

# 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.

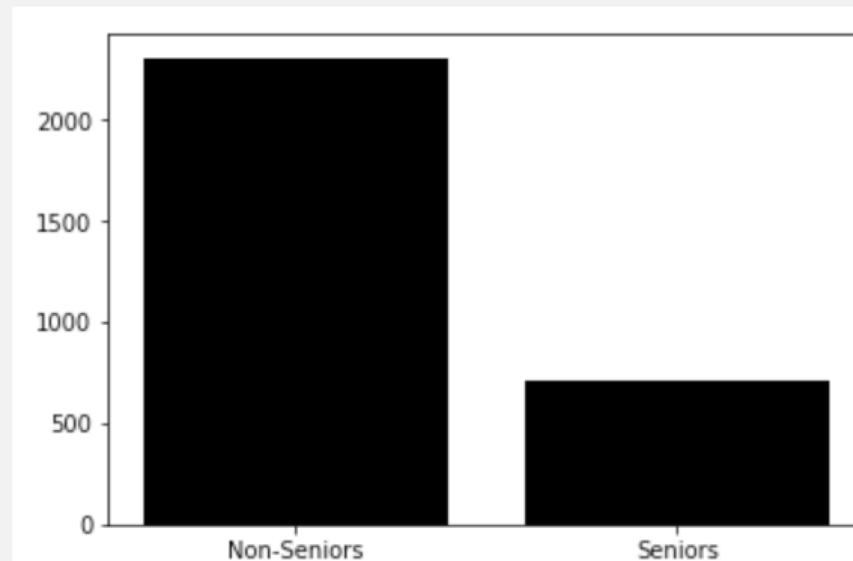
- A. Introduce the business problem of our fictional client, Edugames.org
- B. Introduce public survey data to help us solve the problem
- C. Present our recommendations to Edugames based on analysis
- D. Step through the process for arriving at those recommendations

## 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 2308 Non-seniors (age  $\leq 65$ ) and 707 Seniors (age  $> 65$ )
- Questions On:
  - Time and Attention
  - Growth Mindset
  - Technology Engagement
  - Demographic Information



# THE IDEAL CUSTOMER

- We'll assume that the ideal customer is a senior (Age of 66+) 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.

## RECOMMENDATIONS

- We have high confidence that younger (ages 66-71), college-educated seniors are more likely to fit the product.
  - We encourage Edugames to begin its 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.

# ANALYSIS METHODOLOGY

- Where did the results we just saw come from?
  - A. Data cleaning
  - B. Feature Engineering
  - C. Correlation Analysis
  - D. Linear Regression Model

