Turing Admissions and Outcomes Analysis

Erin Williams

Springboard - Foundations of Data Science

TURING SCHOOL OF SOFTWARE & DESIGN

Overview of problem

Using compiled data from enroll, apply, and our internal records, I sought to answer the following questions:

1. Students with which demographic characteristics are most likely to drop out? (race, gender, logic scores, payment plan)
2. Are there factors which make a student more likely to repeat?
3. Can we calculate a student's likelihood of repeating and support them accordingly?
4. What are the "success rates" of different interviewers?
5. Are students with higher logic scores more likely to be successful completing the program?

Logistic regression models

Based on the data available, I built two prediction models that allowed me to answer these questions:

1. ***What is the likelihood of graduation for students from various demographic groups based on their initial logic score?***
2. ***Are there factors which make a student more likely to repeat?***

Initial statistical analysis

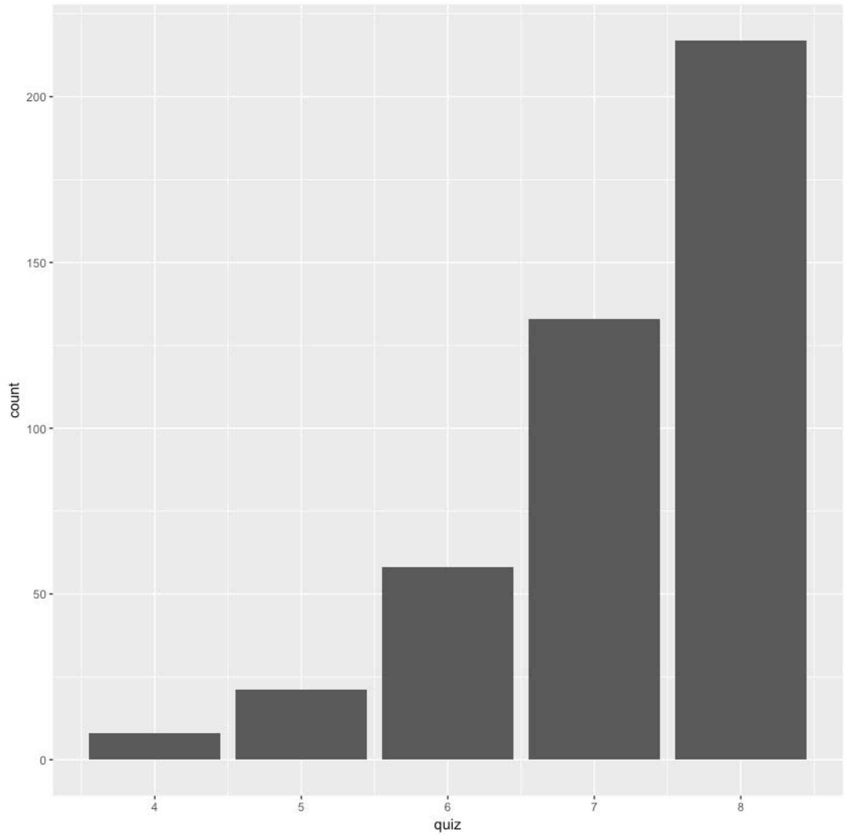
Data set: alldata_safe1

Compiled from enroll, apply, demographic surveys

A count of application quiz scores among enrolled students.

Most accepted students have higher quiz scores.

