Project Final Submission

Behavioural Research and Experimental Design

Dhruvee Birla & Tejasvi Chebrolu

2019115006 & 2019115008

Original Project Overview

Goal

The goal for our project was “Analyzing the impact of belief in Machine Learning Models by predicting gender from confessions posts.

Overview

For this project, we define a few terms before we can understand our goal.

Belief

Belief is defined as the mental acceptance or conviction in the truth or actuality of some idea.

Gender Prediction

Gender prediction is a machine learning task where the aim of the model is to infer the gender of a user based on some data.

Psychometric Functions

A psychometric function is an inferential psychometric model applied in detection and discrimination tasks. It models the relationship between a given feature of a physical stimulus, and forced-choice responses of a human or animal test subject.

There is a lot of ongoing debate about the potential merging and blurring of the lines between artificial intelligence and human knowledge. There are many aspects of human life where AI is used to improve the ease of working and make our lives simpler. However, not all the uses of AI have been beneficial, there have been cases like the infamous Google Photos Caption Generator where the AI called a group of African Americans a “group of gorillas”. Clearly, this is an example of where it is not beneficial to trust the AI. There have been various examples of gender prediction tasks based on data that would not be traditionally associated with a preference relating to a gender. Psychometric functions have been used extensively for

statistical research involving behavioural research tasks. Via this project, we aim to integrate these three major discplines to understand at what stimulus a person is likely to start trusting a machine learning model over their own instincts.

Experimental Facts

Hypothesis

People's beliefs are likely to be influenced because of their belief in Machine Learning models. This is a non-directional alternate hypothesis.

Variables

Independent Variable: Model Accuracy

Dependent Variable: Truth Value

Confounding Variable: Knowledge of Machine Learning Models

Data Collection

The confessions were obtained from the comments of the following SubReddits :

- AskMen

  A subreddit which has users anonymously ask questions which they would want answered by a man. There are around 4.6 million members in this subreddit.

- AskWomen

  A subreddit for people to anonymously ask questions which they want answered by a woman. There are around 4.5 million members in this subreddit.

We took the top posts of the subreddits for the last twelve months and looked at answers to questions that were ambiguous. An example of such questions were:

> What's the best advice you have been given?

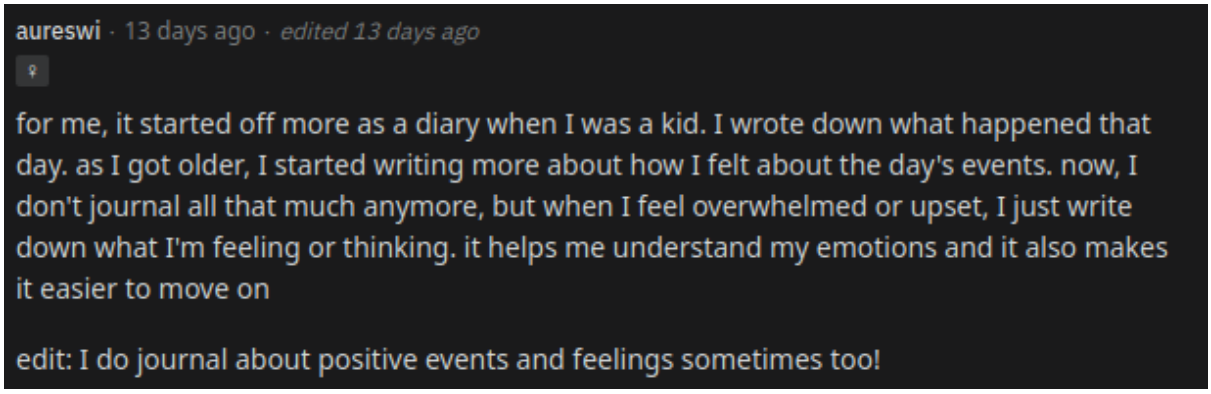
> What is your favorite food?