# DATA CLEANING

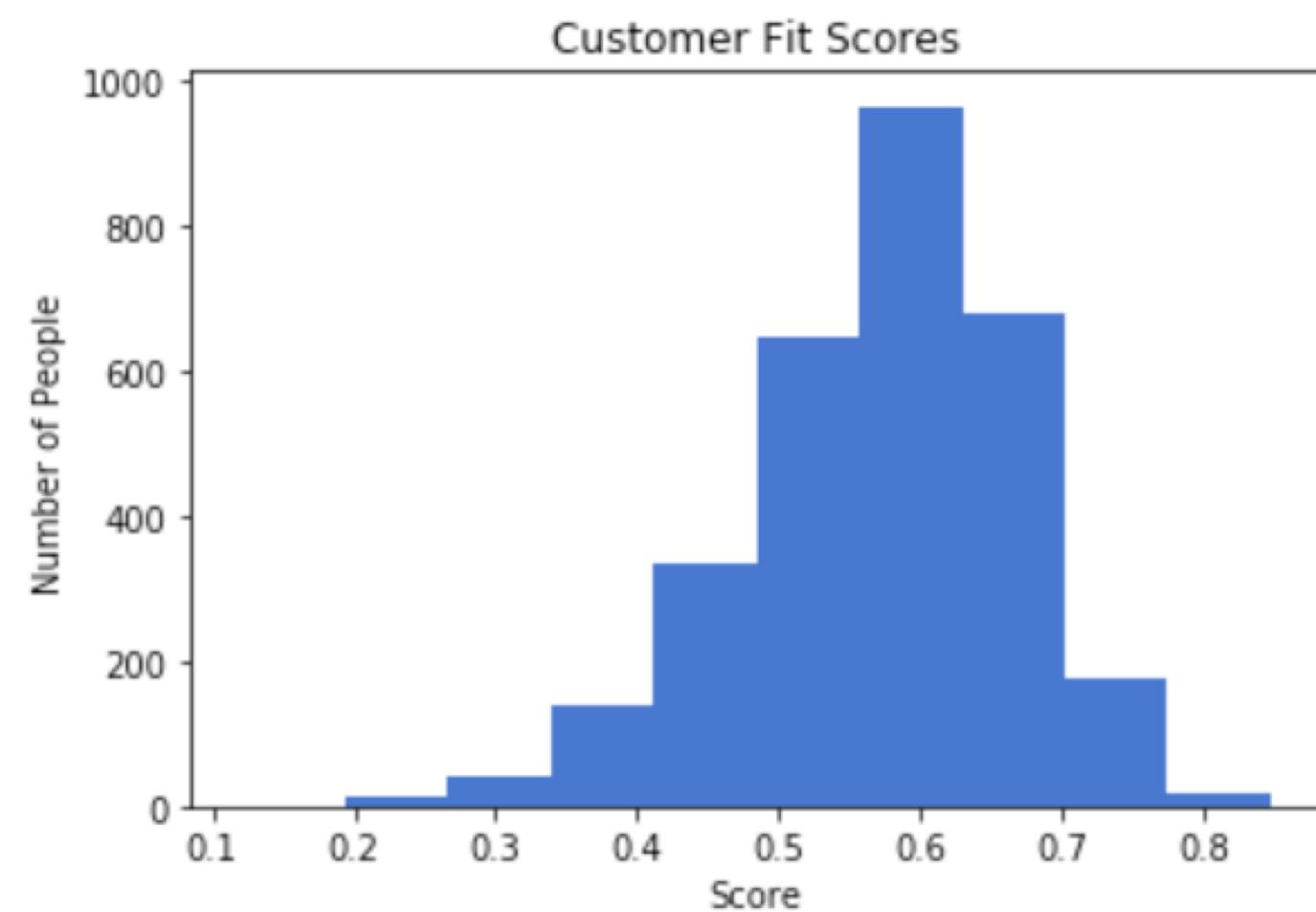
- 117 features encoding categorical responses to individual questions.
- No missing values – very clean data set.
- Some empty strings if a question is only asked under a condition.
- Some non-responses- either “I don’t know” or “I don’t want to tell you.”
- Non-response was dealt with in three ways.
  - Selecting a subset of complete responses for demographics (about 89% of seniors)
  - Giving 0 points for no response in target features. (assumes equal distribution)

# FEATURE DESIGN: CUSTOMER FIT

## Recall That:

- We'll assume that the ideal customer is a senior (Age of 66+) 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 used and how weightings were made is available in the methodology notebook.**