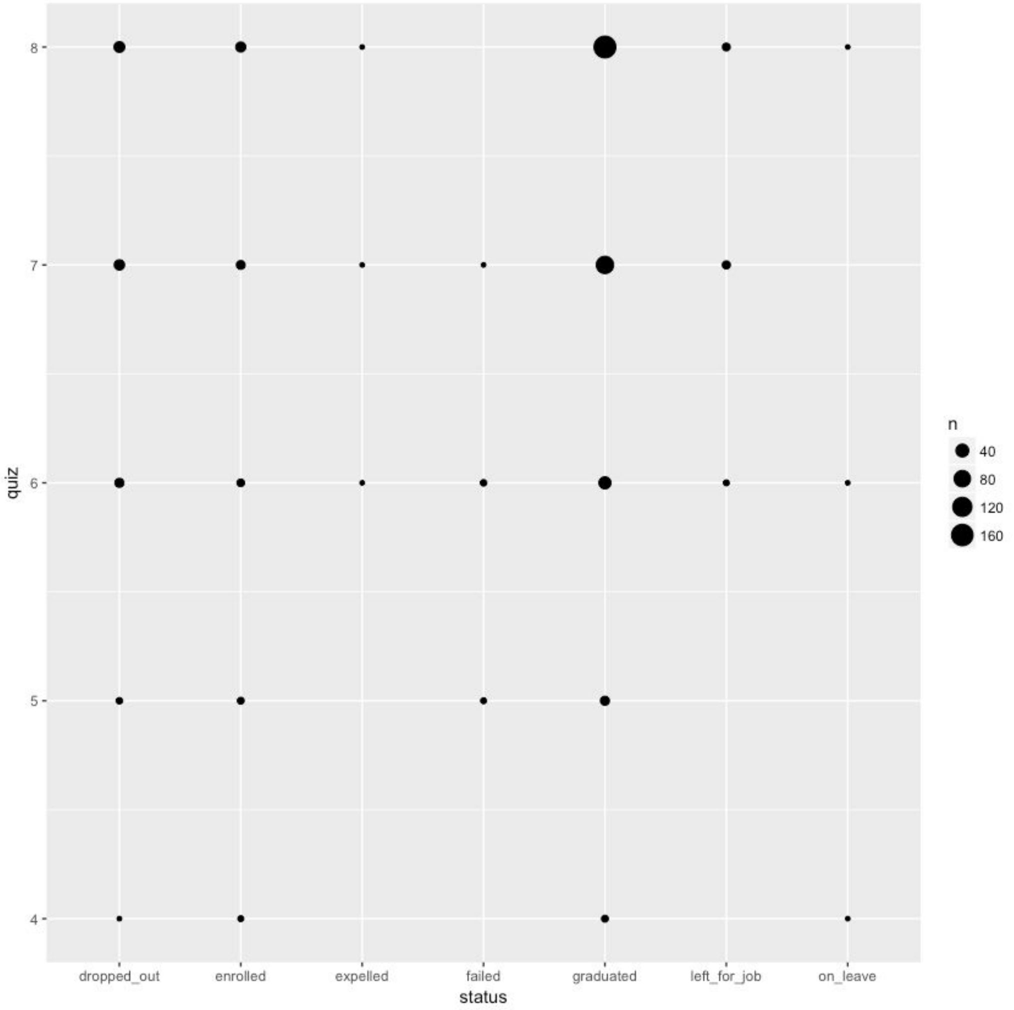
```
ggplot(alldata_safe1$status, aes(quiz)) +
    geom_bar()
```



Initial quiz scores compared to graduation status.

No students with 4 or 5 have been expelled, and no students with 8 have failed.

