CAPSTONE 2: PROJECTING PRODUCT-CUSTOMER FIT WITH PUBLIC DATA

By: Sam Fisher

Mentor: Evan Elg

WHO AM I?

- A student making a career transition into data science via Springboard.
- A fellow human with curiosity, the internet, and some time to study.
- Spent most of my adult life as a professional composer after completing a bachelor's in music technology at Oberlin Conservatory.
- Worked in two tech start-ups: Seller Labs and Brain.fm.
- Not scary.

Please feel free to ask questions, challenge things, and engage!

OBJECTIVES FOR THIS TALK.

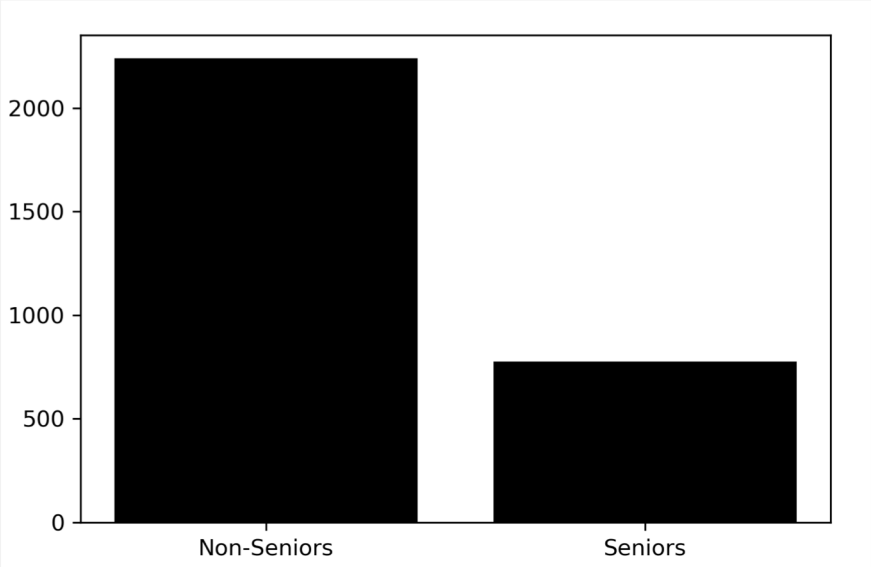
A. Introduce the business problem of our fictional client, Edugames.org
B. Introduce public survey data to help us solve the problem
C. Step through the analysis of this data
D. Present our recommendations to Edugames.org based on this analysis

WHICH SENIORS WOULD PLAY AN ONLINE MATH GAME?

- Edugames.org is seeking to create online math games for seniors.
- Retired seniors are sometimes bored. Why not have an engaging hobby?
- We want to predict what specific demographics fit the product.
- There is no hard data on sales or product engagement yet.

PUBLIC SURVEY DATA PROVIDES RELEVANT INFORMATION.

- Disclaimer: This project uses public data from Pew Research Center's "Information Engaged and Information Wary" survey. Pew Research Center bears no responsibility for interpretations presented or conclusions reached based on analysis of the data.
- Sample includes 2240 Non-seniors (age < 65) and 775 Seniors (age >= 65)
- Questions On:
  - Time and Attention
  - Growth Mindset
  - Technology Engagement
  - Demographic Information



ANALYSIS METHODOLOGY

- **We'd like to use demographic factors to predict for things that we assume will make an ideal customer.**
  A. Data cleaning – is this valid input?
  B. Feature Design
     1. How do we quantify customer fit?
     2. How can we use the demographic data in a linear model?
  C. Correlation Analysis – what covaries with what?
  D. Linear Regression Model – how do demographic factors impact customer fit?

DATA CLEANING

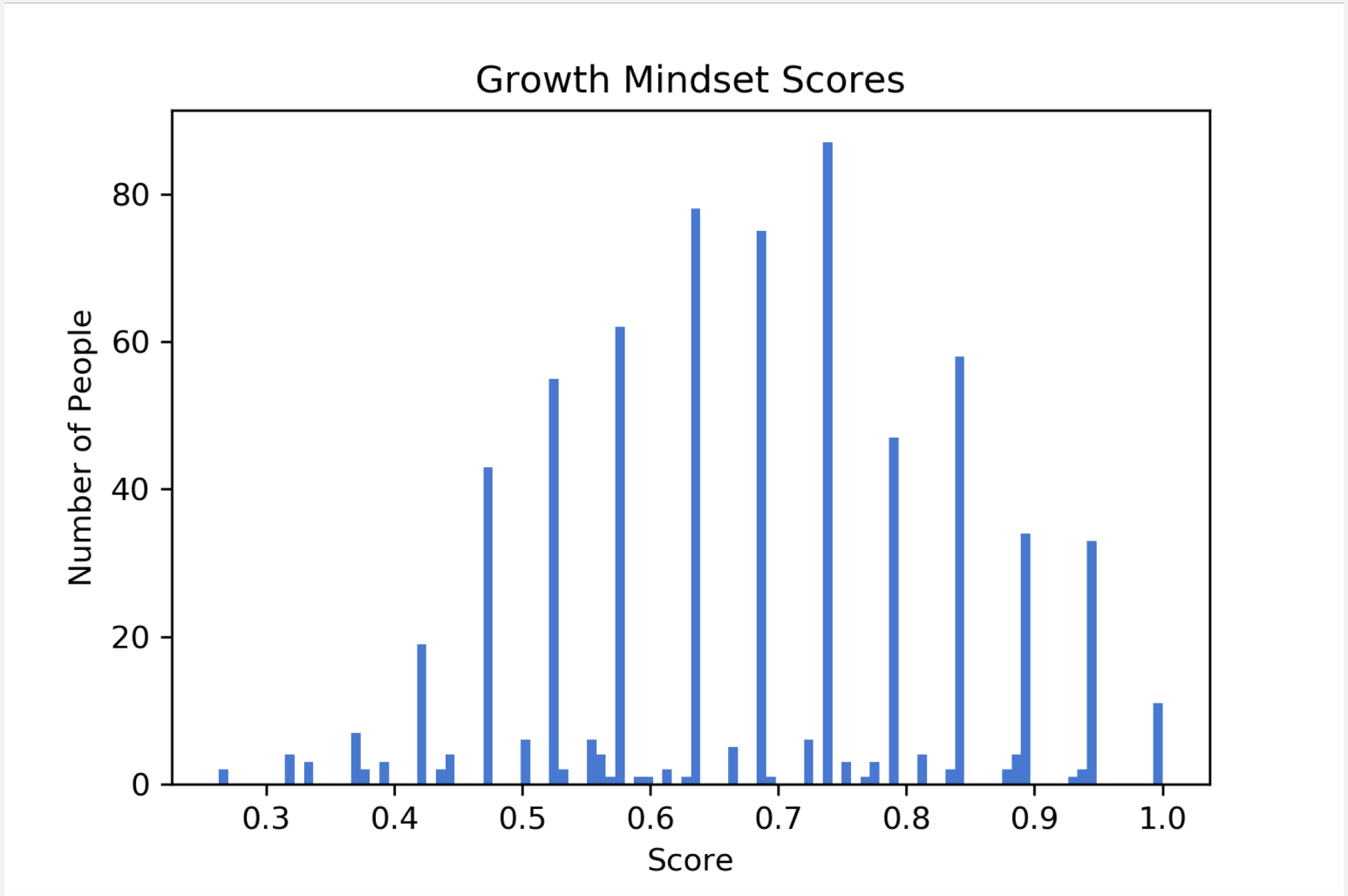
- 117 features encoding categorical responses to individual questions.
- No missing values – very clean data set.
- Some empty strings if a question is only asked under a condition.
- Some non-responses- either "I don't know" or "I don't want to tell you."
- Non-response was dealt with in two ways.
  - Selecting a subset of complete responses for demographics (about 89% of seniors)
  - Removing questions that individual respondents did not answer from their score calculation (assumes that non-response is fairly random)