# FEATURE DESIGN: CUSTOMER FIT

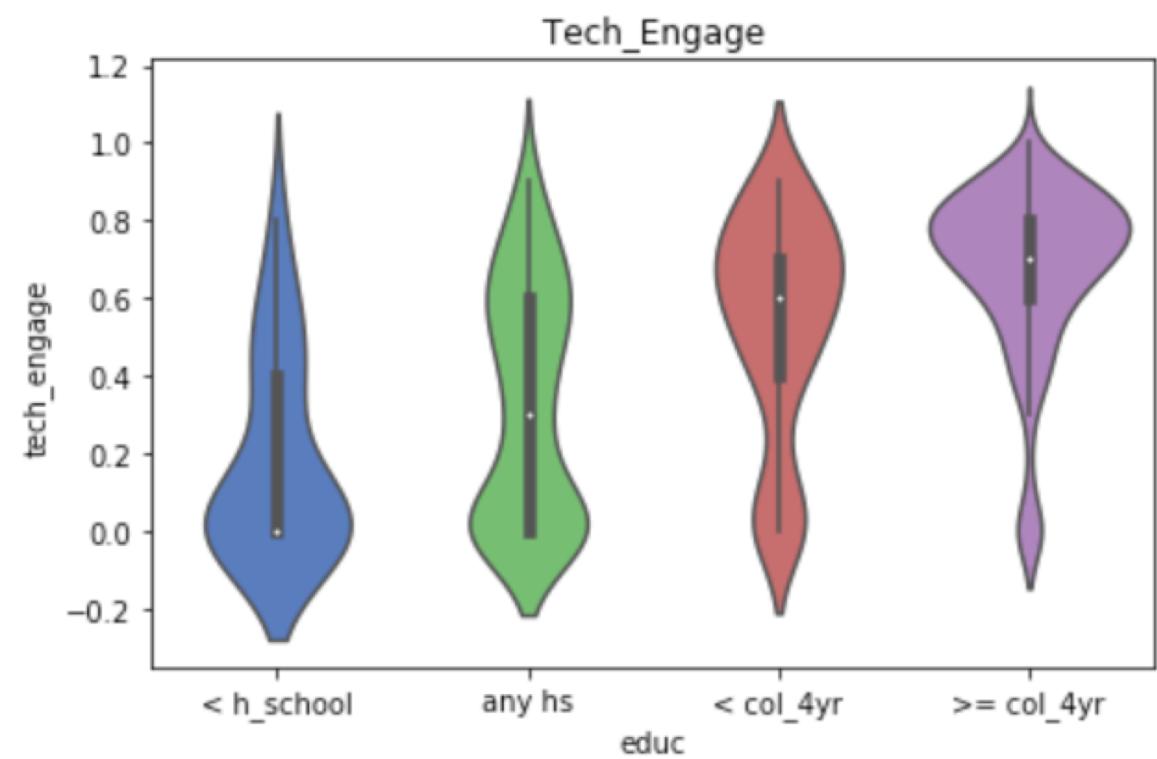
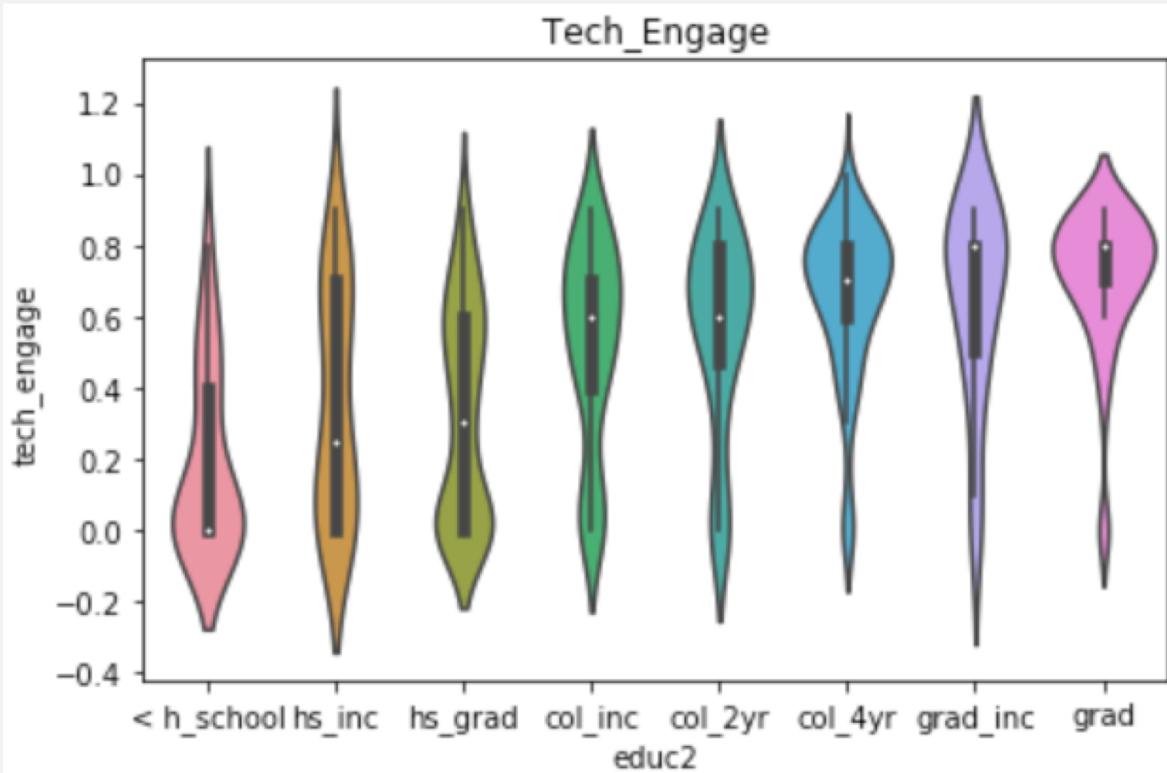


## FEATURE DESIGN: DEMOGRAPHICS

- 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.