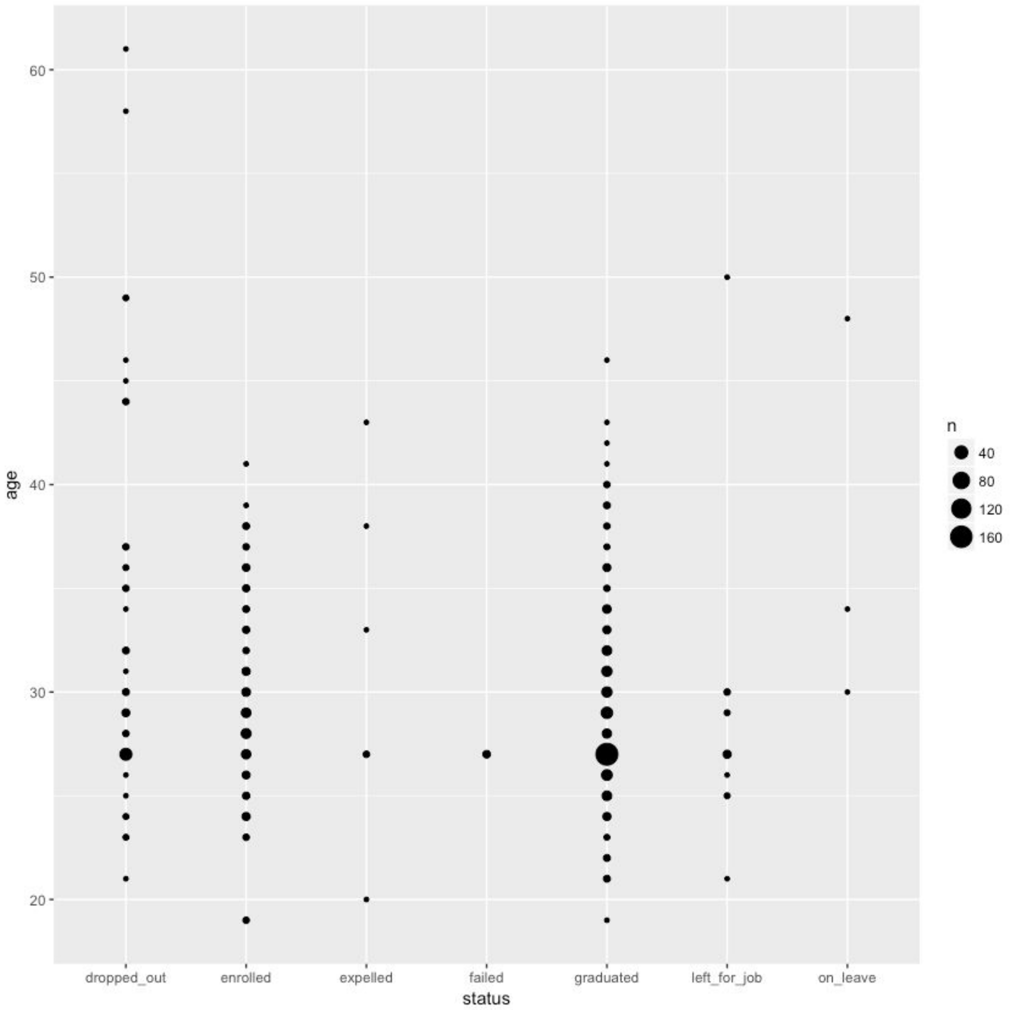
```
ggplot(alldata_safe1, aes(status, quiz)) +
    geom_count()
```



Age and status

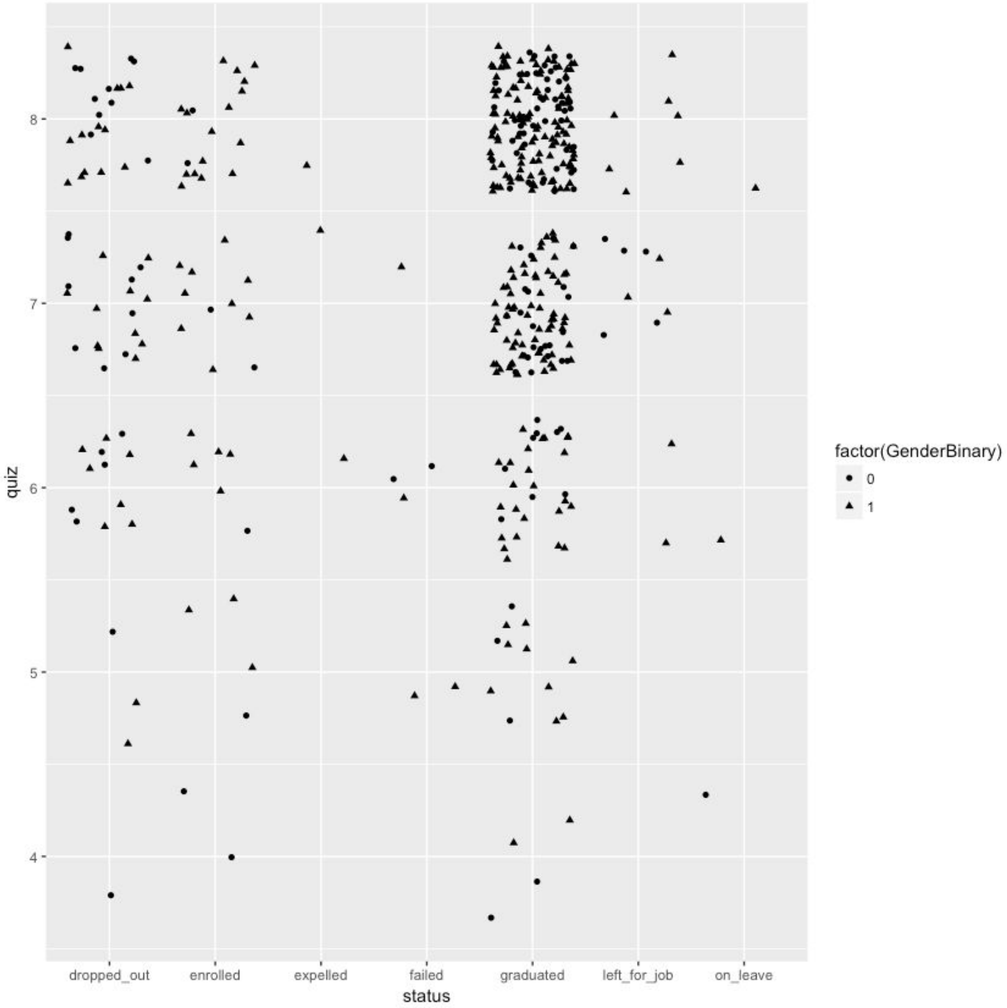
Older students have higher incidences of dropping out.

```
ggplot(alldata_safe1, aes(status, age)) +
  geom_count()
```



