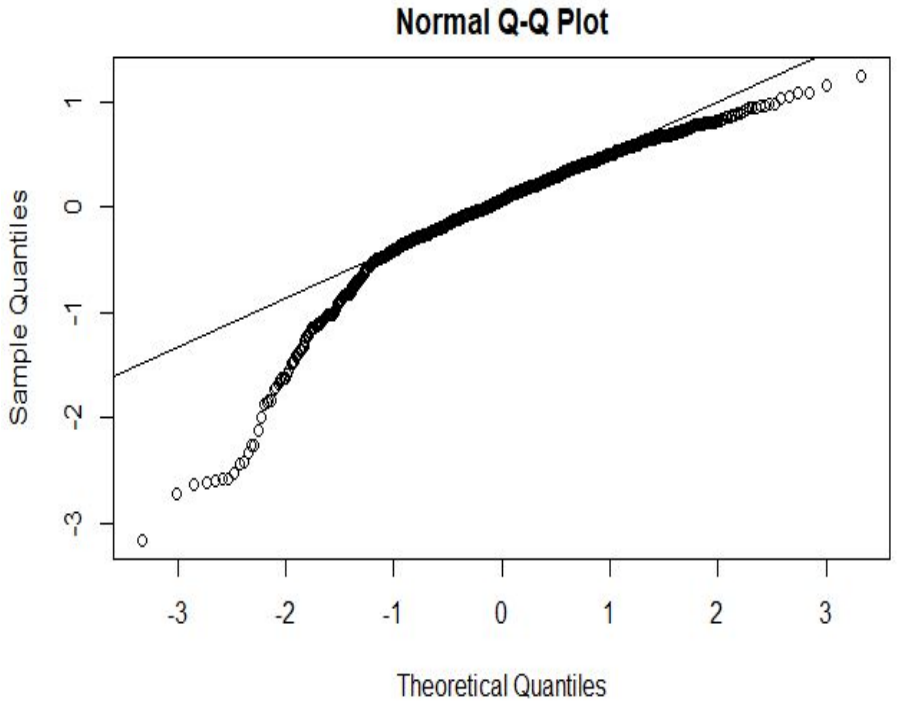
- The avg student completed 111 credits
  - Only 72% completed over 128 credits
  - 28% did not earn more than 128 credits
- The avg student earned 104 GPA credits
- The avg student had a GPA of 3.19

GPA Explanatory Model

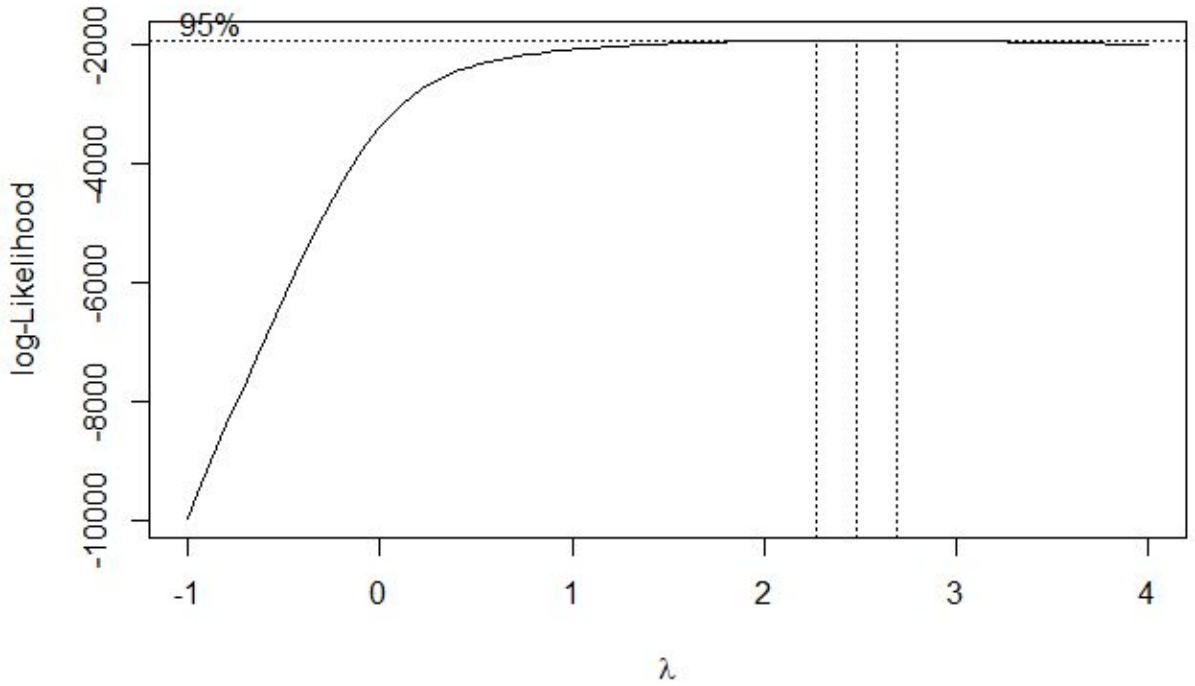
- Used a LASSO regularization to find important predictors
- Eliminated IN_STATE and UNMET_TUITION_PERCENTAGE
- Hard to eliminate certain dummy variables because of nature of our study
- Decided to include all Majors, Residence Halls, and Ethnicities
- Now for some diagnostic plots...