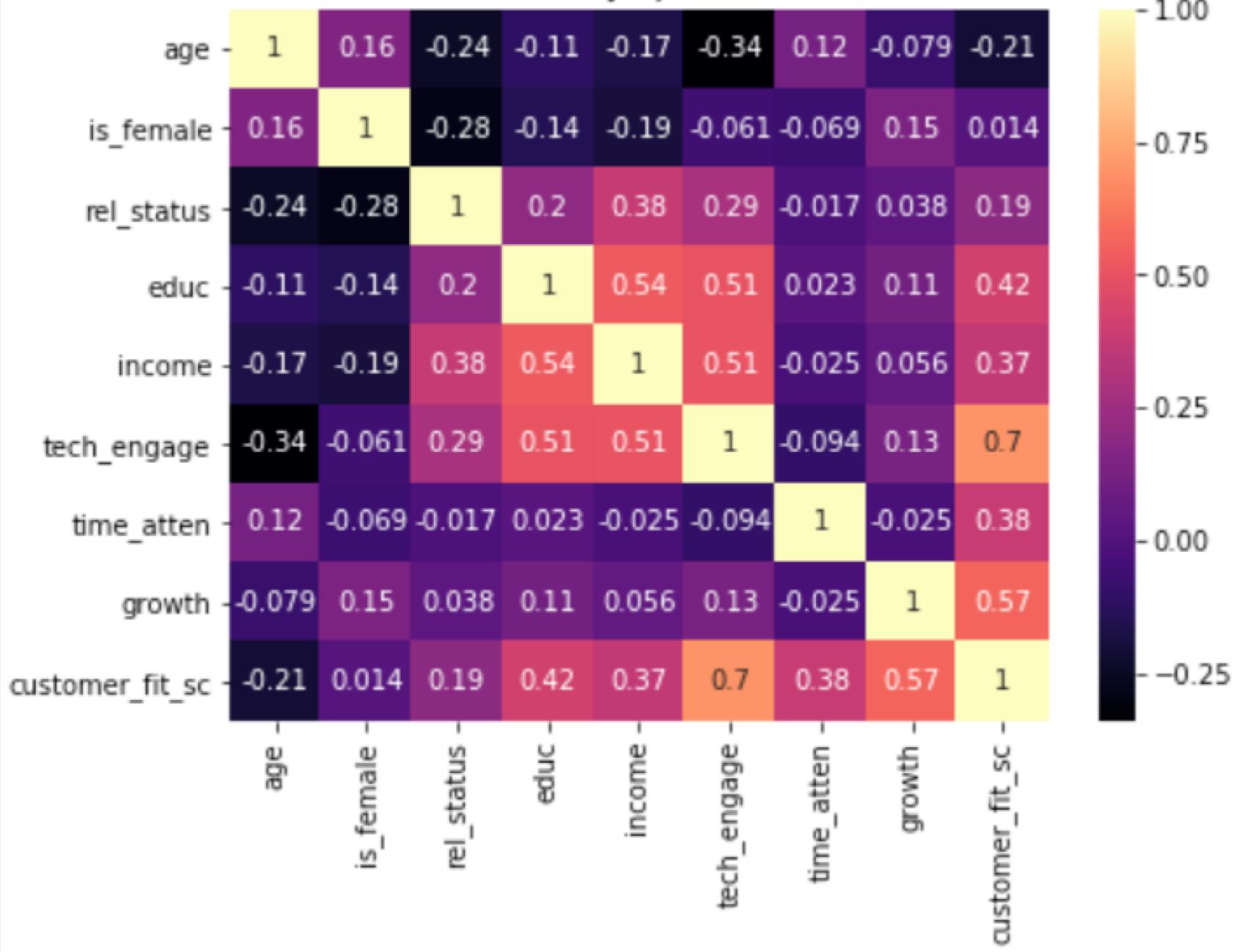


# CORRELATION ANALYSIS

- 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

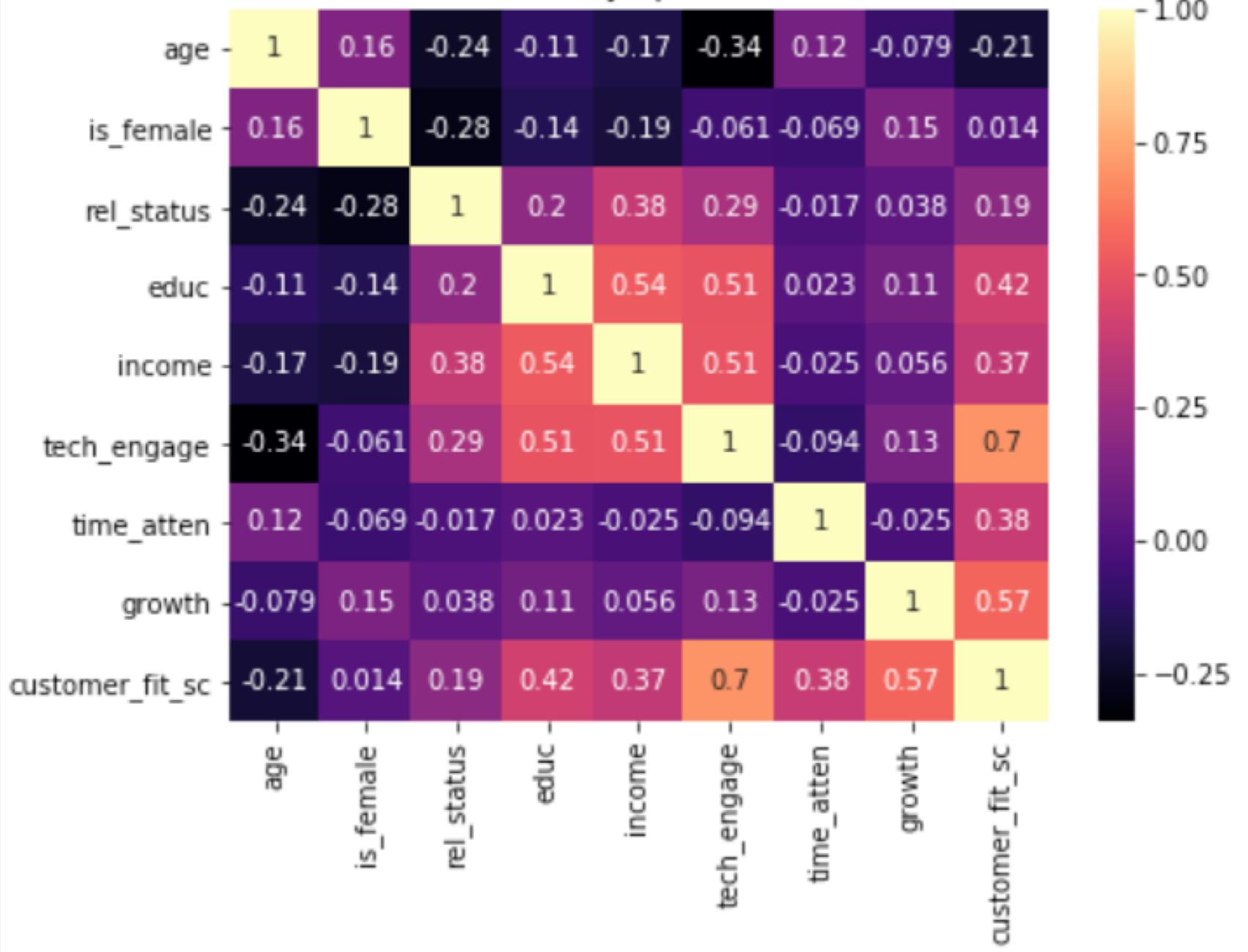


# CORRELATION ANALYSIS

- 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. This ends up with a small net positive 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