THE IDEAL CUSTOMER

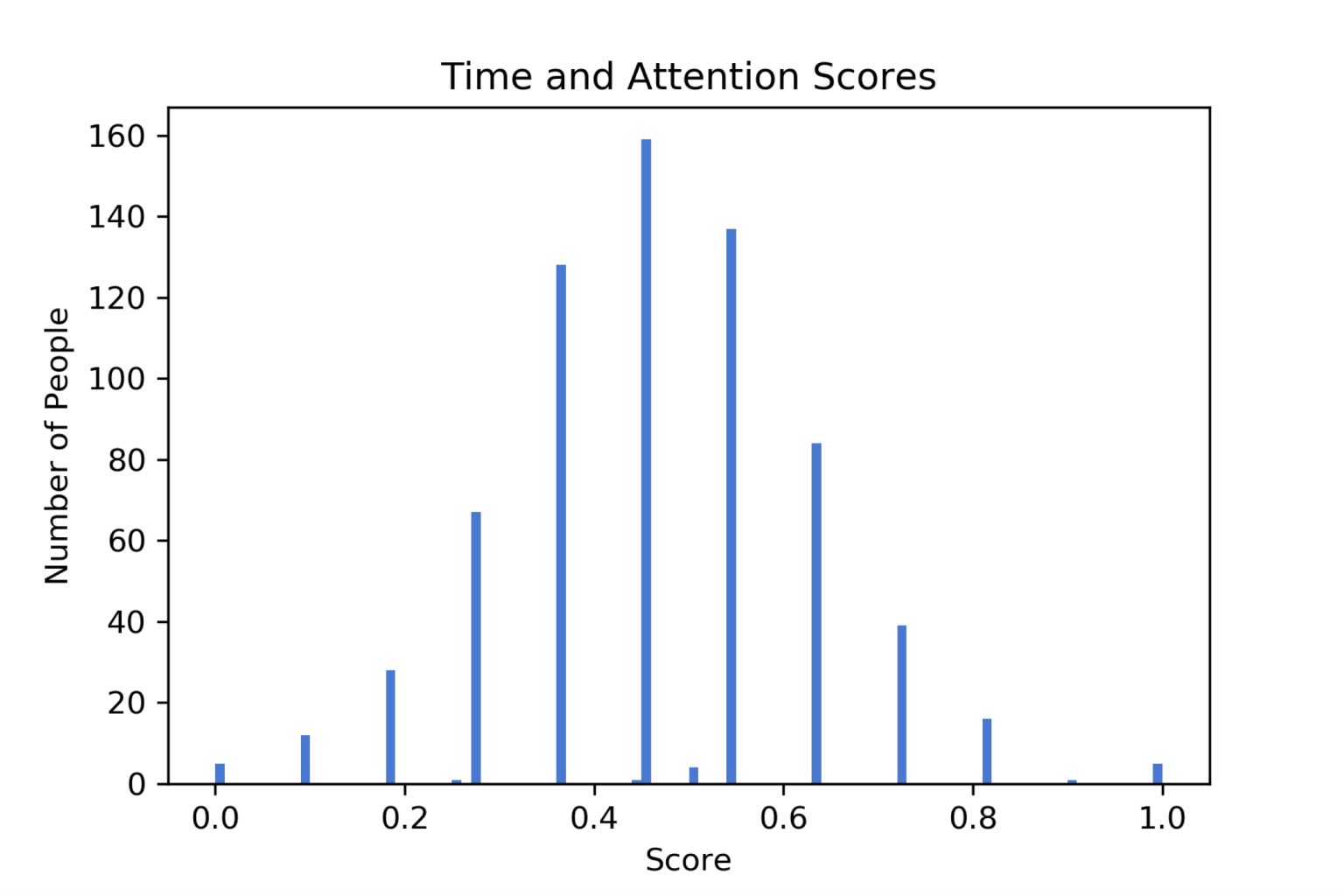
- We'll assume that the ideal customer is a senior (Age of 65+) who:
  - Exemplifies the belief that their learning pursuit will be fruitful and that a new experience can be rewarding. [Growth Mindset]
  - Has the time and attention necessary to play an involved and challenging game, potentially for multiple hours per day. [Time and Attention]
  - Engages frequently with the internet and technology and has access to the necessary hardware to play the game. [Technology Engagement]
- We measure each of these by creating a sum of ordinal responses to relevant survey questions.
- The customer fit score is a scaled sum of these 3 features.

For more information on how these are computed, including the specific survey questions, see the "Creating Target Customer-Fit Features" section of the methodology notebook.