GPA Explanatory Model

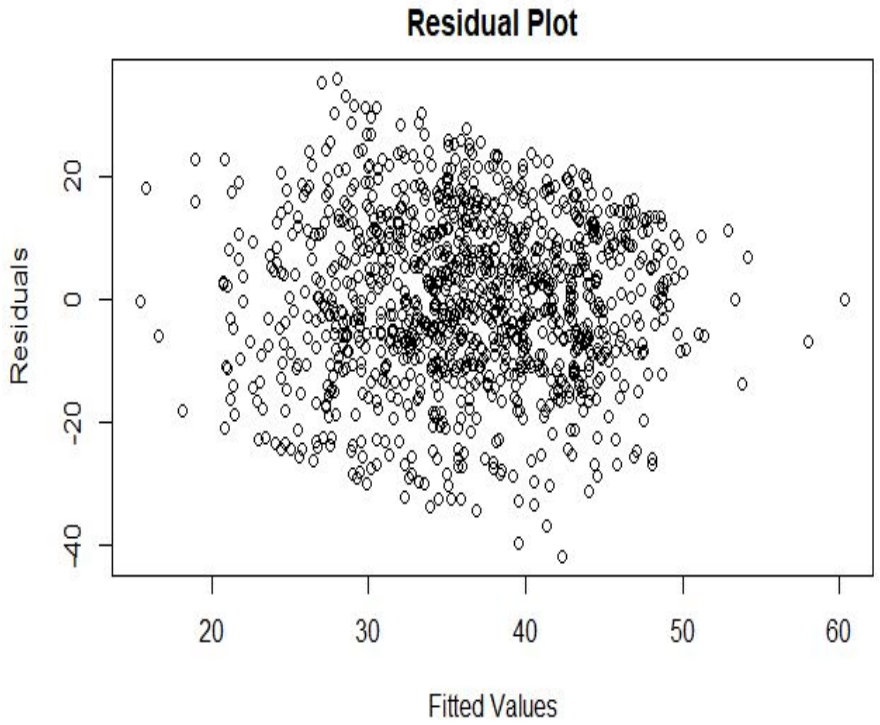


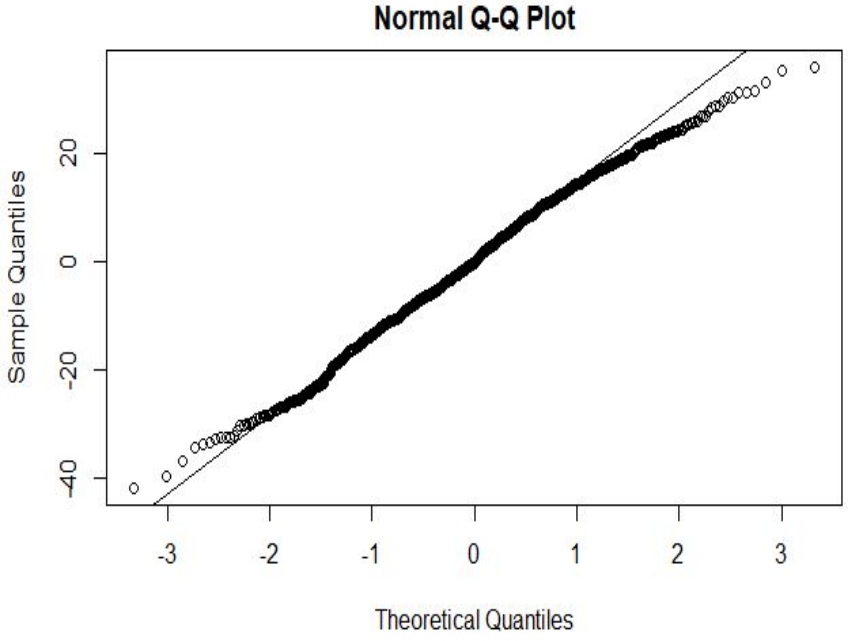


Box Cox Transformation



Post Transformation





How to improve?