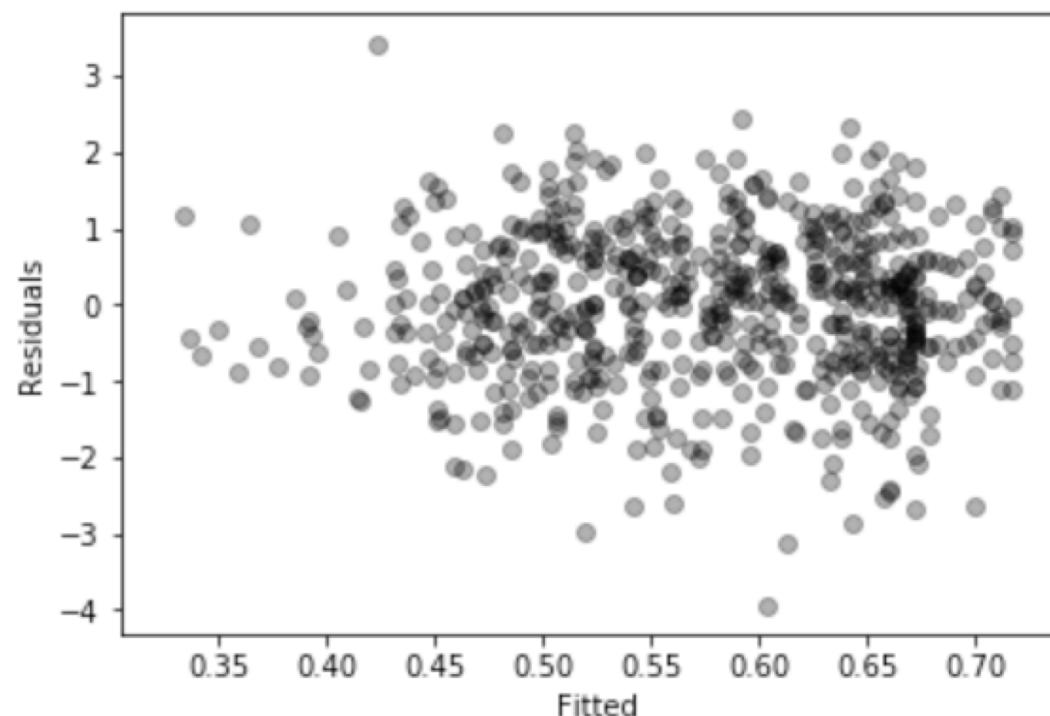


## LINEAR REGRESSION MODEL

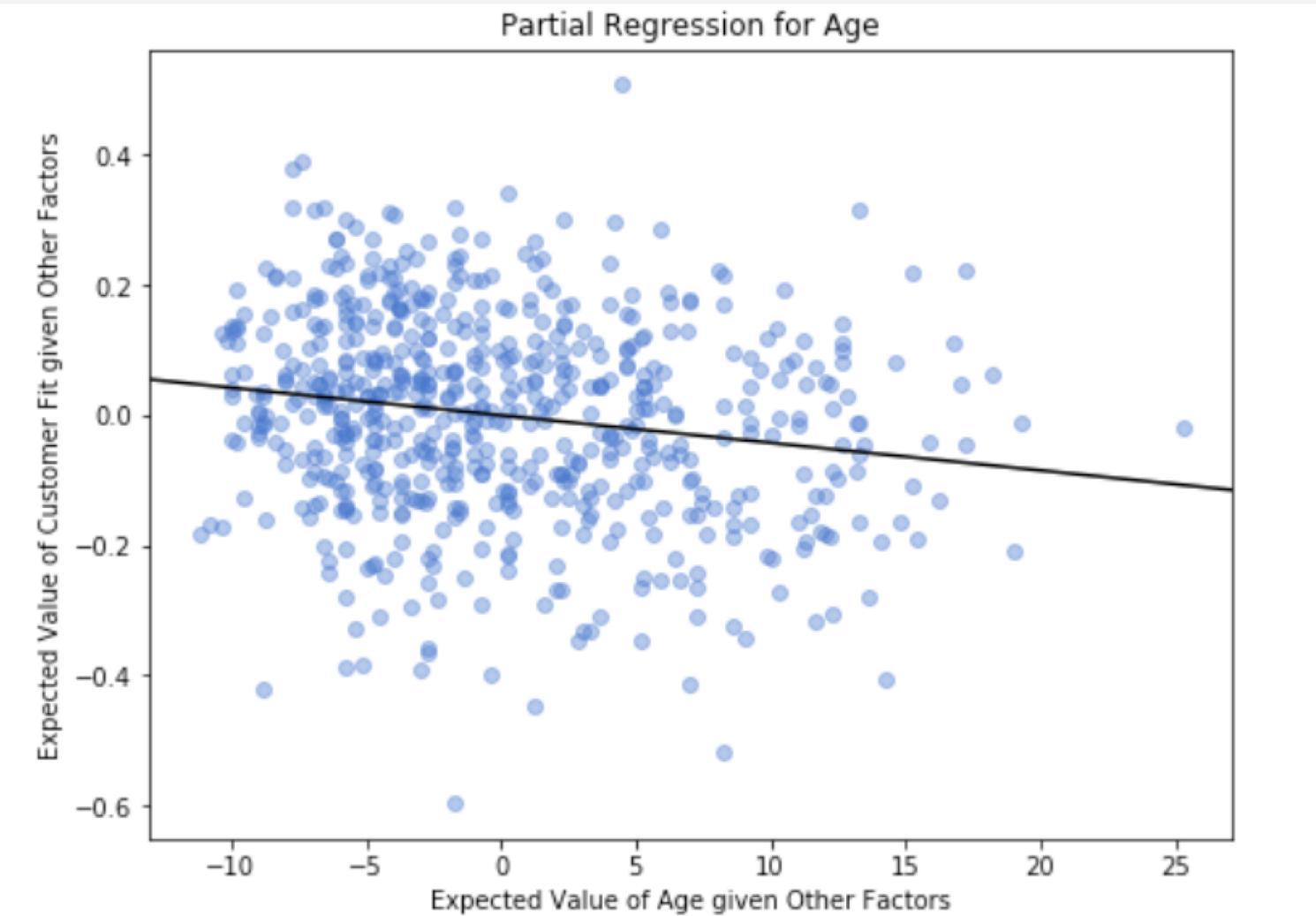
- Ordinary Least Squares method was employed, using StatsModels
- Statistically significant covariates were kept and others discarded.
- The target variable, Cust\_Fit ranges from 0 to 1.
- $Cust\_Fit = C - 0.004 * Age + 0.04 * Is\_Female + 0.039 * Relationship + 0.074 * Education$
- R-squared = 0.226
- Because we want *directional* insight, optimizing predictive power does not yield a more valuable recommendation.

