Capstone Report

Motivation:

Understanding a person’s characteristics is very important for any hiring manager. Difficulties can arise though in finding this information. Directly asking questions like “what kind of work environment do you feel most comfortable in” or “where do you see yourself in 5 years” can give an interviewer a rough idea of a person’s personality; however, there are a couple of drawbacks. First of all, job applicants expect such questions, giving them time to prepare their answers and learn to answer what the interviewer is looking for. Second of all, there are many upon many applicants. It would be useful to be able to filter out candidates before even getting to the behavioral interview if their characteristics are unfit for the job.

Being able to give a metric on a person’s behavior is useful not just from a business perspective, but also an individual. A person would be able to see their weaknesses and have an objective way of seeing their improvement.

This project looks in particular at “workaholism,” defined as a person’s tendency to work. Workaholism is measured on a scale of 1 to 5. A rating of 5 indicates a person who enjoys working and will neglect social interactions in order to get his job done. A rating of 1, on the other hand, indicates the opposite: someone who tends to postpone getting his work done on time. It is of interest for recruiters to be able to find applicants that strike a balance, that is a value between 2 and 4. These people value getting their work done but also value things besides work.

Goals and Methodology:

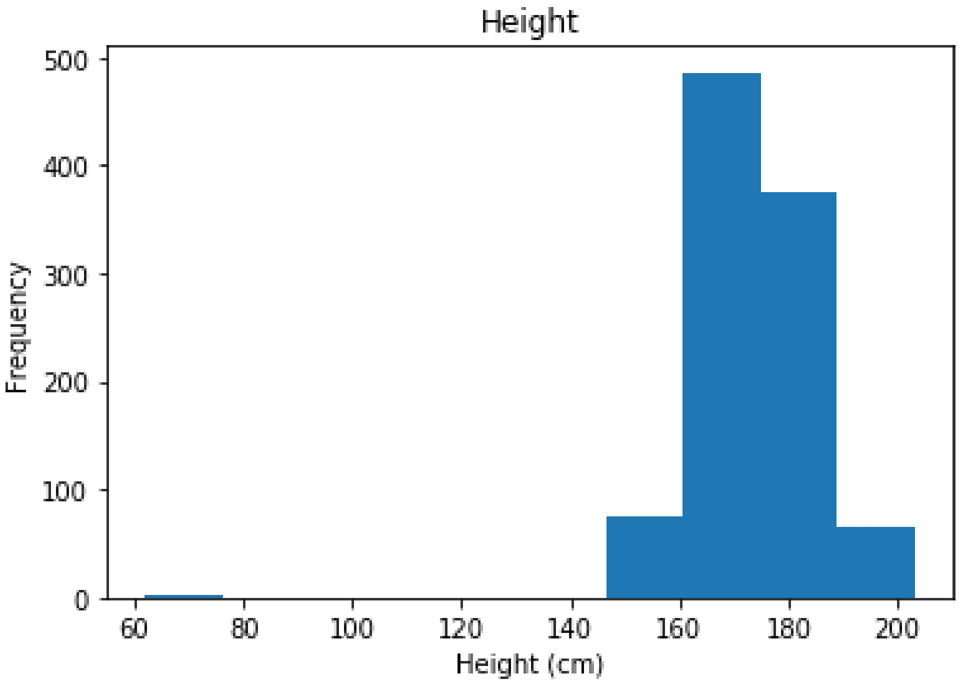
This project will use a dataset from kaggle found here. The survey contains responses from young people, that is people in their young 20, about their background including things like the setting they grew up in (urban vs rural), number of siblings they have, and education level. It also contains responses about their interests in music and movies, and behaviors like how much they enjoy shopping. It also contains responses towards the target variable, workaholism. Each response can be viewed as another feature. The data set is fairly small with only around 1000 people taking part in the survey. There are over 100 features.

There are 2 main goals in the project. The first is to identify which features are most closely related to workaholism. The second is to build a model with machine learning to predict how likely it is that someone would say that they are workaholic.

Data Wrangling

After importing the data, I selected which features to use in my model. While doing this, I filled in null values with the average of the features. Next, I noticed that while most data points were numerical values between 1 and 5 describing the person's affinity to the variable in question, others were strings. To correct for this, I mapped the string values to numeric. For example, I mapped the string ‘social drinker’ value under ‘Alcoholism’ to 5 since that indicates a strong affinity.

