REGRESSION ANALYSIS

OF STUDENT ENGAGEMENT WITH UNIVERSITY HEALTH PROMOTION DEPARTMENT

Data Science research by Elisabeth Starr at the University of San Francisco

ABSTRACT:

The University of San Francisco sponsors an annual challenge called "Go Dons Get Fit" in which students and faculty compete to see who can log the most exercise minutes in the month of October. The goal is to encourage an active lifestyle and reduce stress among students. An exploratory analysis of student data from 2021 conducted in R found that student engagement decreased over the course of the month. The purpose of this analysis is to identify variables that contribute to student engagement over the month so the department can incentivize consistency.

BACKGROUND

A regression analysis is a statistical method for evaluating the relationship between a dependent variable and one or more independent variables. OLS(ordinary least squares) is a method which provides the best unbiased estimator that minimizes the distance between observed Y values and predicted Y values.

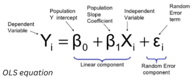


Fig.1 OLS equation

Lasso (least absolute shrinkage and selection operator) is a regression technique that adds an L1 penalty to the OLS equation. The lambda in the equation is selected by k-fold cross validation.

$$L_{lasso}(\hat{\beta}) = \sum_{i=1}^{n}(y_i - x_i^T\hat{\beta})^2 + \lambda\sum_{j=1}^{m}|\hat{\beta}_j|$$

Fig.2 LASSO equation

OBJECTIVE

An exploratory data analysis found that students disengage with the challenge over time, resulting in a drop in exercise over the course of the month. The purpose of this regression analysis was to look for variables with a positive correlation to exercise so the department could better identify strategies to prolong the engagement of students.

Some of the variables examined were the types of exercise recorded by students such as cardio, weights, sports, and yoga.

Another variable examined was the students' academic department, including nursing, law, and arts and science. The intention was to identify if participation was distributed evenly across majors and to observe whether some academic departments exercised more or less as a whole.

METHODOLOGY

For this analysis an anonymized set of student data was loaded into an R coding environment. Cleaning this data set was the most time consuming part of the process as the data set was not formatted ideally and variables that should have been quantitative were collected as categorical data.

Once the data was organized, variable selection was performed using Lasso and k-fold cross validation. Because Lasso does not deliver unbiased estimators, an OLS linear regression model was applied afterwards in order to quantify the relationships between the selected variables.

Both of these regression models were built using glmnet[1], an R package designed for regression algorithms.

It is important to note that the 2nd and 3rd weeks of October coincide with midterms and Fall break, respectively. However, it is not possible to discern if that was a contributing factor to the decline in exercise. It is also not possible to know if students stopped exercising or just stopped logging minutes.

RESULTS

The results of the regression analysis showed that while student engagement did drop over the course of the month, this wasn't universal across activities. Students who recorded activities classified as "cardio" or "sports" finished the month with higher and more consistent minute counts than students who logged "strength". And students who recorded "yoga" minutes had a below average total minute count by the end of the month. This could be because more active students tend to already participate in group sports or it could be because activities that lend themselves to lengthier sessions, like walking, are categorized as cardio.

ANALYSIS

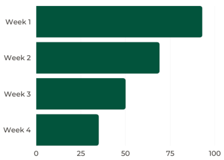


Fig.3 Total student participation by week October 2021

In figure 3 it is clear that by week 4 student engagement tapered off by nearly 70% from week 1. The health promotions department was aware of this trend but was surprised by this graph which quantified the decline at about 20% per week. Figure 4 demonstrates a breakdown of minutes logged per week per academic department. From this it is clear that the decline is distributed evenly across majors.

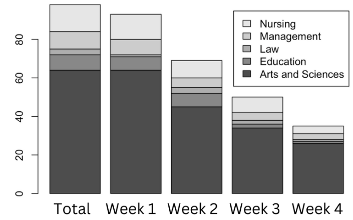


Fig.4 Student minutes logged per week by academic department October 2021

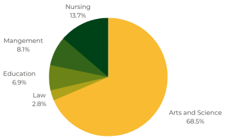


Fig.5 Student participation by academic department

Figure 5 demonstrates the percentage of participants by academic department. It appears that the college of Arts and Science makes up the majority of participants, which is consistent with the data in figure 4.

It was hypothesized that nursing students with a heavier academic load might exercise less than other majors, but the data does not support that hypothesis.

Lasso was used to observe the relationship between amount of exercise and type of exercise. This requires tuning the lambda parameter for the L1 penalty. Using k-fold cross validation and plotting the MSEs is helpful.

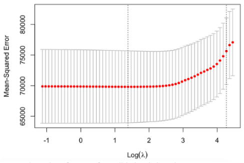
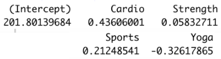


Fig.6 Plot of MSEs of possible lambda values

In figure 6, the plot of the MSEs of possible lambda values confirms the lambda selection given by Lasso of 3.9. Once the lambda has been chosen the regression can be run. The coefficients show that cardio appeared to be most correlated with logging more minutes than average, followed by group sports. They also show that logging yoga minutes had a negative relationship with total time logged.



```
 (Intercept)      Cardio    Strength 
201.80139684  0.43606001  0.05832711 
                  Sports        Yoga 
              0.21248541 -0.32617865 
```

Fig. 7 Regression coefficients from the OLS model

When we then ran the same data on an OLS model we found that these relationships were even stronger and clearer than in the Lasso model.

REFERENCES

[1] Lasso and Elastic-Net Regularized Generalized Linear Models https://glmnet.stanford.edu/index.html

RECOMENDATIONS

Key findings of this analysis are:

1) Confirmation that student engagement decreased significantly over the course of the month. This was consistent across majors and genders.

2) The majority of students who participated were from the Arts and Science department.

3) Students who logged cardio or group sports minutes were more likely to continue the challenge and to have a higher average number of minutes.

Based on these findings, our recommendations for the health promotion department are to offer prizes, not just for most minutes logged, but for consistency or exercise streaks. We also recommend doing more outreach to try to recruit more participants from outside the Arts and Sciences department so more students can have access to this program.

UNIVERSITY OF SAN FRANCISCO