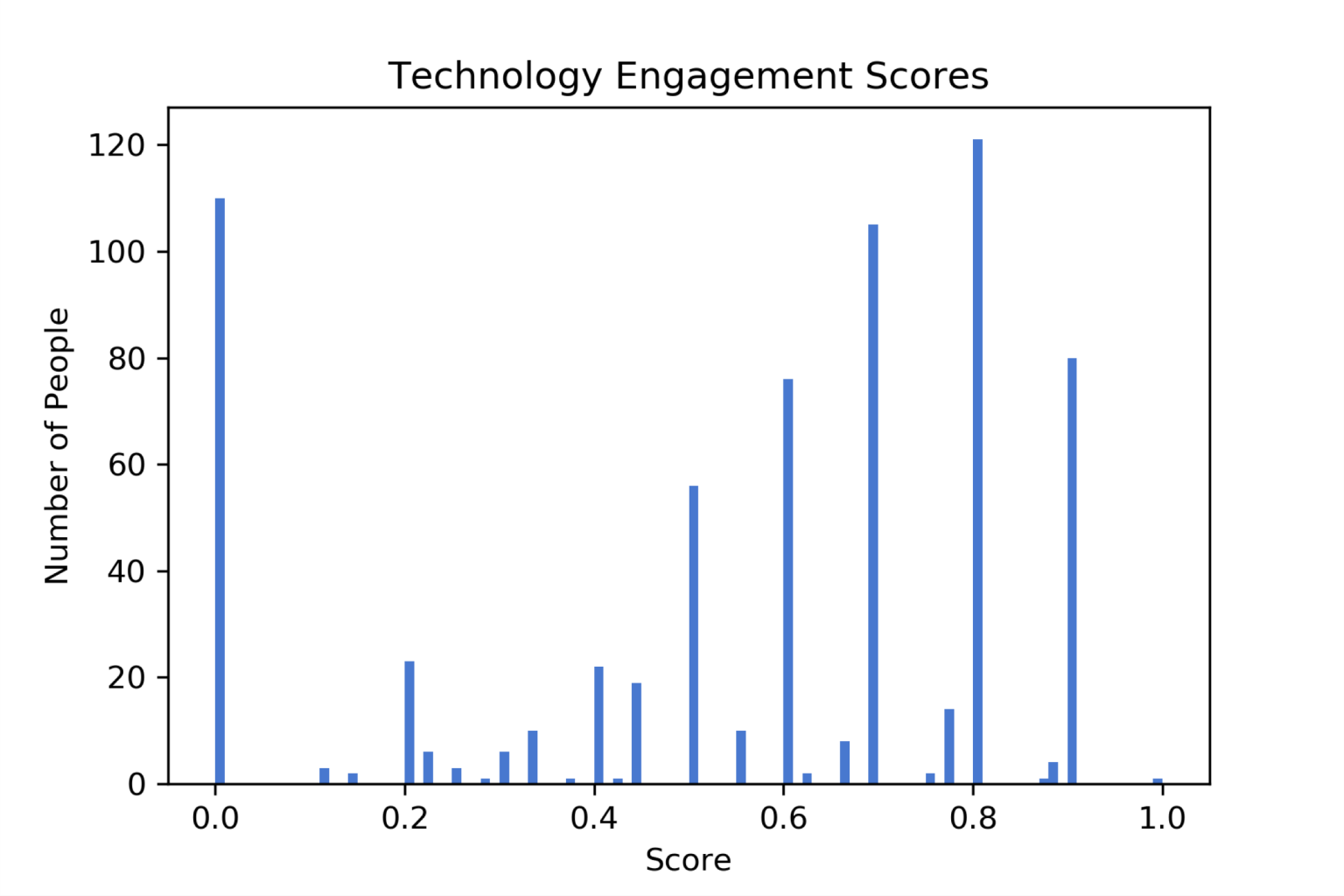
FEATURE DESIGN: GROWTH MINDSET



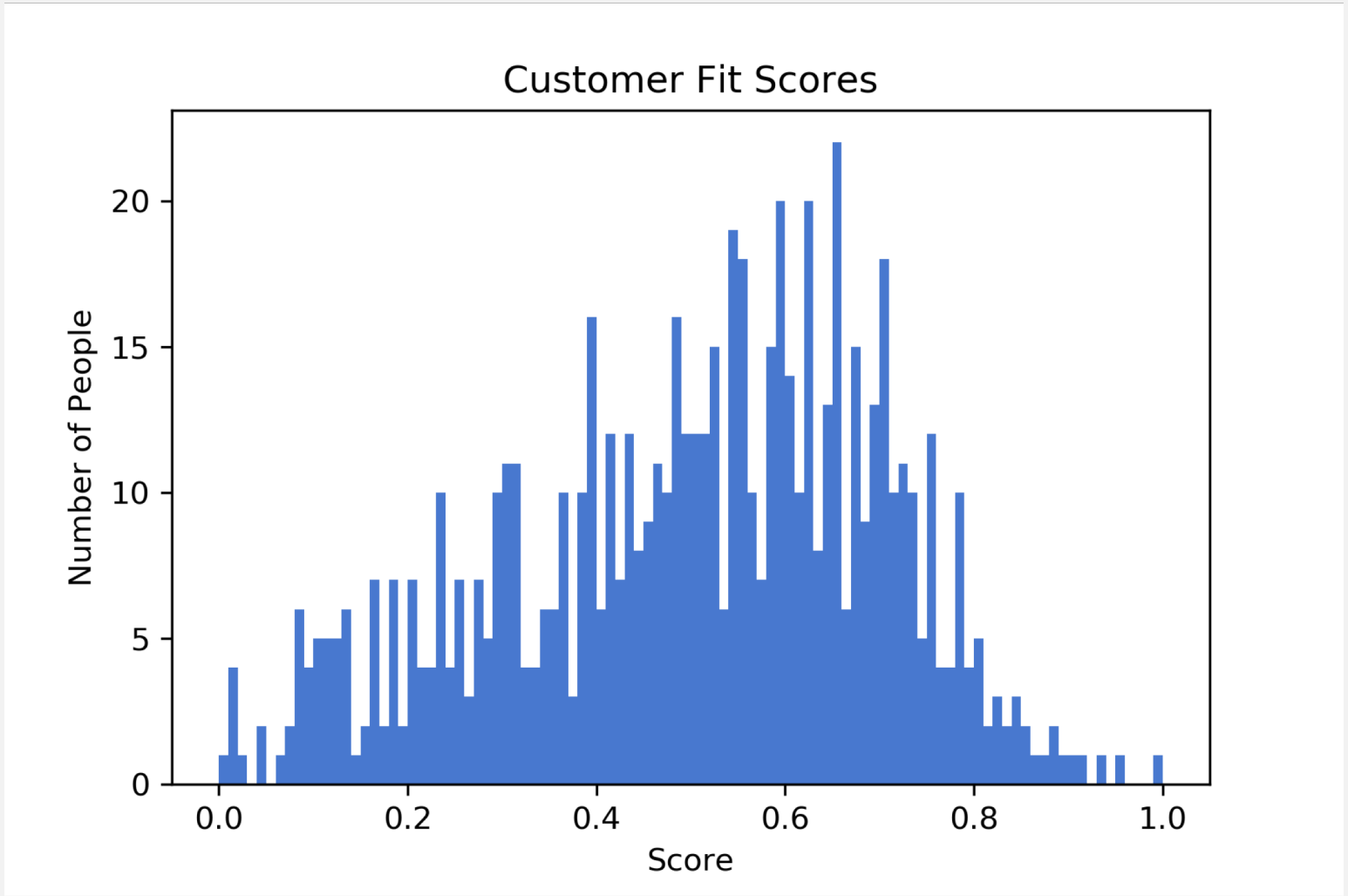
FEATURE DESIGN: TIME + ATTENTION



FEATURE DESIGN: TECH ENGAGEMENT



FEATURE DESIGN: CUSTOMER FIT

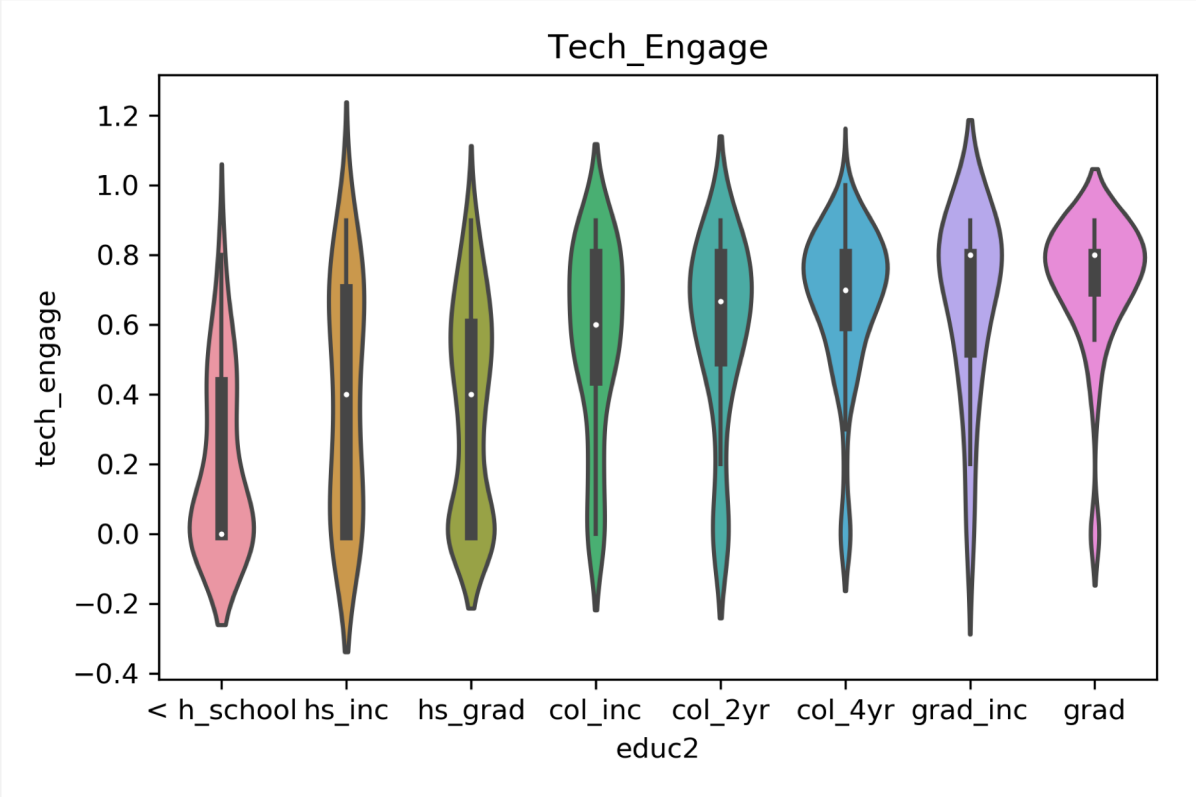


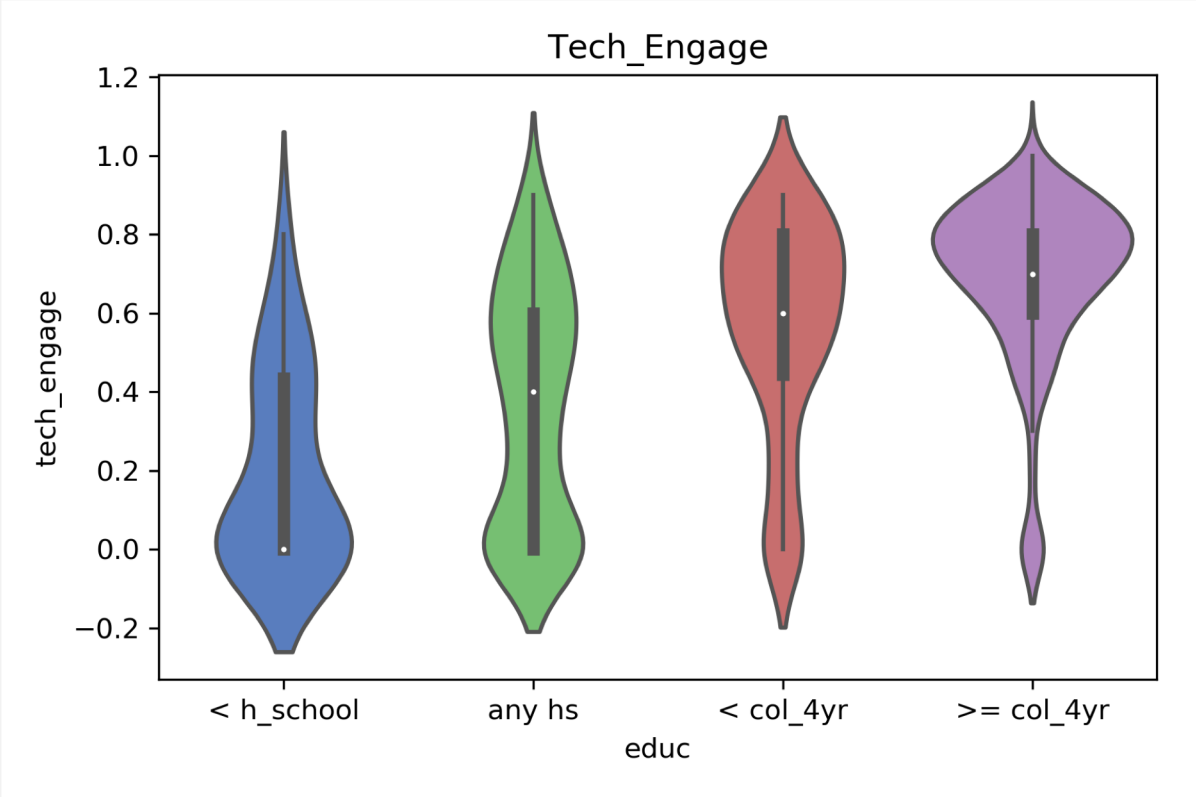
FEATURE DESIGN: DEMOGRAPHICS

- Initial covariates included:
  - Age, Sex, Education, Relationship Status, Race, and Political Ideology
- Dealing with categorical variables for linear regression is neat and nuanced.
- Each categorical was encoded as either:
  - Indicator variable - binary on/off, multi categories encoded as one-hot.
  - Ordinal variable – have a rank order and fit the "evenly spaced" assumption
- The evenly spaced assumption proved challenging for education level!

FEATURE DESIGN: DEMOGRAPHICS

Linearizing the effect of education level on technology engagement.