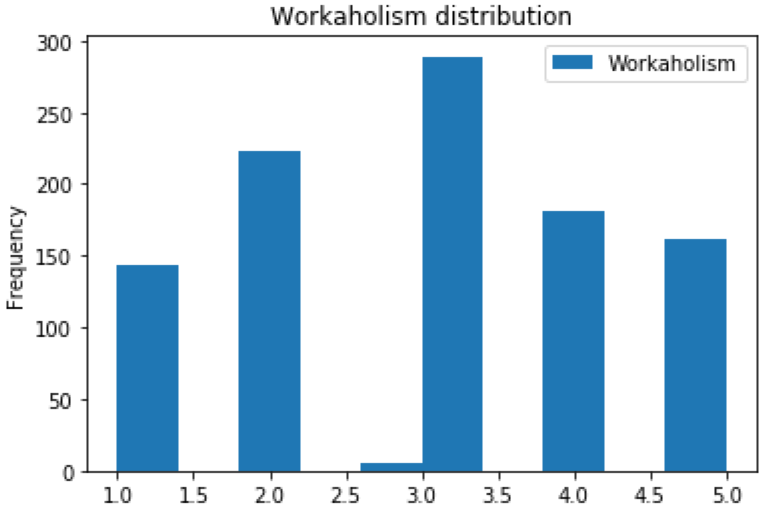
After doing this, I checked for outliers. The only columns not rated 1 -5 were height and weight. Upon plotting I saw that there was 1 height between 60-80 cm (see below).



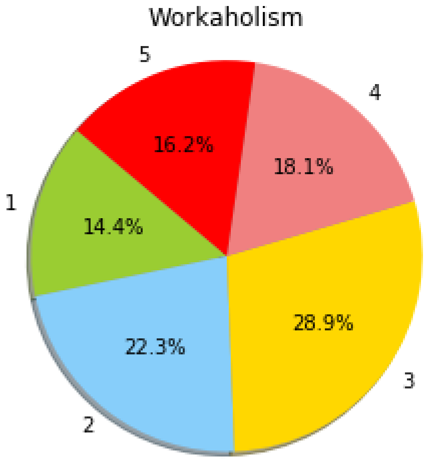
I corrected this point by noticing that it’s corresponding weight was in the normal range. I therefore filled the point with the mean height of people with that given weight. Finally, I confirmed that all the other columns had values between 1 and 5.

Exploratory Data Analysis

I next visualized my data to find trends in the relationship between variables. First, I plotted the distribution of our target parameter, Workaholism.

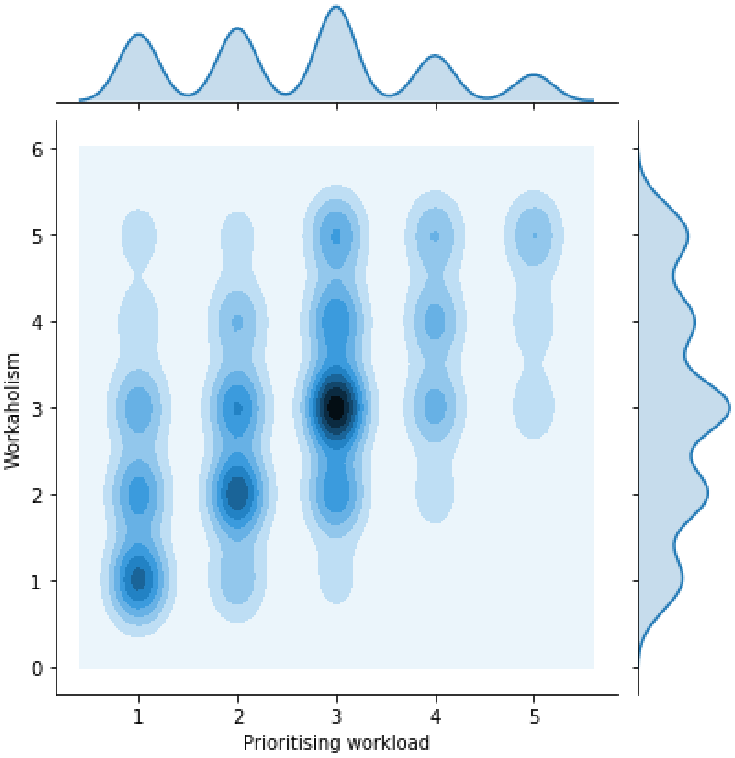


The distribution looks normal with most of the values being 3.