We used the above questions to divide our posts into five categories:

- Food
- General Advice
- Life Hacks

- Animals
- Mental Health

For each category, we used reddit flairs (users can use flairs to showcase their gender on the particular subreddit) to weed out the random replies to questions that might not be of the required gender. Furthermore, we also took the data from users who have above a certain threshold of karma from that particular comment (around 200). Examples of flairs can be seen below:



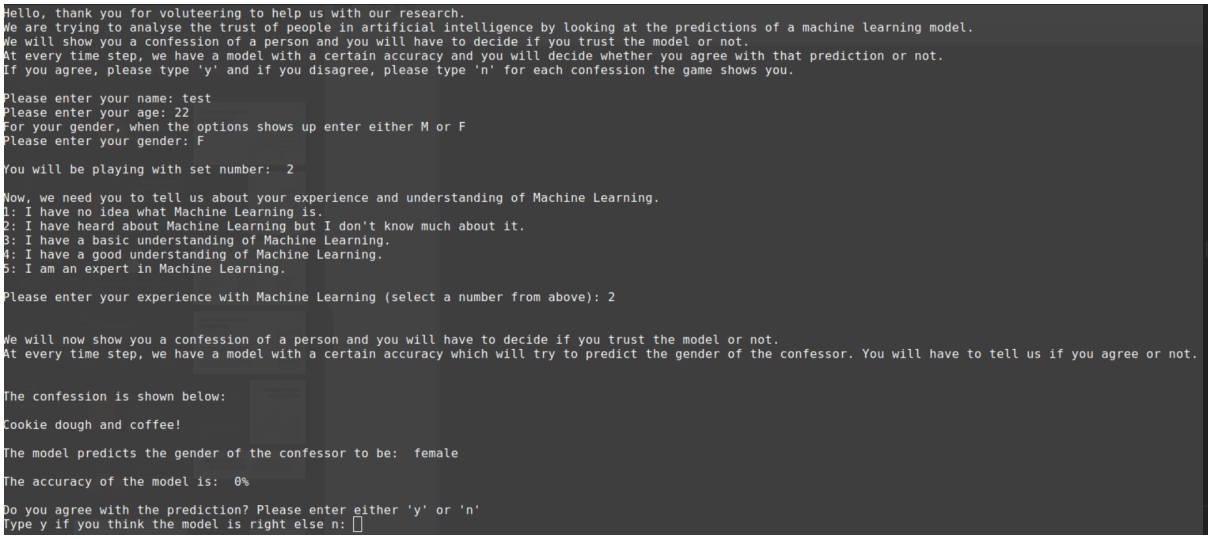
The confessions obtained from these subreddits were then shown to a group of students across batches through random sampling. They were also asked to rate their knowledge of ML on a likert scale.

Experimental Design

Procedure

- The participants were shown 10 randomly generated sentences for each accuracy.
- The model predicted the gender with 0%, 25%, 50%, 75% and 100% accuracy.
- They had to reply with a 'y' (yes) or 'n' (no) depending on whether or not they agreed with the model's prediction.
- An interactive terminal game was used to achieve all the above steps.

A screenshot of the game is attached:



Problems & Solutions

Dataset

The dataset could have been biased with a particular category. This would result in a hidden confounding variable. To counter this the sentences given to each individual were randomly generated and multiple sentences were given for each model accuracy to increase the strength of our evidence.