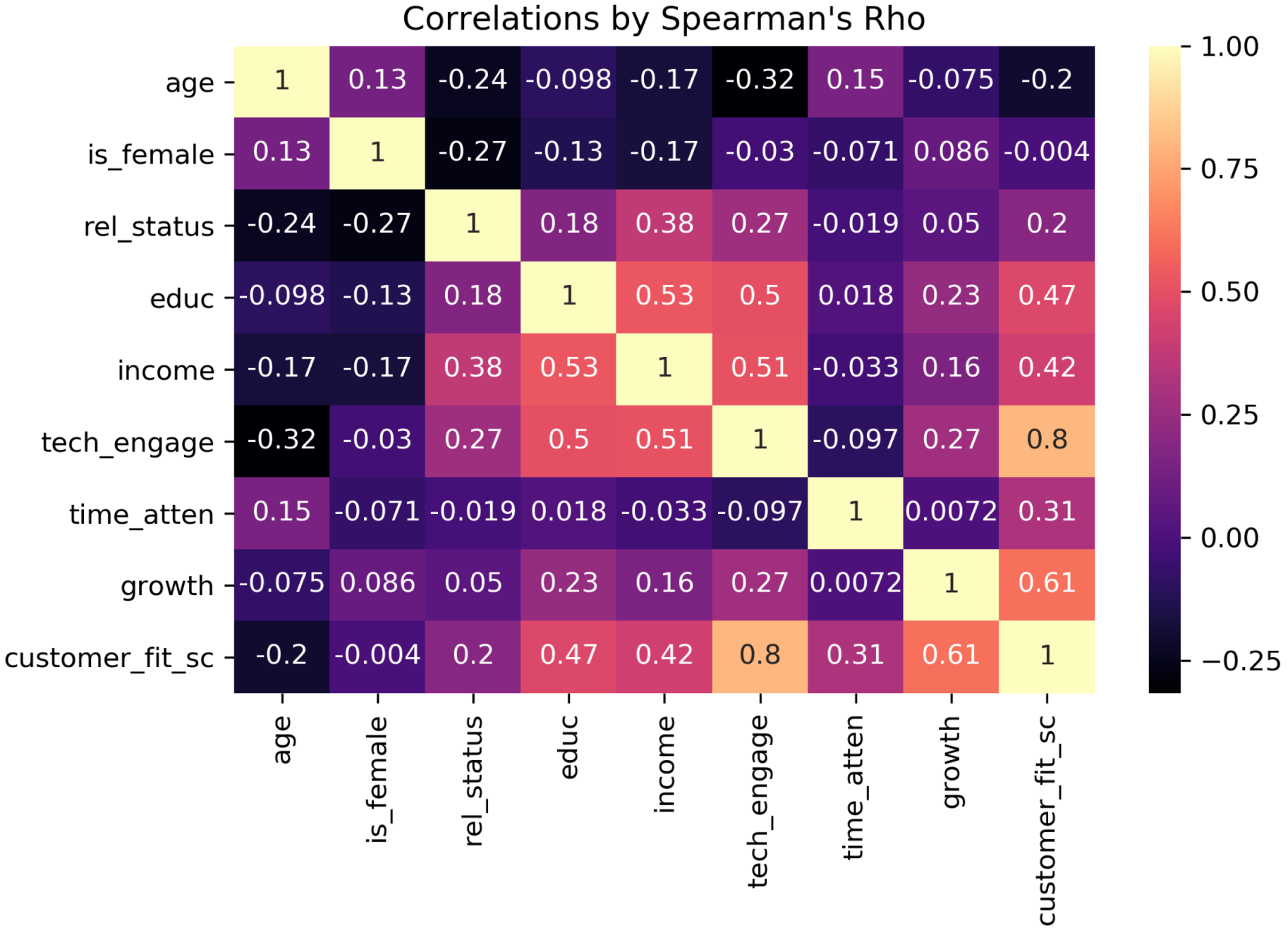




CORRELATION ANALYSIS

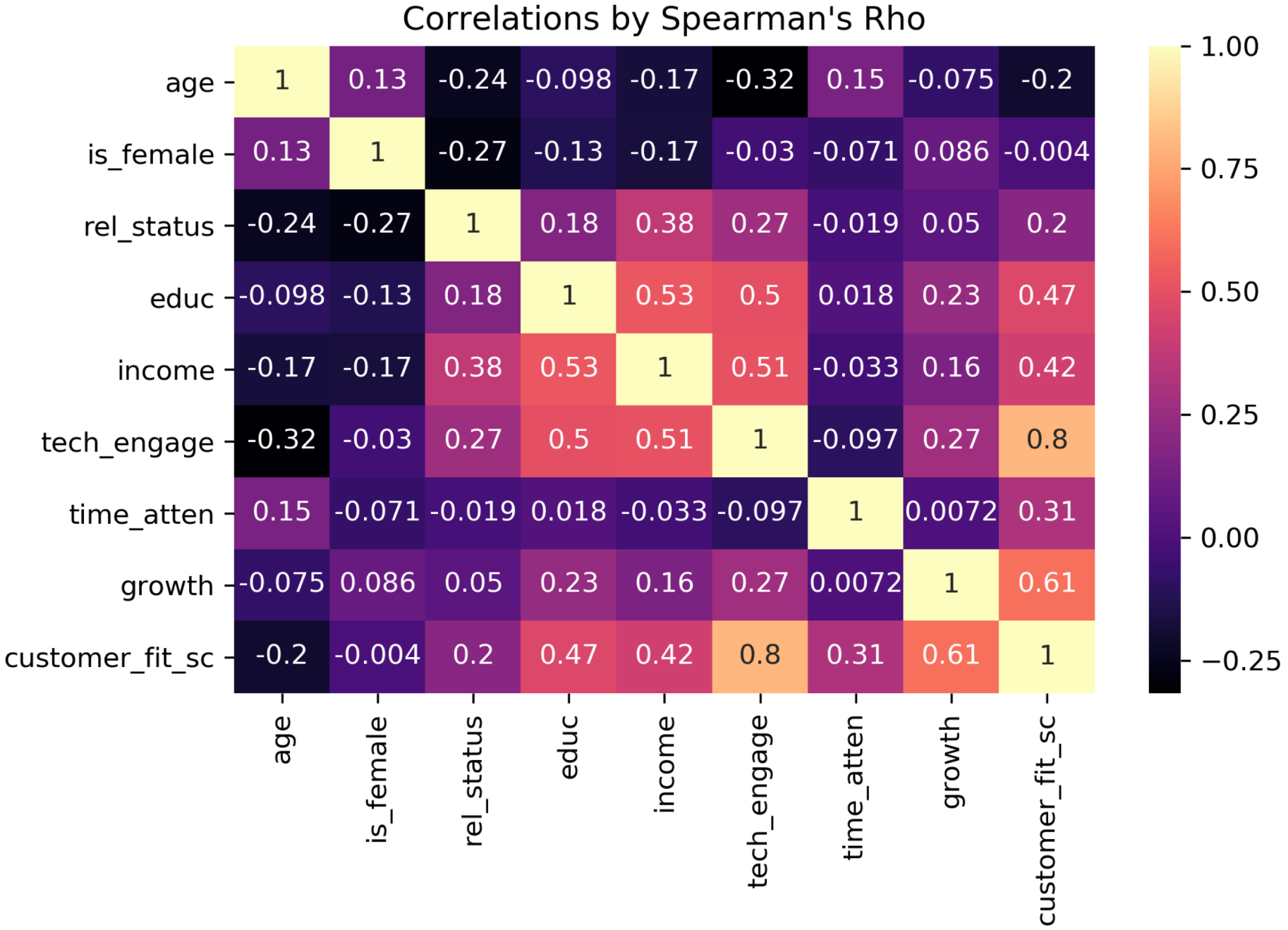
- In the following picture, note the following:
  - Income is positively correlated with education, relationship status.
  - Education and relationship status positively correlate to technology engagement.

  Since technology engagement is far more predictable than growth mindset or time and attention, income has a noted confounding effect on the final analysis.


Correlations by Spearman's Rho
age 1 0.13 -0.24 -0.098 -0.17 -0.32 0.15 -0.075 -0.2
is_female 0.13 1 -0.27 -0.13 -0.17 -0.03 -0.071 0.086 -0.004
rel_status -0.24 -0.27 1 0.18 0.38 0.27 -0.019 0.05 0.2
educ -0.098 -0.13 0.18 1 0.53 0.5 0.018 0.23 0.47
income -0.17 -0.17 0.38 0.53 1 0.51 -0.033 0.16 0.42
tech_engage -0.32 -0.03 0.27 0.5 0.51 1 -0.097 0.27 0.8
time_atten 0.15 -0.071 -0.019 0.018 -0.033 -0.097 1 0.0072 0.31
growth -0.075 0.086 0.05 0.23 0.16 0.27 0.0072 1 0.61
customer_fit_sc -0.2 -0.004 0.2 0.47 0.42 0.8 0.31 0.61 1
age
is_female
rel_status
educ
income
tech_engage
time_atten
growth
customer_fit_sc
1.00
0.75
0.50
0.25
0.00
−0.25

CORRELATION ANALYSIS

- Next look for the following:
  - Age is negatively correlated with technology engagement and positively correlated with time and attention. This ends up with a net negative on customer fit.
  - Being female is positively correlated with growth mindset but negatively with income and technology engagement, perhaps through the confounder of age. Gender has a complex effect on customer fit.
  - These two variables have interesting and useful correlates with the behaviors and beliefs we care about. They also bring out important social considerations.


Correlations by Spearman's Rho
age 1 0.13 -0.24 -0.098 -0.17 -0.32 0.15 -0.075 -0.2
is_female 0.13 1 -0.27 -0.13 -0.17 -0.03 -0.071 0.086 -0.004
rel_status -0.24 -0.27 1 0.18 0.38 0.27 -0.019 0.05 0.2
educ -0.098 -0.13 0.18 1 0.53 0.5 0.018 0.23 0.47
income -0.17 -0.17 0.38 0.53 1 0.51 -0.033 0.16 0.42
tech_engage -0.32 -0.03 0.27 0.5 0.51 1 -0.097 0.27 0.8
time_atten 0.15 -0.071 -0.019 0.018 -0.033 -0.097 1 0.0072 0.31
growth -0.075 0.086 0.05 0.23 0.16 0.27 0.0072 1 0.61
customer_fit_sc -0.2 -0.004 0.2 0.47 0.42 0.8 0.31 0.61 1
age
is_female
rel_status
educ
income
tech_engage
time_atten
growth
customer_fit_sc
1.00
0.75
0.50
0.25
0.00
−0.25

MODELING CUSTOMER FIT

Recall That:

- We'll assume that the ideal customer is a senior (Age of 65+) who:
  - Exemplifies the belief that their learning pursuit will be fruitful and that a new experience can be rewarding. [Growth Mindset]
  - Has the time and attention necessary to play an involved and challenging game, potentially for multiple hours per day. [Time and Attention]
  - Engages frequently with the internet and technology and has access to the necessary hardware to play the game. [Technology Engagement]
- We measure these by creating a sum of ordinal responses to survey questions.
- A weighted sum of these three scores makes the customer fit score.
- **A full key for which questions were utilized in each feature is available in the methodology notebook under "Creating Target Customer-Fit Features."**