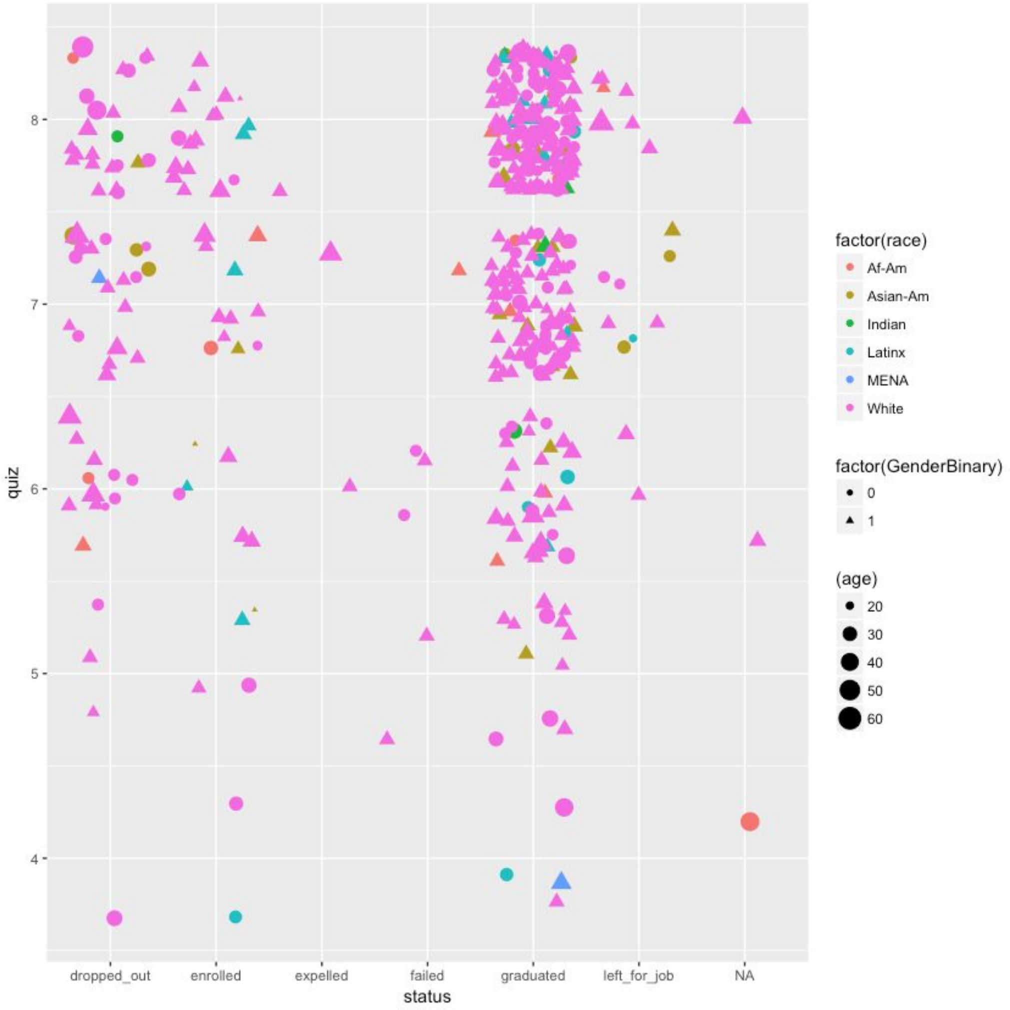
Quiz, gender and status

```
ggplot(alldata_safe1, aes(x = status, y = quiz,
shape = factor(GenderBinary))) +
geom_jitter()
```