Participant

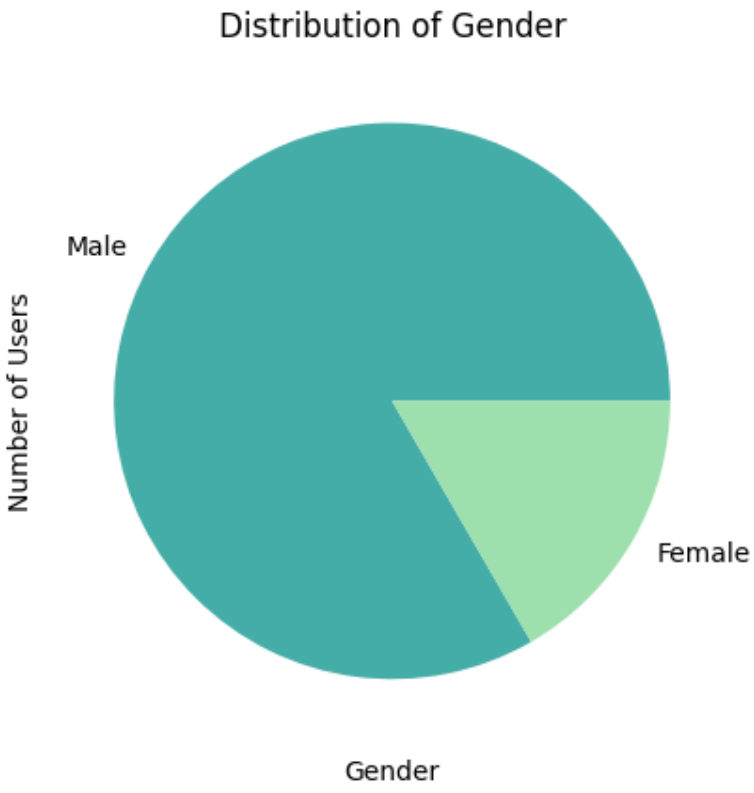
There could have been a participant bias due to improper sampling. This was avoided by choosing participants through random sampling thus avoiding bias. Participants were called from across batches thus involving participants with different levels of ML knowledge.

Outlier

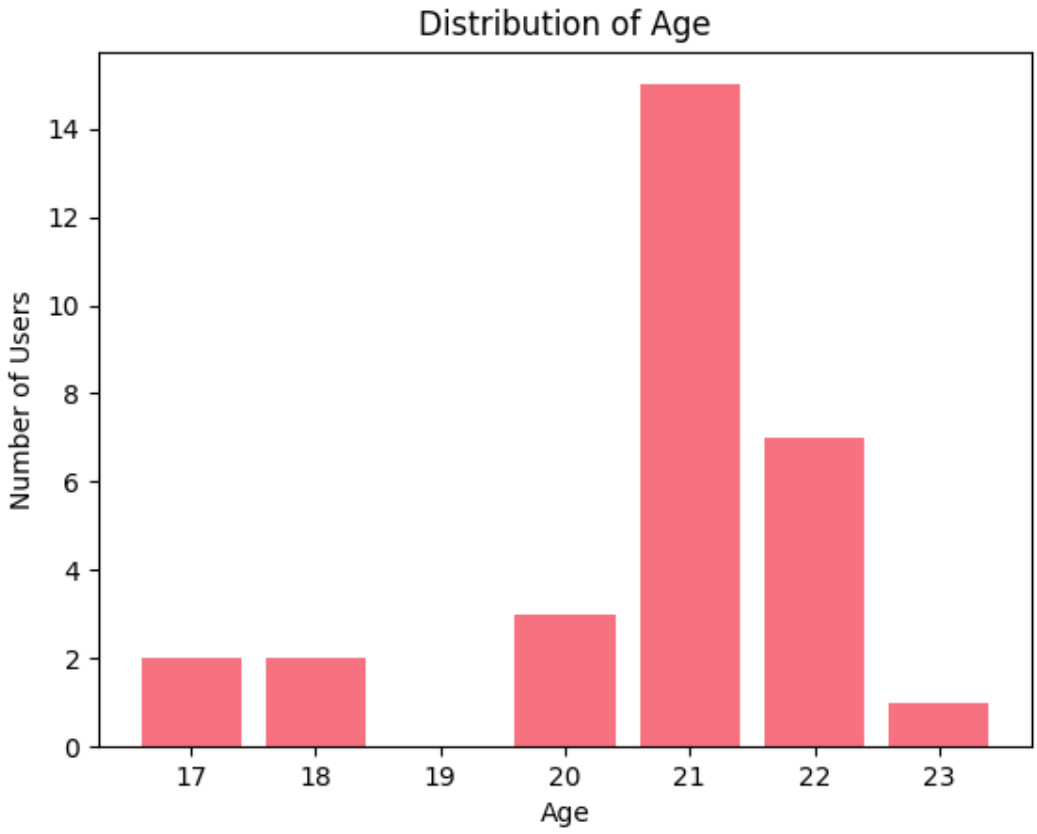
We used manual inspection to remove claims which felt absurd and who had answers which did not seem to match data similar to others with similar characteristics. We had 1 outlier in our collected data. The outlier had answers that did not seem to match the answers of people with the same experience level.