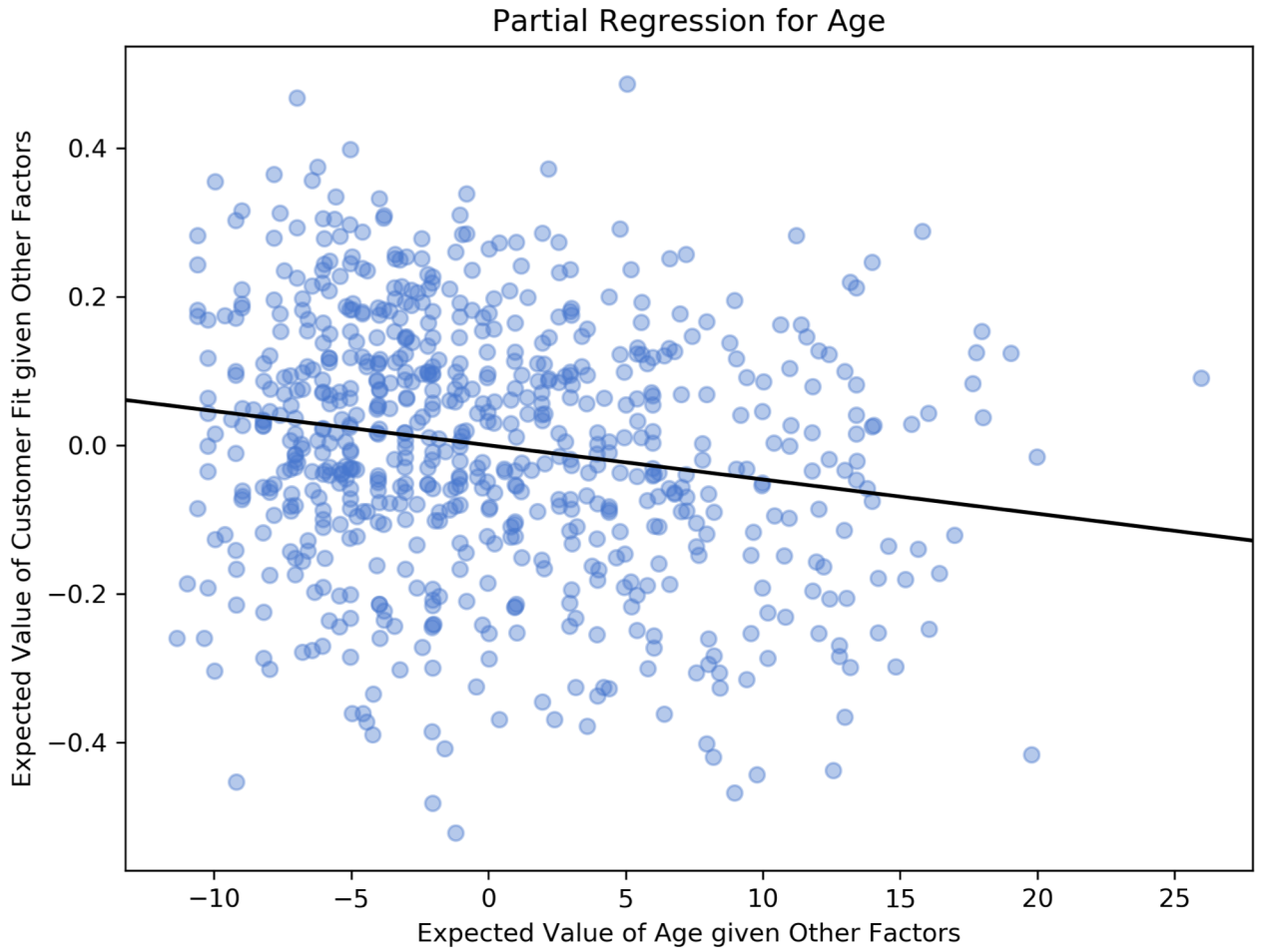
COVARIATE SELECTION FOR REGRESSION

- What better way to select covariates for linear regression than with linear regression?
  - Lasso shrinks the coefficients of less relevant predictors towards zero with a regularization term, alpha, that acts on the L1 norm of the coefficients.
  - Alpha works as a threshold for which covariates have their coefficients shrunk to zero.
- With alpha of 0.002, we decide to keep age, sex, domestic partnership, and education. We toss out race and political ideology as less useful covariates.

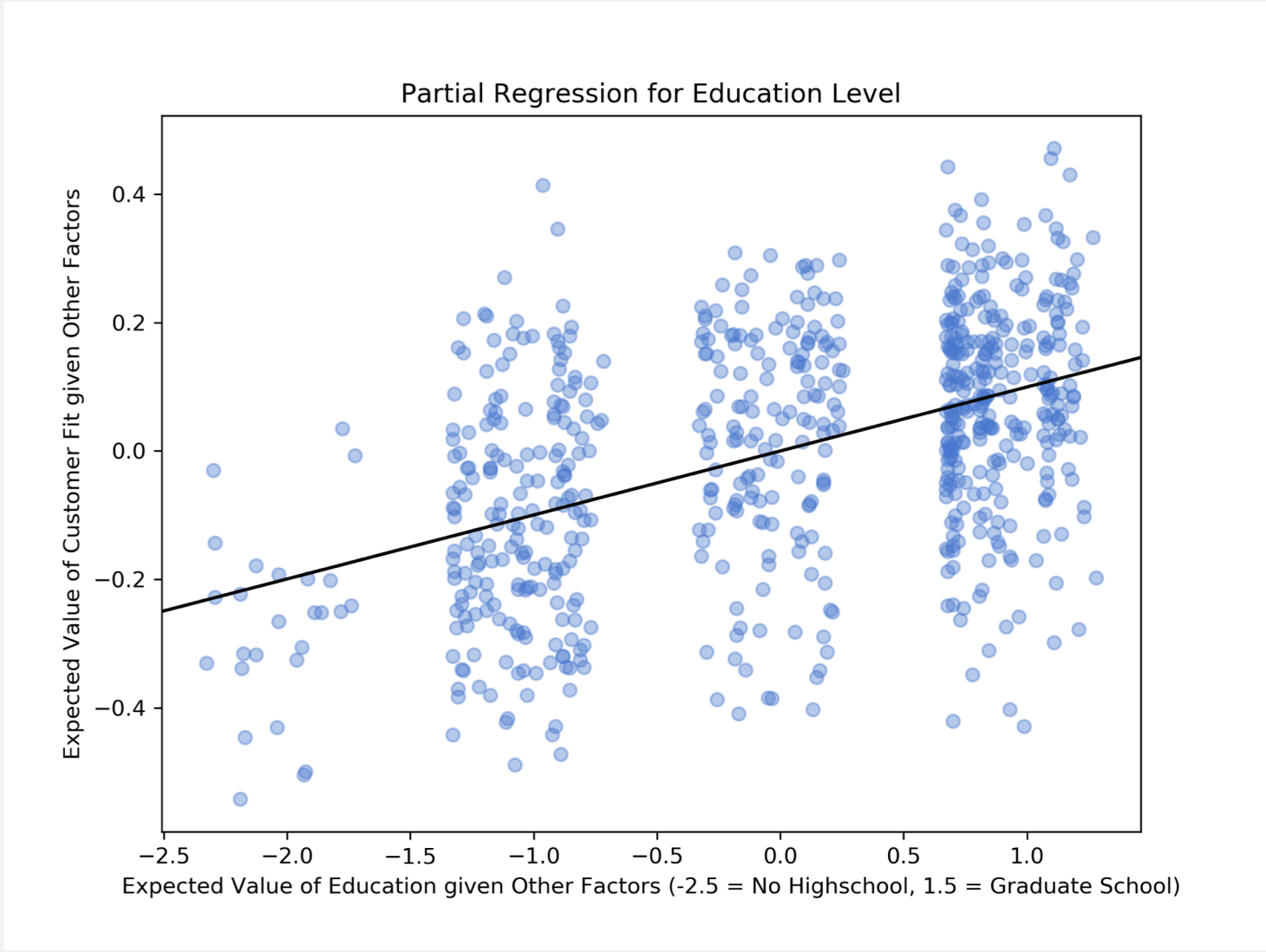
LINEAR REGRESSION MODEL

- Ordinary Least Squares method was employed, using StatsModels
- Statistically significant covariates were kept and others discarded.
- The target variable, Cust_Fit ranges from 0 to 1.
- Cust_Fit = 0.58 – 0.005 * Age + 0.04 * Is_Female + 0.04 * Relationship + 0.01* Education
- Largest coefficients are Education and Age (these variables have greater scale than the others)
- R-squared = 0.284
- Because we want *directional* insight, optimizing predictive power does not yield a more valuable recommendation.

