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 that provides the best unbiased estimator that minimizes the distance between observed Y values and predicted Y values.

\[ Y = \beta_0 + \beta_1 X + \epsilon \]

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.

```latex
\lambda_{opt} = \frac{1}{\sum \lambda_{i}^2 / \sum \lambda_{i}^2} \sum \lambda_{i}^2 / \sum \lambda_{i}^2
```

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 estimates, 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(), 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.

RECOMMENDATIONS

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.

REFERENCES

[1] Lasso and Elastic-Net Regularized Generalized Linear Models
https://github.com/stanford-smartcity/