Dataset Representations

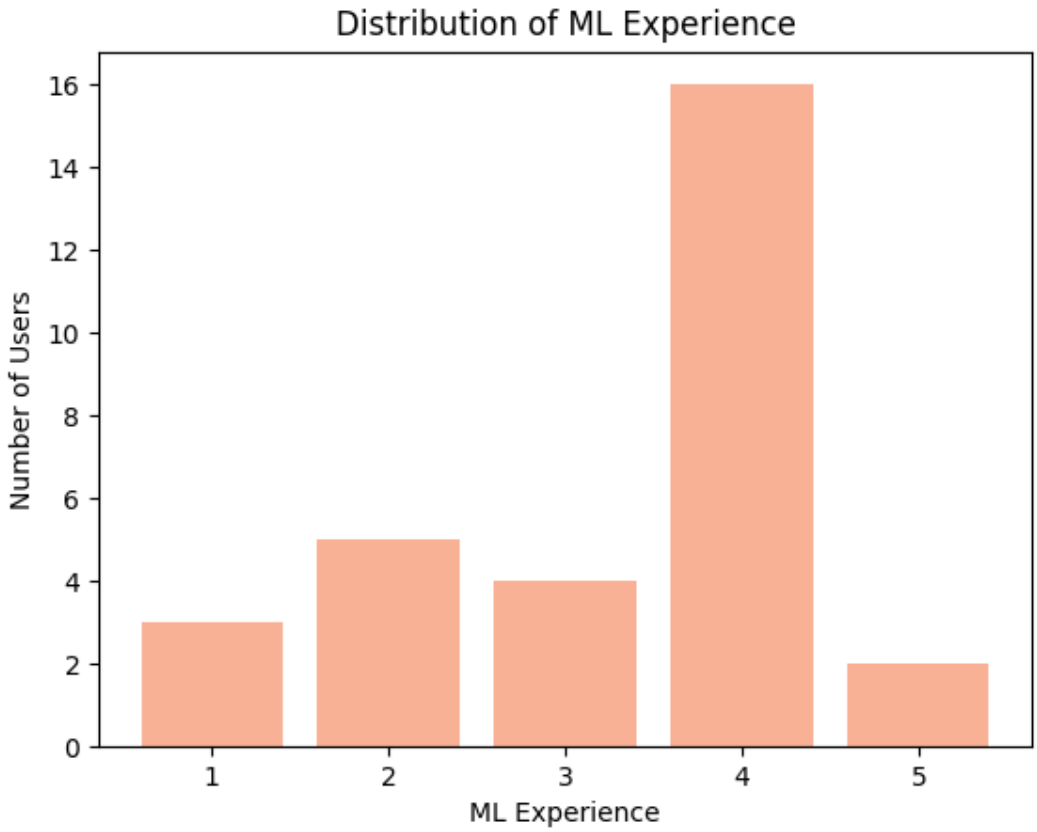
Gender



Age



Machine Learning Experience



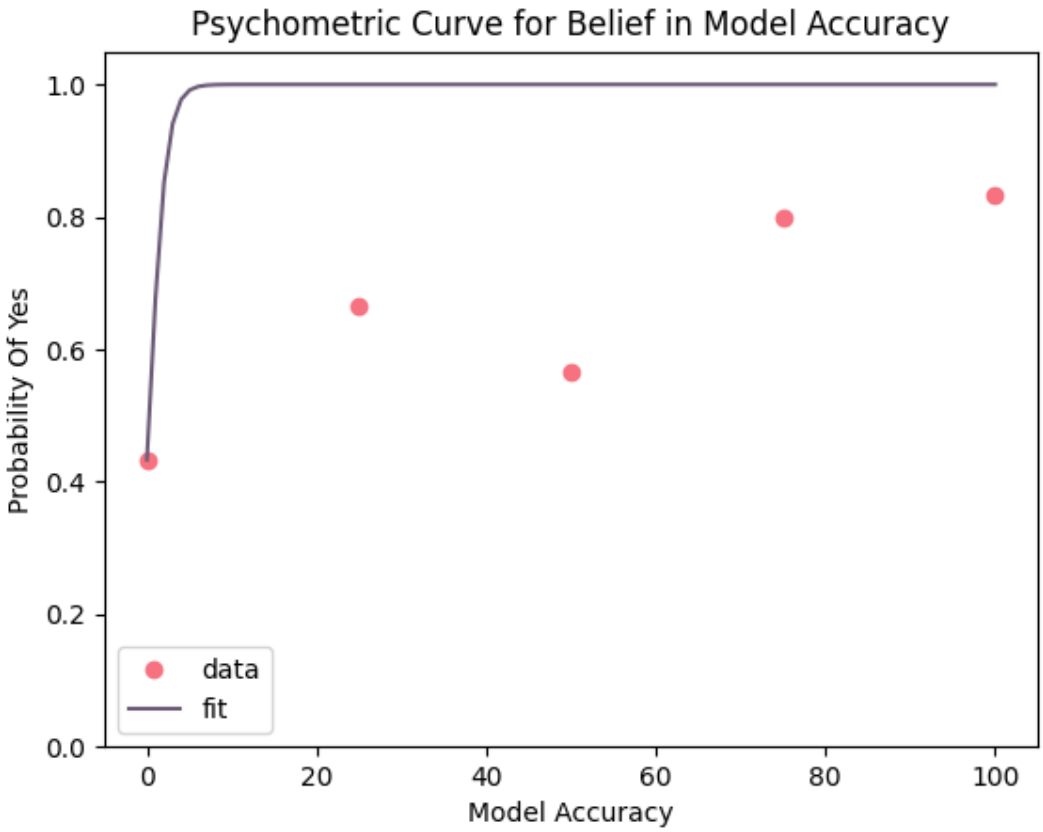
Results and Analysis

The confessions obtained were divided into 2 categories - confessions from men and confessions from women.

The data collected from the users were used to plot the psychometric curve. The x-axis represented the model accuracies shown to the user and the y-axis represented the proportion of the user agreeing to the model's prediction.

We used a simple forced choice experimental paradigm along with a sigmoidal function to fit the curve. This was beneficial as we were able to cover the range 0 to 1 instead of 0.5 to 1. This moves from a subject being certain that the stimulus was not of the particular type requested to certainity that it was.

The curve for the inital experiment can be seen below:



Analysis

▼ The point of subjective equality obtained is around 10% model accuracy. In other words, at 10% accuracy the participants started to differ from the predicted output of the model.

▼ This showed that people had a lot of faith in machine learning models. This was quite counter-intuitive as we assumed that people would choose their beliefs over the belief of the algorithm. Even at around 10 percent model accuracy, people start to feel the stimulus provided by the machine as enough to sway their decision.

▼ It is important to have a significant number of people from different backgrounds to find a generalizable correlation.

Drawbacks

Our major drawbacks were:

- More model accuracies could have been included rather than just 5 accuracies.
- More people with no experience in ML could have been included in this study.
- A different function could have been used for fitting the curve.