# MODEL VALIDATION

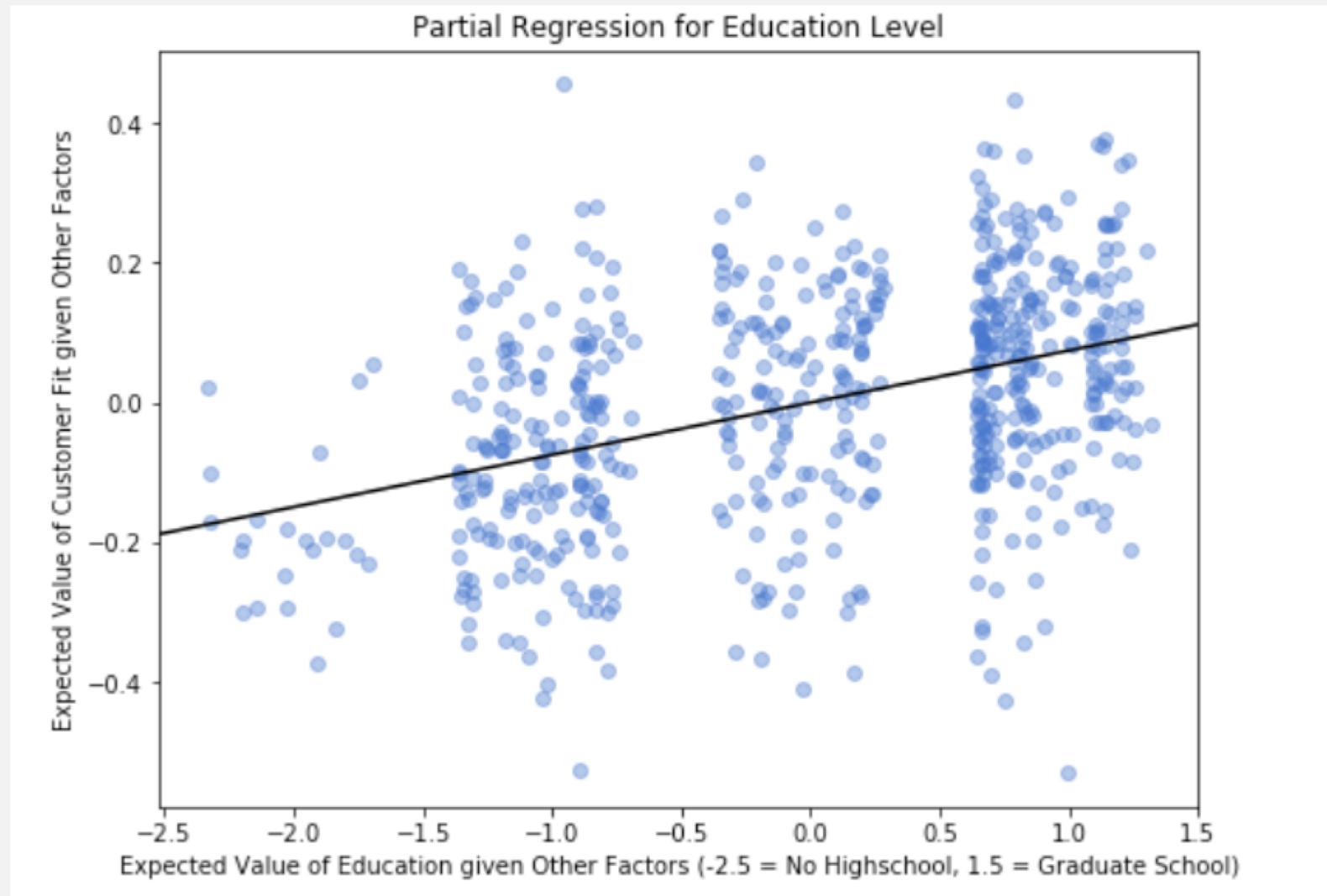
- Standardized residuals were normally distributed on visual inspection.
- The conditional distribution of residuals to fitted value appears normal.
- No individual data points had unusual leverage.