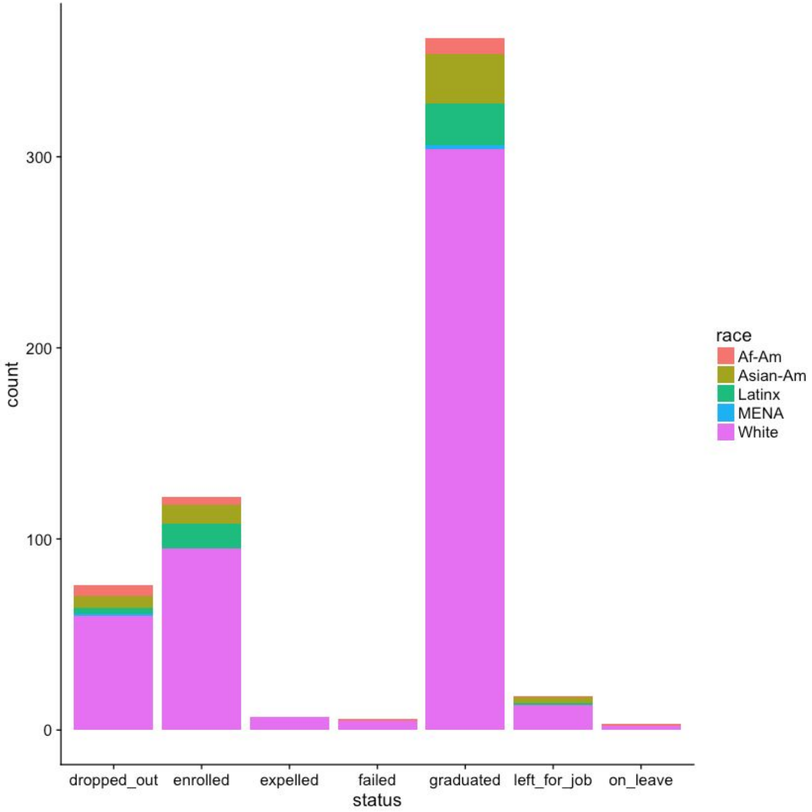
Quiz, race, gender, age and status

ggplot(alldata_safe1, aes(x = status, y = quiz, col = factor(race), shape = factor(GenderBinary), size = (age))) + geom_jitter()



Status count by race

```
ggplot(alldata_safe1, aes(x=status, fill=race)) +
geom_bar()
```



Applying the prediction model

What is the likelihood of graduation for students from various demographic groups based on their initial logic score?

I created a logistic regression model using the dependent variable ‘graduated’, where graduated = 1, and every other status = 0.

```
glm(graduated~age+quiz+GenderBinary+White+veterans+BE+int_logic_score+enrollments,
data=alldata_safe1, family="binomial")
```

Logistic Regression Code

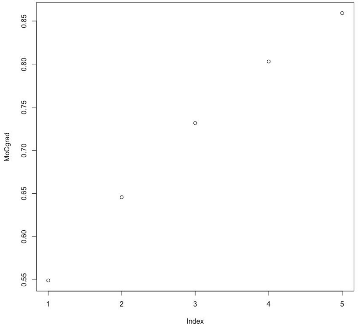
```
Call:
  glm(formula = graduated ~ age + quiz + GenderBinary + White +
        veterans + BE + int_logic_score + enrollments, family = "binomial",
      data = alldata_safe1)

Deviance Residuals:
  Min        1Q    Median       3Q       Max
-3.2680  -0.3566   0.4492   0.5757   1.8671

Coefficients:
  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -7.94399    1.83012  -4.341 1.42e-05 ***
  age             -0.02190    0.02931  -0.747  0.45509
quiz             0.40282    0.14299   2.817  0.00485 **
  GenderBinary     0.20024    0.30797   0.650  0.51558
White            0.47126    0.35764   1.318  0.18761
veterans        -0.38063    0.99086  -0.384  0.70088
BE              -0.44465    0.38311  -1.161  0.24578
int_logic_score  0.16355    0.10257   1.595  0.11080
enrollments      1.30982    0.13857   9.453  < 2e-16 ***
  ---
  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 530.29  on 436  degrees of freedom
Residual deviance: 343.17  on 428  degrees of freedom
(157 observations deleted due to missingness)
AIC: 361.17

Number of Fisher Scoring iterations: 5
```