Improvements

Ethical Perspective

Our initial survey consisted of participants who were under 18 (minors). In our new results, we excluded participants who were below 18 and collected data from people who claimed to have similar machine learning experience level to ensure the same amount of data.

Better Usage of Categories

One drawback our experimental design faced was the ability to show consistency of data from similar questions from the same categories which had both the genders being predicted. For example, there were no questions in the sub-category “food” that had male as an answer. So, we chose a similar question from the other subreddit (askMen in this case) to ensure a balanced dataset with approximately similar male and female answers. Therefore, our new design ensures that each set shown to the participants has a roughly similar female/male answer prediction; a similar category ratio; and a similar gender per category ratio as well.

Better Usage of Available Data

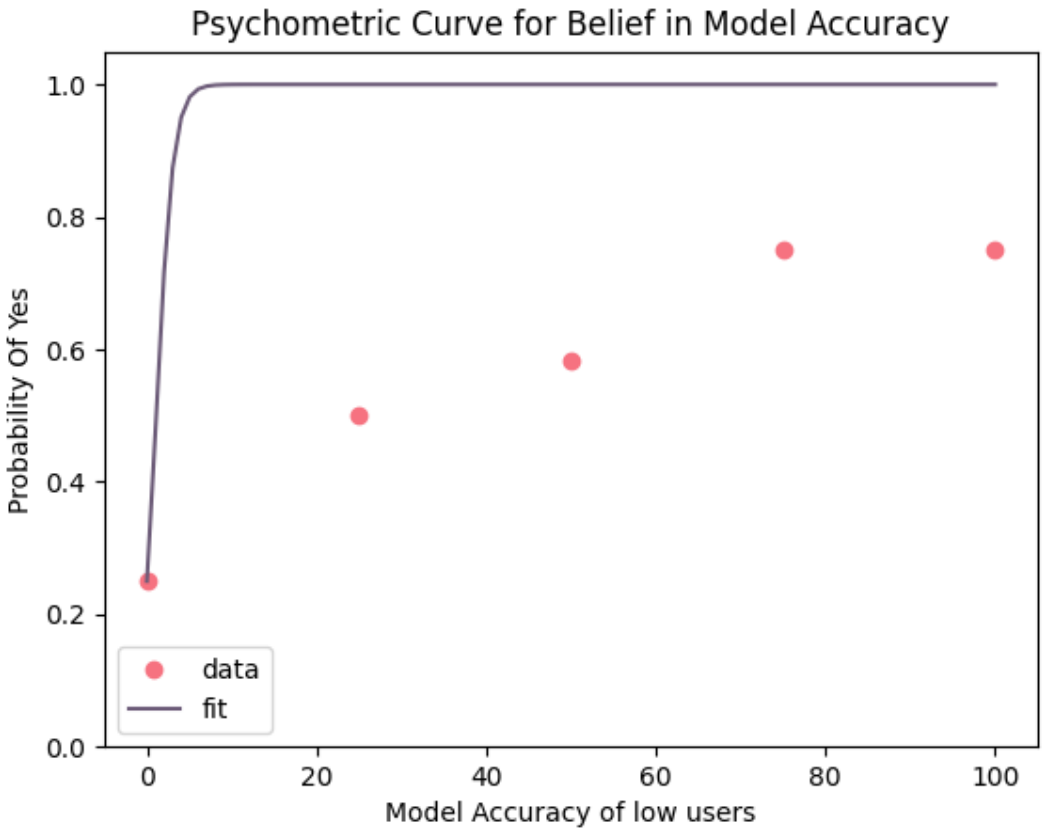
We divided the users with Machine Learning Experience with values in [1, 2, 3] into the low group and the remaining into the high group to draw comparisions within different understandings of Machine Learning. This helped us understand our data better and draw more conclusions from the available data. There were 16 participants in the low group and 18 in the high group. Any conclusion we drew based on the psychometric graph was repeated for both the groups.

