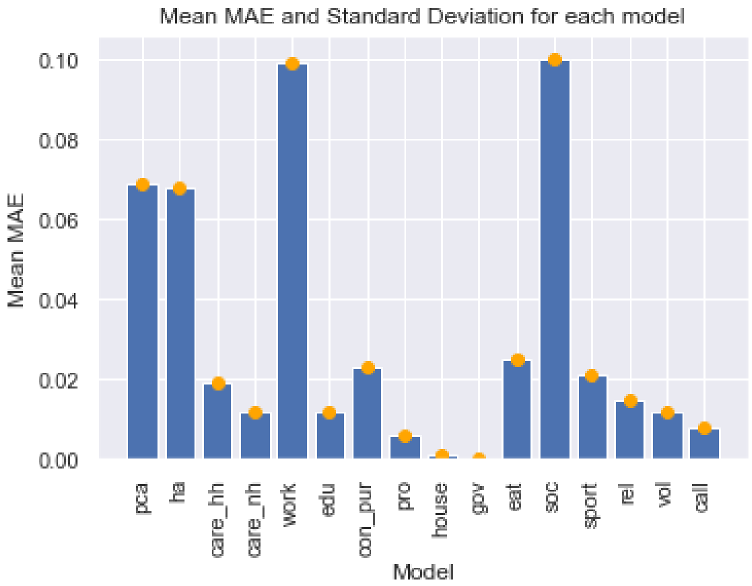
The problem we are trying to solve is understanding how different factors, such as age, gender, income, etc., influence an individual's time usage. This can help viewers understand how different factors influence people's daily activities and how they allocate their time, which can have important implications for productivity, quality of life, and overall well-being.

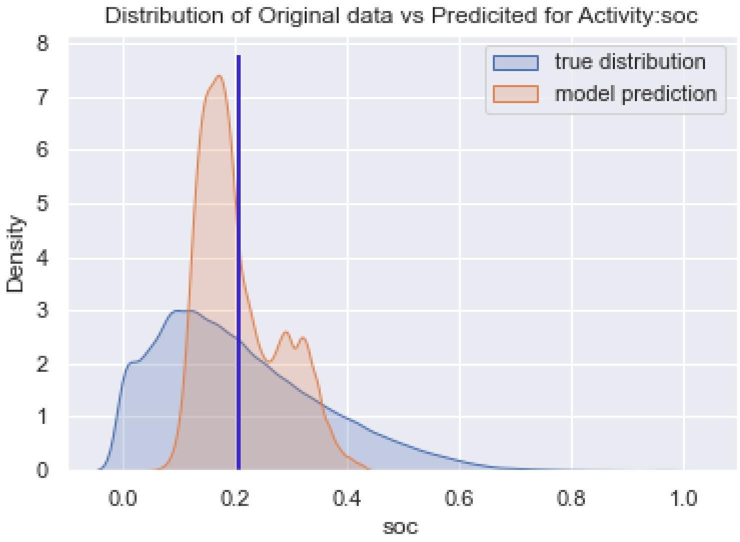
To solve this problem, we can use data visualization techniques to create visual representations from the predictors that allow us to easily see patterns and trends. For this project, I used xgboost to implement gradient-boosted decision trees, where I trained a regression model for each major category of time usage. The models were trained on an input of 19 features for each respondent and the output for each separate model was the proportion of time that activity was predicted to take. Since the output is a proportion, the original data had to be transformed by dividing the total number of minutes in a day.

Some visualizations used to show the product of the predictors by showing their mean absolute error and standard deviation to show accuracy and the predicted distribution compared to the original data, including both all groups and specific subgroups.