Predictions

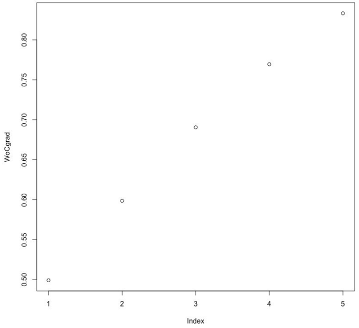
Using this model, I wanted to predict the likelihood of graduation based on an applicant's race, gender, and initial quiz score. I got an array of numbers that looked like this. This basically shows a positive correlation to graduation rates for all groups.

Likelihood of graduating with a quiz score of 8:

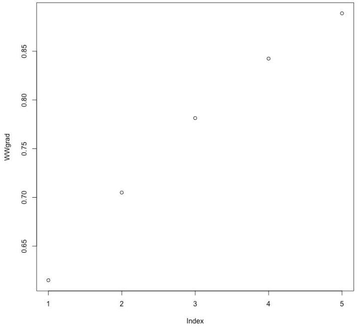
Men of color: 86%
Underrep: 88%



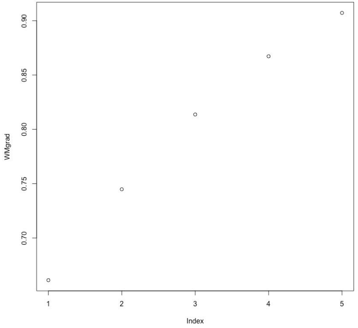
Women of color: 83%
Underrep: 85



White women: 89%
White & Asian: 88



White men: 91%
White & Asian: 90



Predictions - positive outcomes

Using this model, I wanted to predict the likelihood of graduation based on an applicant's race, gender, and initial quiz score. I got an array of numbers that looked like this. This basically shows a positive correlation to graduation rates for all groups.

Likelihood of positive outcomes with a quiz score of 8:

Men of color: 93%	Women of color: 91%	White women: 93%	White men: 94%
Underrep: 94	Underrep: 92	White & Asian: 93	White & Asian: 94

Let's try it! Predict a student's likelihood of graduating from Turing:

The factors available in this model are:

- Age
- Gender (0=F, 1=M)
- Race (1=white or 0=PoC in this model)
- Quiz score (4 through 8)
- Vet status
- Program
- Interview Logic Score (6-14)
- Number of enrollments (4-7)

Basic code to make a prediction

- new.pred <- data.frame(age=XX, White=X, GenderBinary=X, quiz=c(4:8), veterans=X, BE=X, int_logic_score=XX, enrollments=X)
- predict(log.reg.white, new.pred, type = "response")

Recommendations

1. Continue using the logic test as an admissions tool. This is strongly correlated with student success and likelihood to graduate. Require all applicants to take the logic test, refining the Fast Track process to include the logic test.
2. Collect anecdotal, qualitative data from students of color and female students who graduated and those who did not in order to deepen the analysis and determine why students from underrepresented groups are less likely to graduate.
3. Based on results of surveys above, offer support immediately to underrepresented students who score a 7 or 8 on their initial logic evaluation.

Ideas for further research

Idea 1: For the second prediction model, I wanted to answer the following question: *Are there factors which make a student more likely to repeat?* I built another model using repeats as the dependent variable, but haven't analyzed it in depth.

Idea 2: Further research that would be valuable for our admissions process would be to review the success rates of various interviewers (question 4 above). This would require some data wrangling to make the list of interviewers into a readable format for logistic or linear regression. I would probably make new binary columns or assign each interviewer a number then run a similar regression model to the one above.

Idea 3: Create a linear regression model by scaling the 'behavior' column from 1-5, then backtracking that to interview red flags (this would require significantly more data collection than is currently available).

TURING

SCHOOL OF SOFTWARE & DESIGN