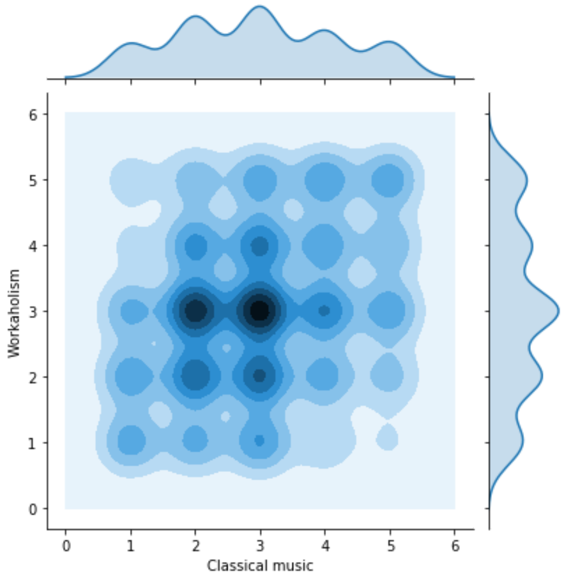


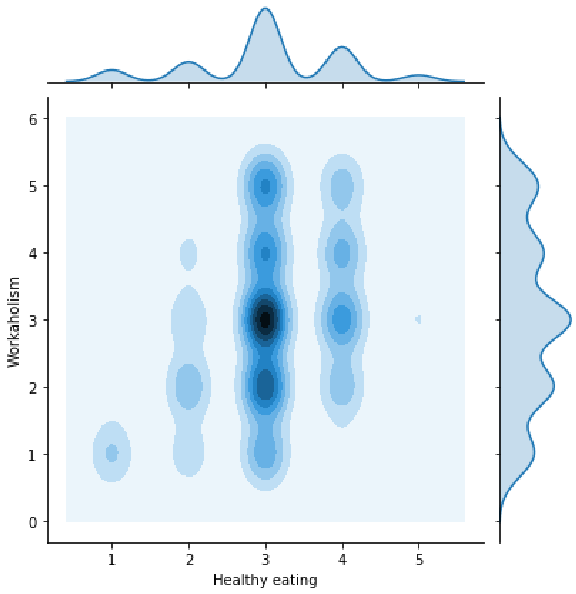
Note that 1 value is between 2 and 3 as null values were filled in with the mean. Also note that 30 percent of the values fall out in the extremes, 1 and 5.

Next, I visualized correlations by creating join plots to get an idea of which features to include. Most of the graphs did not show any strong correlations. A couple did though. One correlation that made a lot of sense intuitively was with ‘prioritizing workload’.