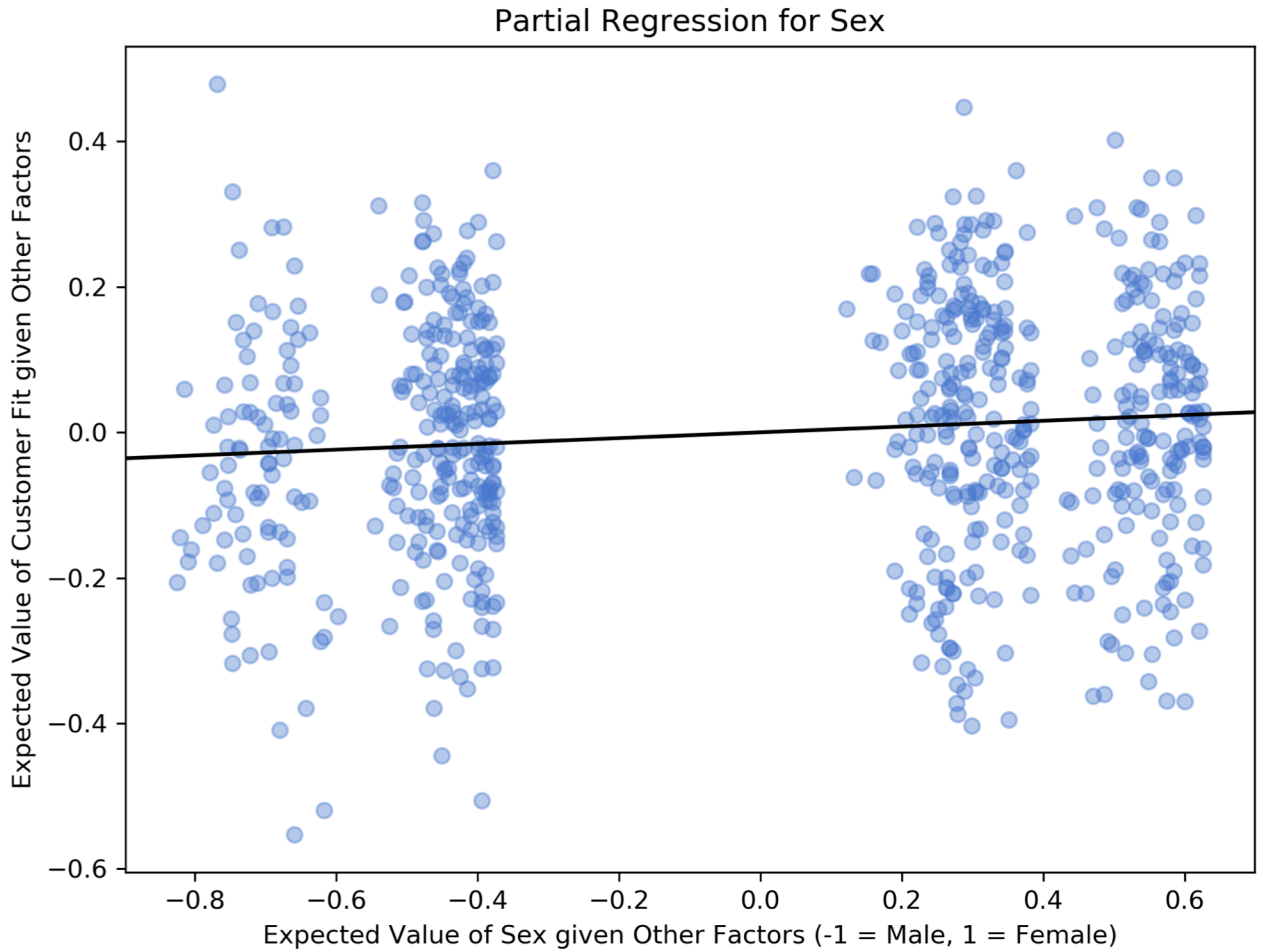
PARTIAL REGRESSION FOR AGE



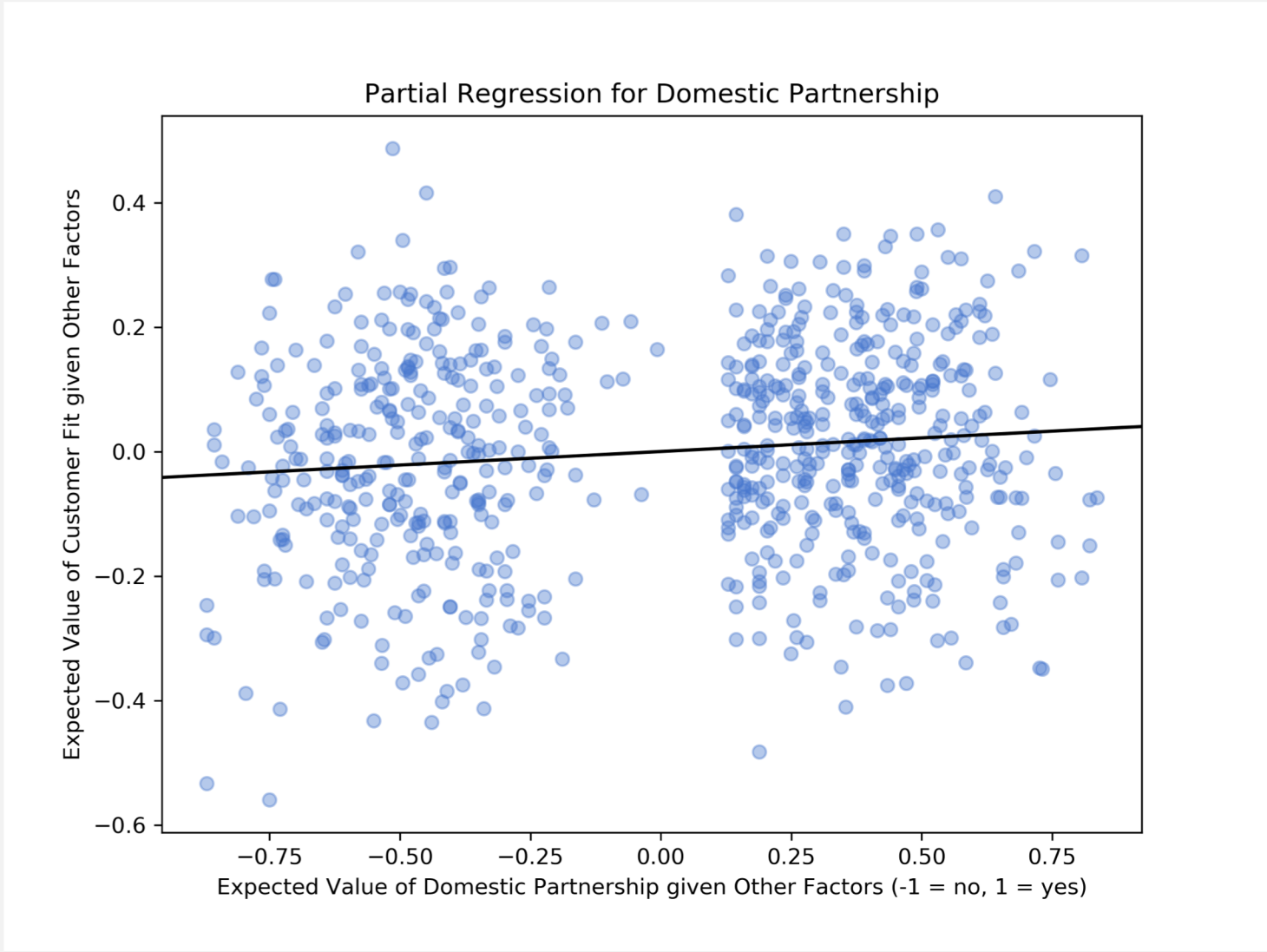
PARTIAL REGRESSION FOR EDUCATION



PARTIAL REGRESSION FOR SEX

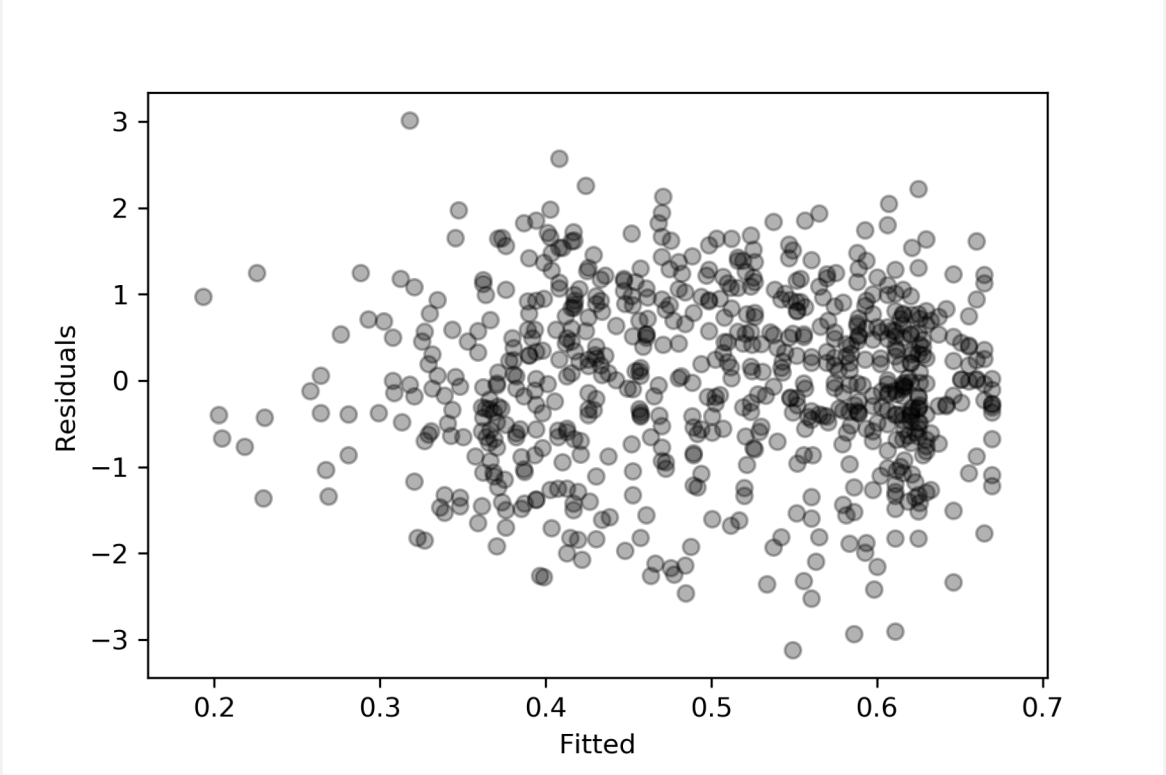


PARTIAL REGRESSION FOR DOMESTIC PARTNERSHIP



MODEL VALIDATION

- Standardized residuals were normally distributed on visual inspection.
- There conditional distribution of residuals to fitted value appears normal.
- No individual data points had unusual leverage.



MODELING OBSERVATIONS

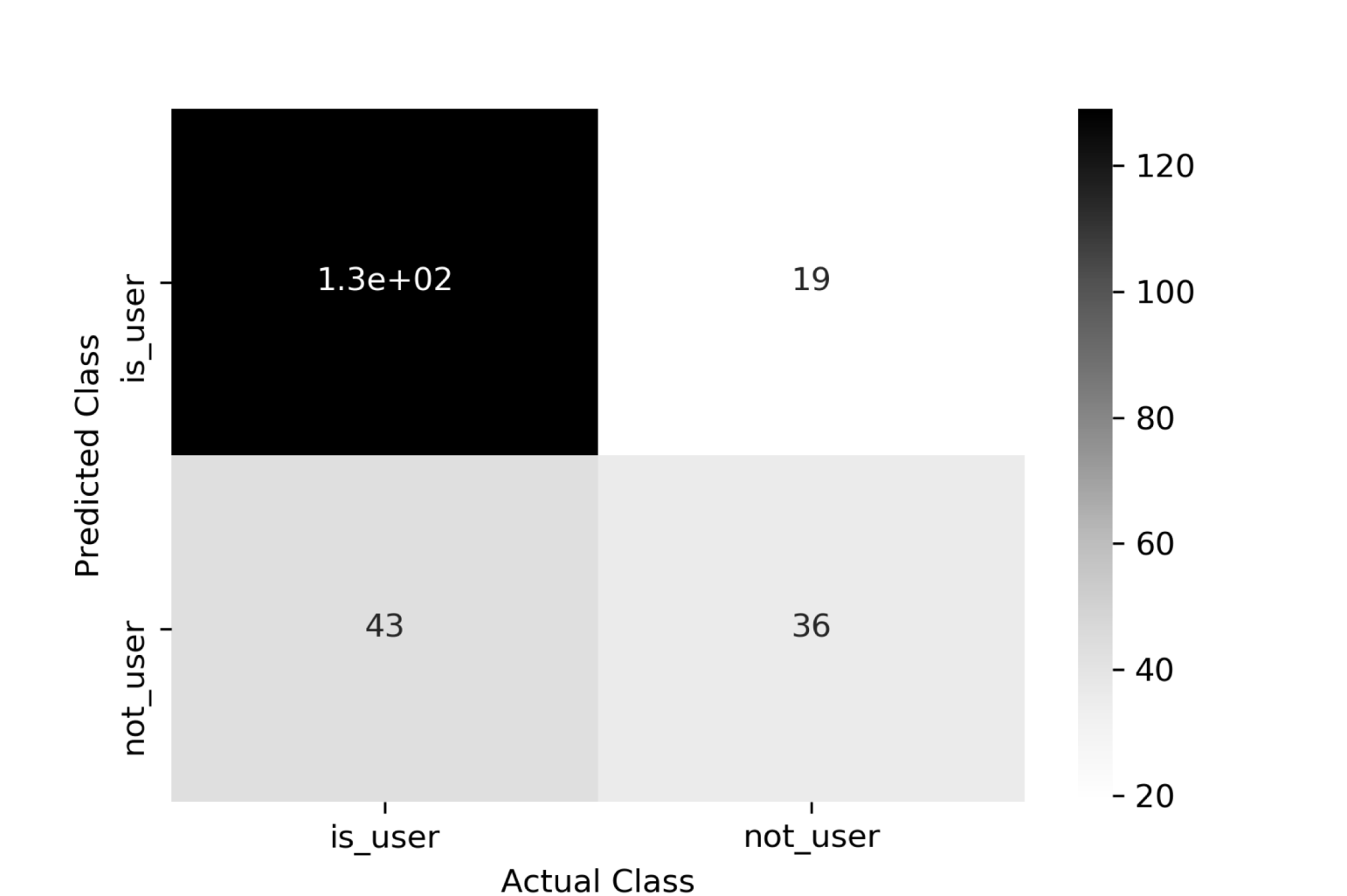
- We saw that greater income correlates positively with being in a domestic partnership (spearman's rho value of 0.38) and education level (rho of 0.53).
- Demographics explained 36% of the variance in technology engagement, but 8% of the variance in growth mindset 2.5% of the variance in time and attention.
- As a consequence, much of what we observe in the final results are effects related to wealth, which we *assume* will not be available to marketers.
- If we could measure wealth, how would we want to action on it?
- What is our attitude towards marketing to the wealthy?

NOW FOR A BRIEF LOOK AT A SECOND MODEL…

CLASSIFYING FOR MOBILE INTERNET USE BY DEMOGRAPHIC

- Logistic Regression is able to classify whether or not a senior uses the internet on a mobile device (at least occasionally) with 72% accuracy using education, age, sex, and domestic partnership as covariates.
- The marginal effect of each covariate, if the covariates were of equal scale are:
  - Education: 0.57
  - Age: -0.36
  - Sex: 0.06
  - Partnership: 0.14
- The orientation and relative weight of each covariate is the same as in our regression model for fit.
- This model reinforces our qualitative understanding of how these factors relate to technology engagement.

LOGISTIC REGRESSION CONFUSION MATRIX



RECOMMENDATIONS TO EDUGAMES.ORG

- We have high confidence that younger (ages 65-70), college-educated seniors are more likely to fit the product.
  - We encourage Edugames to begin it's marketing process by targeting these groups.
  - Social benefit could be derived by a second "outreach" marketing effort towards lower income and less well-educated seniors.
- We find that being female and being in a domestic partnership (relationship with cohabitation or marriage) may indicate a slightly better fit to the product.
  - We encourage Edugames to utilize split testing to estimate the effect of these variables in early marketing efforts.
  - The observed effect of domestic partnerships on customer fit may be due to wealth differences.

THANK YOU!

Check out the full repository for more information:

https://github.com/swfisher/springboard/tree/master/capstone_2/submission

Feel free to get in touch.