- Failed both Shapiro Wilkes and Breusch Pagan tests
- Examine extreme values
  - Found 103 influential observations using threshold 4/n .
  - What is interesting about these values?
  - Contain more rare majors such as Theology Studies and lack popular majors such as Psychology and Business Administration
- Try removing influential points

Final Fit

- Interpretability is loose due to transformation
  - For a unit change in X, Y is shifted by approximately a cubed root of each coefficient estimate
- High variance of many coefficient estimates
  - Using Ridge penalty would remove all interpretability
- Low adjusted *R*-squared value
- Still fails BP and Shapiro Wilkes tests
- A few coefficient estimates and interpretation:
  - Those living in Lomo and Hayes have lower GPA than Fromm
  - Environmental Science majors had the highest negative impact on GPA, at -2.60 compared to Advertising majors
  - Asian Americans have 1.91 higher GPA than African American students

Predictive Model for GPA

- Compared models using cross-validation on both L1 and L2 regularization penalties
- LASSO:
  - MSPE: 0.405
- Ridge:
  - MSPE: 0.407

Final Fit

- Ridge Regression model is preferred here
  - Preservation of all factor levels
  - LASSO eliminates predictions for a number of student types
  - Small loss in accuracy preferred in order to keep all variables

Logistic Regression

Steps

1. Clean the dataset and remove influential points
2. Create a new discrete variable that describes if the student graduated or not
    a. 1 if student earned more or equal to 128 credits
    b. 0 for otherwise
3. Remove the CREDITS_EARNED variable since we would like to know what other variables predict whether the student graduated or not
4. Create a balanced dataset since the dataset is unbalanced
5. Split both datasets into a train and test dataset
    a. 75% of data as training data
    b. 25% of data as testing data
6. Run two models on both datasets

Balanced Data Trouble

- After cleaning there is 1393 students left
  - 1052 Graduated
  - 341 Did not graduate

Based from above, it is clear that the data is unbalanced. The way I thought to make it balance was to take 341 graduated students and then append that data to the others who did not to make a dataset with 682 observations. However, when I would test the model that I created, the model would see majors that it was not trained on creating an Error.

These majors are the less represented since there are less students studying them.

How to fix this problem?

Two Options:

1. Delete the Major Row - We would lose insight of what majors lead to success
2. Oversampling - Generate new data that would balance the number of majors, but would bring bias into the model

What I did:

I used set theory to identify what majors existed in the testing data that did not exist in the training data and removed those majors in the testing dataset. Then I ran the model and tested it. Some observations were deleted so the data was not perfectly balanced, but it was still very close

Model 1 Inference

- GPA and GPA credits contribute most to whether if a student graduates or not
- Most majors contribute positively
  - Japanese Studies contributed the most
- Loyola Village is the best dorm to live
- If you are Pacific Islander, Hispanic or Unknown, it would contribute to whether you will graduate or not
  - This does not insinuate that some ethnicities perform better than others since it is the number of observations between ethnicities are not equal.
- 45 positive coefficients

Model 1 Assessment

This was created with the unbalanced dataset

True-Positives: 257

False-Positives: 6

True-Negatives: 82

False-Negatives: 4

Remember we want to increase the number of TP and TN, and to decrease the number of FP(Type I error) and FN(Type II error)

Model 1 Assessment

- Accuracy is 97% - Testing accuracy
- Sensitivity is 98% - Performance of testing true positives correctly
- Specificity is 93% - Performance of testing true negative correctly
- Sensitivity is 98% - Positive Predictive Value

Model 2 Inference

- GPA and GPA credits contribute most to whether if a student graduates or not
- If the student is male, it is more likely that they would graduate
- If the student is from California, it is more likely that they would graduate
    - Makes sense for out-of-state students will have a higher probability of transferring
- Students with higher unmet need percentage are more likely to graduate
    - Students who pay more are more likely to stay
- Students with Pell Grants are less likely to graduate
- Students who take more Lab Classes are more likely to graduate
- International Studies is the best major
- Living in Pac-Wing or Loyola Village contributes positively
- 36 positive coefficients

Model 2 Assessment

This was created with the balanced dataset

True-Positives: 74

False-Positives: 5

True-Negatives: 86

False-Negatives: 5

Remember we want to increase the number of TP and TN, and to decrease the number of FP(Type I error) and FN(Type II error)

Model 2 Assessment

- Accuracy is 94% - Testing accuracy
- Sensitivity is 94% - Performance of testing true positives correctly
- Specificity is 95% - Performance of testing true negative correctly
- Sensitivity is 94% - Positive Predictive Value

Which Model is Better?

Model 1 performed better in all relationships except the specificity. The second model has higher specificity rate, which is basically the ability to rightly predict if the student would not graduate.

This is surprising!

It seems that the number of observations matters more rather than the whether the data being used is balanced. My first guess was that the first model overfitted, but it returned a greater testing accuracy than the second model. For predictive purposes, I would proceed using the first model. But for inference, I would work with the second model, especially if we had more balanced data.