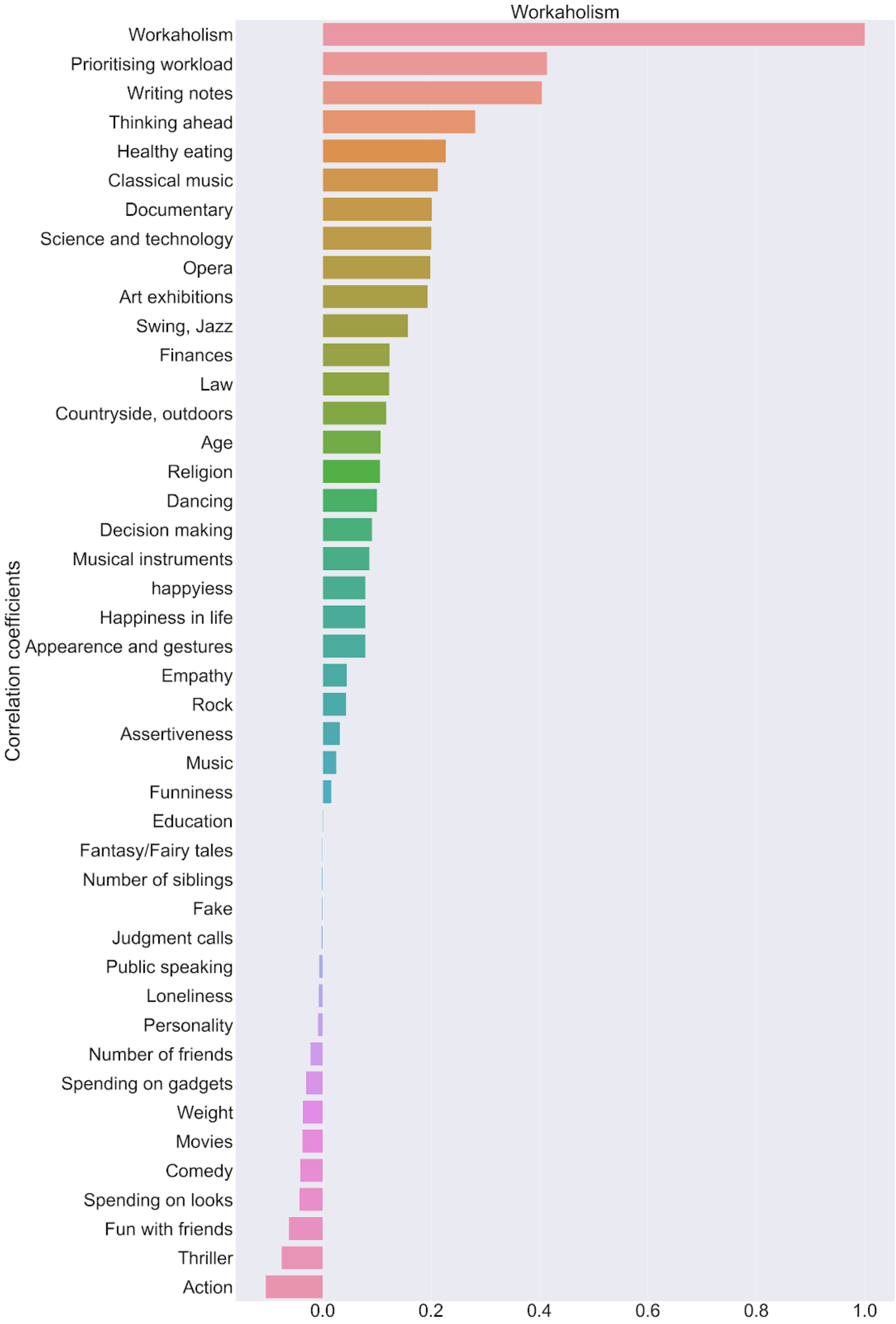


We see that people who don’t prioritise their workload tend not to be as workaholic.





People who eat healthier tend to be more workaholic and tend to like classical music more. I then used statistical methods to be able to more easily identify the relationships. To do this, I calculated the Pearson coefficients



This graph gives the correlation between workaholism and our various variables.

We see as expected that there is a strong correlation with ‘Prioritising workload’, and ‘thinking ahead’. We also see some intuitive negative correlations with ‘fun with friends’. Finally the lack of correlation between some of the features like ‘education’ is somewhat surprising.

After calculating these coefficients I evaluated whether these values were significant or just by chance. First I set up a hypothesis test. The null hypothesis was that there was no significant relationship between ‘Workaholism’ and ‘ Prioritising workload The alternative was that there was. I set a significance level of .05 and split up the dependant feature by the young people who said they strongly had it (rated themselves higher than 3) and those who said they didn’t have it (had less than 3). I found that the p value was very close to 0, allowing me to conclude that there was a significant relationship. Instead of repeating this for each feature, I set up the same test in a loop for all of the features. I found that 23 of the 42 columns I was looking at had significant correlations.

After doing this, I also found a 95 percent confidence interval for mean difference of workaholism and ‘fun with friends’ as another way to double check for significance. The fact that 0 was not included in the CI, (-.144,-.03) shows that it is unlikely that there is no correlation. That being said though, a value of -.03 would be a weak correlation.

This step was very useful in accomplishing the first goal of the project, seing which features were most related to workaholism. The stronger the correlation coefficient, the more closely related the variables are. It also helped me choose which features to choose for the next step in the project, using machine learning to generate a model.

Machine Learning

First I chose the MSE as the method to evaluate the model performance. That makes the most sense here since I want to impose a greater penalty on predictions that are farther away.

Next, I set a baseline performance by evaluating how accurate we would be if we just select the average every time. The MSE for this was 1.6.