# PARTIAL REGRESSION FOR AGE



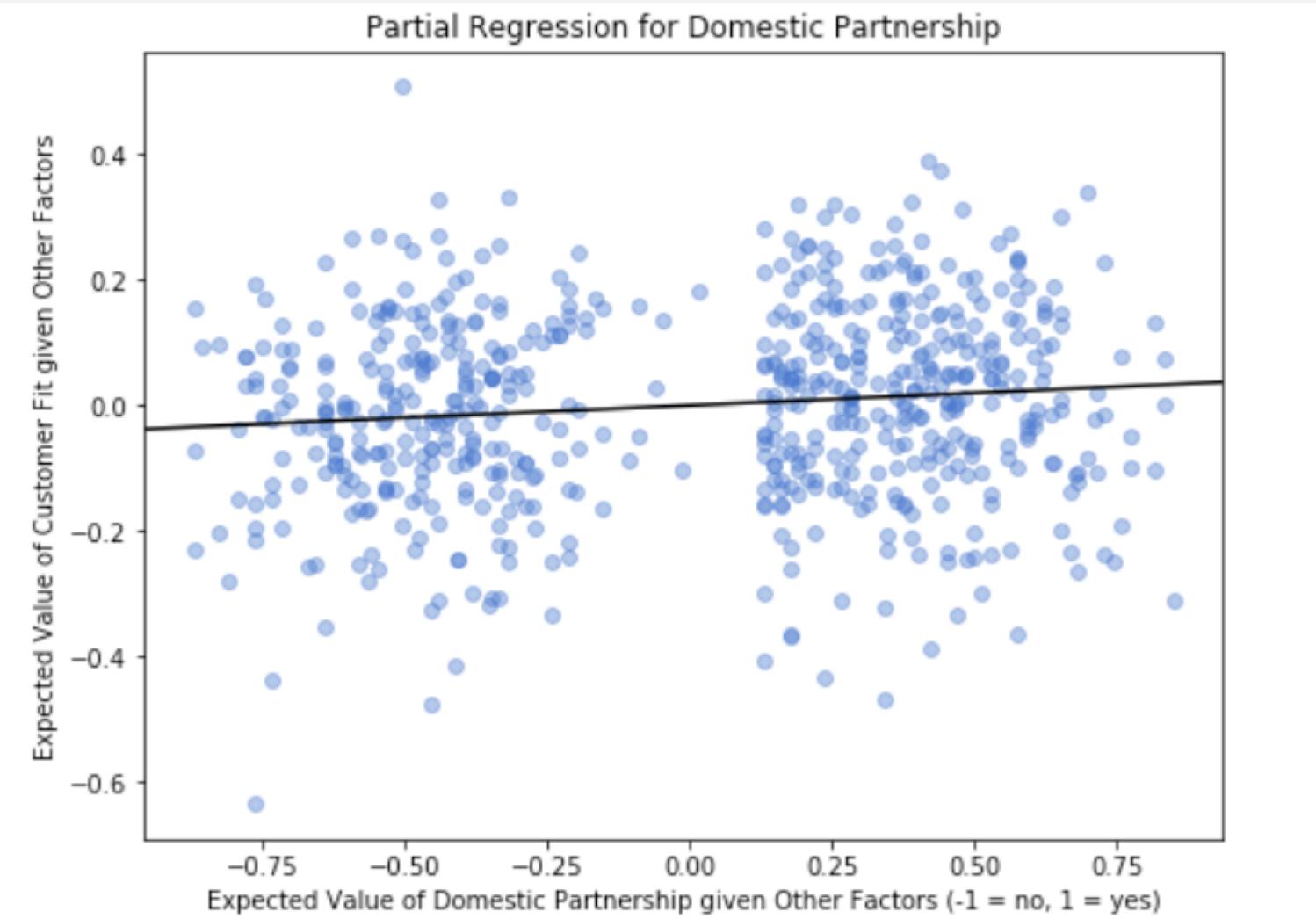
# PARTIAL REGRESSION FOR EDUCATION



# PARTIAL REGRESSION FOR SEX



# PARTIAL REGRESSION FOR DOMESTIC PARTNERSHIP



## 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 0.54).
- Demographics explained 36.5% of the variance in technology engagement, but 5% of the variance in growth mindset 2% 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?

# THANK YOU!

Check out the full repository for more information:

[github.com/swfisher/capstone\\_2/submit](https://github.com/swfisher/capstone_2/submit)

Feel free to get in touch.