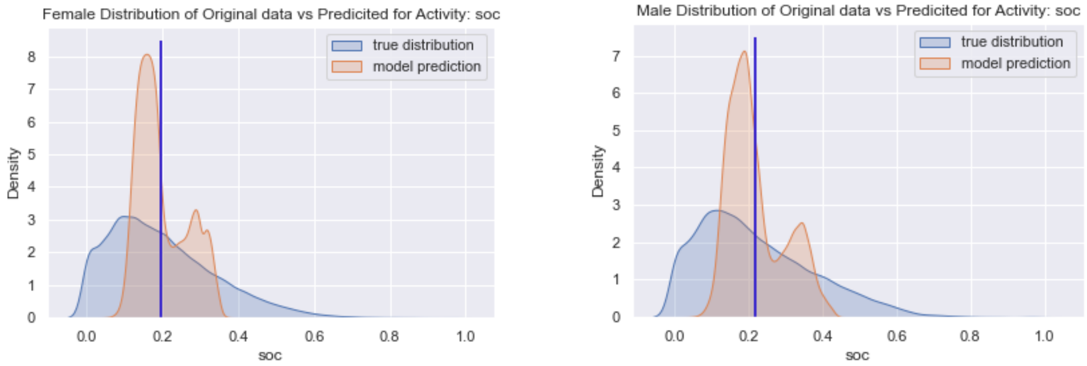
Showing the mean absolute error and its standard deviation can help us understand how well the predictors are doing. Since this is a fairly simple metric, we want to be able to easily compare the different models' performances. With that being the case, I chose to simply graph these values using a bar chart with the standard deviation being the error bars. This allowed for easy comparisons and being able to tell which models performed the best. In this case the lower the better for MAE, so government services and housework were two of the best models according to the graph. Here’s an example:



Of course, the mean absolute error only discusses the calculation of how well the model is doing, but another thing to visualize is the distribution of what the predictor is outputting compared to the original data. Although the data is a discrete distribution, I thought a good way to visualize this was by using a kernel density estimation plot, a continuous distribution. And for the task of comparing the two distributions, overlapping them, and displaying the mean seemed appropriate. Here is an example of one of the distribution comparisons:



While the mean absolute error and the comparisons of the distributions helped show the predictor's output on all groups, there might have been certain subgroups that are predicted as more or less accurate than the group as a whole. So, we use a similar visualization as the distribution comparison above, but this time only compare subgroups' predictions compared to their original distribution. Here is an example of male vs female distributions against their original distributions.



Of course, the code can be adjusted to compare all the different subgroups' distributions.

To evaluate the effectiveness of my visualizations, we can consider several factors. The most important factor that I wanted to emphasize, and believe that I did, was that I wanted to ensure that the visualizations are easy to interpret and understand. This means using clear, concise labels and ensuring that the visualizations are not cluttered or overly complex.

We can also consider the overall design of the visualizations. This includes the choice of colors, fonts, and other design elements, as well as the layout of the visualizations on the page. The design should be visually appealing and help to convey the information being presented clearly.

The designs I have presented are not novel, but they get the job done for the given task in each case. For the first task of showing the mean absolute error and its standard deviation, using a bar chart with error bars is a standard design for a reason, it successfully shows and compares the different means, and if the viewer knows a lesser mean is better, then it is very apparent which models are doing well.

For the other task of the comparison of distributions, while it may be misleading to use a continuous distribution, as the data is clearly discrete because it is a count of the number of respondents, I think there's something to be said for using the continuous distribution instead, it looks more visually appealing and can still be easily compared.

For all graphs, I went with the grid background to be able to measure values farther from the y-axis. I think this is an underrated detail that is easily implemented, but for the few people that want to measure the values of peaks and troughs it is there for them, and those who don't need it aren't negatively affected by this design choice.

Lastly, I think the color choice works well in all the graphs, the dark background is easy on the eyes, and the bars and error bars and distributions are different colors providing contrast to distinguish between them. Specifically, this contrast comes from the well-known color wheel where blue and orange are opposites, meaning they are complementary and work well together, again more visually appealing.

Overall, by using data visualization techniques, we can effectively explore and understand the factors that influence time usage and how different groups allocate their time. By carefully designing and evaluating our visualizations, we can create clear, accurate, and easy-to-understand representations of the predictions made by the models that can help viewers better understand how people use their time.