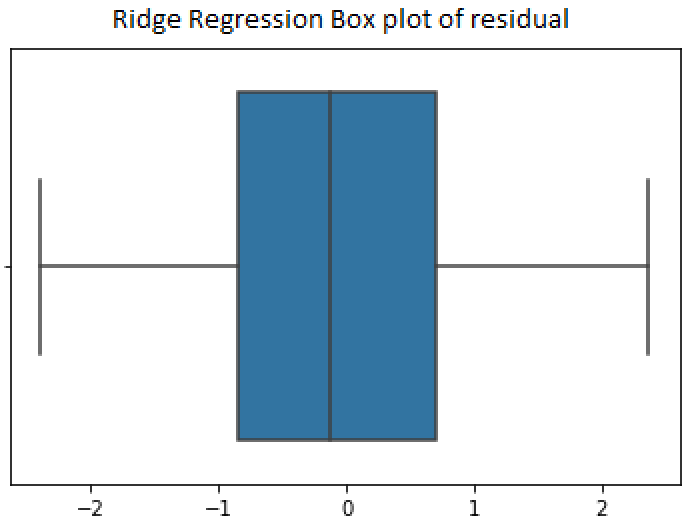
I then evaluated 3 different models. I looked at ridge regression first as a way to see which features were important and which weren't. I then used a random forest regressor and svr to see if those models were better.

In order to cross-validate, I split up my data 20/80 into a test set and a training set. I then ran a grid search with different parameters depending on the model to compensate for overfitting with the training set. I chose the parameters through trial and error. For example, with ridge regression I first chose alpha values of .001, .01, .1, 1, 10, 100, 1000. After seeing that 100 had the best cross_val_score, I redid the grid search, this time with values of 100,200,300… Similarly, for forest, I experimented with the min_samples and max_depth parameters and with SVR the epsilon, gamma, and c parameters. I reported the cross validation score for each in order to compare the models.

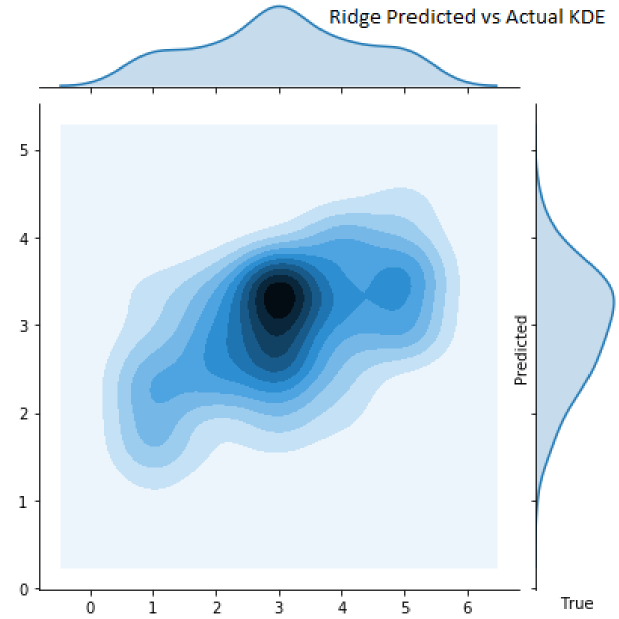
I found overall that the models performed very similarly. All the models achieved an MSE of a bit over 1. The forest model though, had significant overfitting, even after cross-validation. I decided to use the Ridge Regression Model.

Analysis of Model

Our MSE of ridge regression is significantly better than the baseline MSE, 1.6 compared to 1.09.



This graph shows how most of the predictions fall within 1 of the true value.



This graph shows how we can use our model. Note how when the true model is 1, the model predicts no higher than 3.5. Likewise, when the true value is 5, our model predicts no lower than 2.5. Furthermore, 5 percent of the predictions are above 4 and 10 percent are below 2.

We could put this information together to make a conclusion based on the results we get. If we get a value between 2.5 and 3.5, we can't really make any conclusions. As our MSE is 1.09, it is possible that the true value falls in either extreme. If we get a value though, above 4, we could reasonably conclude that their true workaholic is around a 4 considering that our model does not give out that many 4s. Likewise, if we get a value below 2 we could conclude that the true value is either a 1 or 2 since our model does not give out that many 4's.

Practical Use and Further study.

In practice when it comes to filtering out candidates, our model should not be used. Although the MSE is low enough to make some judgments about whether they are or they aren't workaholic, it is too high to make reasonable guesses about where the value truly is out of five.

Further study should involve more samples. 1000 samples are just a bit too small to train our model. Furthermore, the samples were all taken from young adults. A more comprehensive study should use samples from a wider diversity of people.

It also may be of interest to generate predictions for some of the other features, like public speaking.

.