Better Statistical Tools

We used the regular sigmoid function to fit the psychometric curve earlier. To improve our project, we added a bias term to account for the low number of accuracies that were being shown. Additionally, we used a solver to calculate the point of subjective equality instead of a random eyeball. Since the graph is extremely stretched this help us understand the visualisations better.

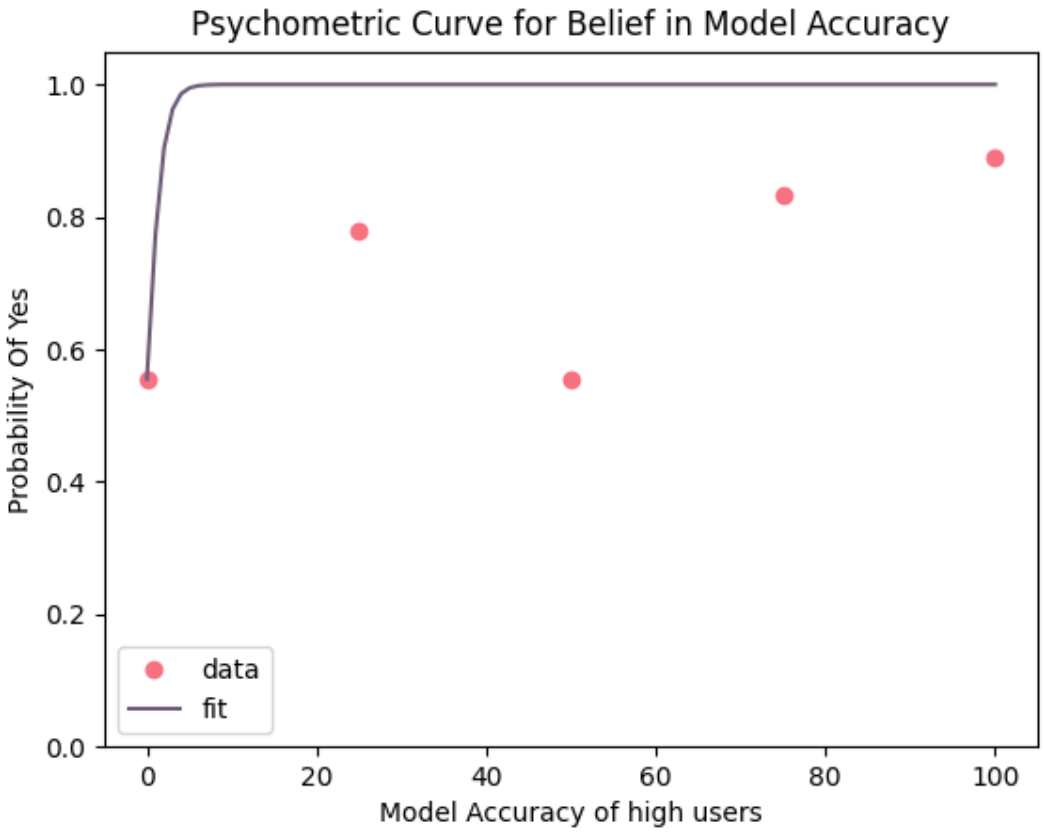
Final Results

Group I (Low)



Here, the measured point of subjective equality was around 15 percent model accuracy. This again reiterated our earlier understanding of machine learning trust.

Group II (High)



Here, the curve can be seen to be fitting differently. This is because of the added bias elements present in the sigmoid. Similarly, our understanding of a very high trust in machine learning is further reinforced.

Future Work

Future Work would include creating an experiment that would ask for the rating of their belief in machine learning models. This belief in machine learning models could be used to perform a sample t-test analysis to check the difference between the mean of their actual belief and the calculated belief. This would ensure another layer of check. Furthermore, the data could have been taken from other universities to ensure datapoints with “actually no knowledge of machine learning”. Increasing the number of model accuracies could also help in better separating the data to understand the psychometric graph in a better manner.