Predicting Alcohol Consumption with Deep Learning methods

Minh Nguyen

MTH 496,
Michigan State University

December 22, 2022

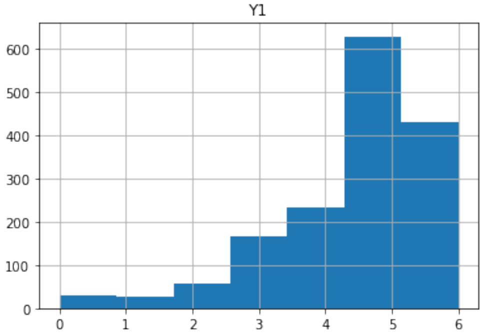
Introduction

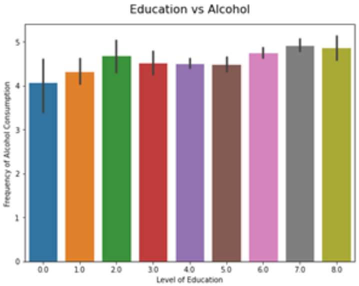
- Predicting whether a person is likely to consume alcohol is an important question in the field of medicine and law.
    - Probation officers and judges need to know whether someone released on bail is likely to consume drugs, and medical experts would benefit from being able to predict a patient's tendency to consume
- Dataset: UCI Data Consumption Data surveys 1855 individuals based off of their personality measures to predict consumption of 18 types of legal/illegal drugs. I focus on alcohol
- Potential concern: algorithmic fair. Does my model discriminate based off of protected characteristics like race and gender?

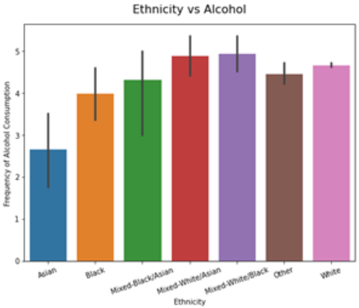
- The data set includes 12 features:
  - NEO-FFI-R (neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness), BIS-11 (impulsivity), and ImpSS (sensation seeking), level of education, age, gender, country of residence and ethnicity
- The data set includes 7 labels for 18 types of drugs:
  - Never Used (0), Used over a Decade (1), Used in Last Decade (2), Used in Last Year (3), Used in Last Month (4), Used in Last Week (5), Used in Last Day (6)
- The dataset if very unbalanced (see figure in next slide). This makes sense because most people have consumed some level of alcohol in their life.
- Therefore, to make for more meaningful (and accurate) predictions, I focus in on whether a person has consumed alcohol in the past month (>3) or not (≤ 3)
  - I turn the problem into a binary classification problem (though I run the model on both the binary and multi-class problem)
  - This has medical and legal implications

INTRODUCTION

DATASET OVERVIEW







- 1,2..,7 corresponds to Never Used to Used in Last Day
- Second figure from user Khadjia on Kaggle.com

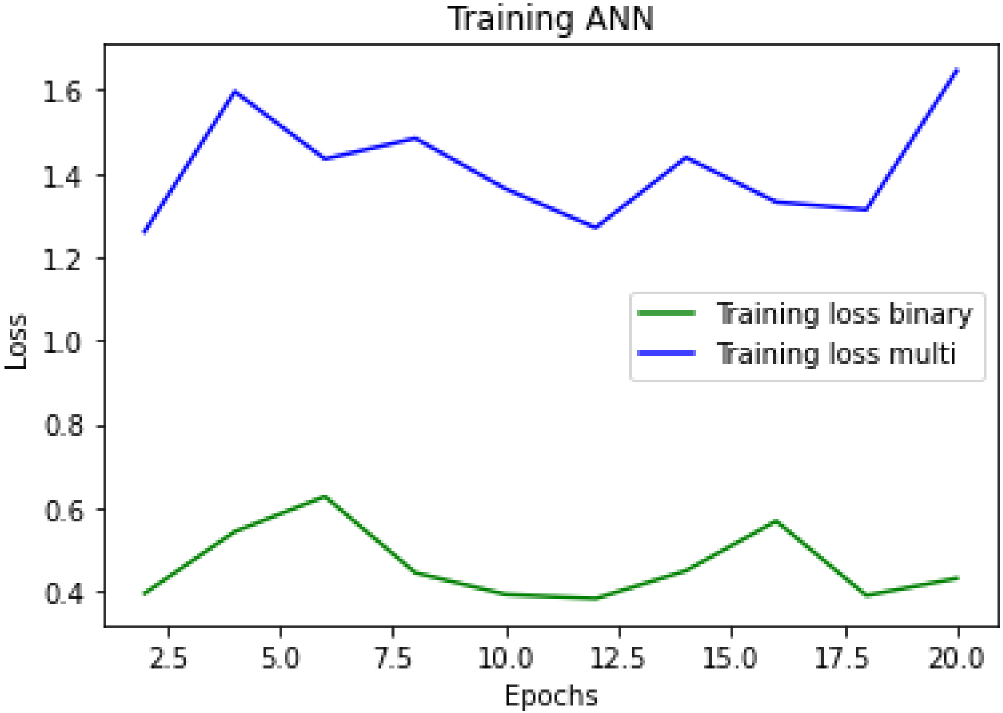
Approaches

- I used 4 models to train the data
  - Logistic regression (for the binary classification problem), Decision Tree (binary and multi-class), Random Forest (binary and multi-class), Artificial Neural Network (binary and multi-classs)

  1. **Logistic regression**: The features are the 12 personality features, while the label is either (1) consumed alcohol within the last month or (0) has not consumed alcohol within the last month. The achieved accuracy is 82.8%
  2. **Decision Tree**: Using decision tree classifier, the achieved accuracy for the binary problem is 69-71%, while the accuracy for the multi-class problem is around 30-35%
  3. **Random Forest**: Building 100 trees with max depth 10 before taking the max voting, achieve an accuracy of 82.5%. The accuracy for the multi-class problem also hovers around 30-35%
  4. **Artificial Neural Network**: Using 1 hidden layer of size 13 and training for 20epochs, I achieve an accuracy of 81.9% for the binary classification. The accuracy achieved with the multi-class was 44.5% after hypertuning parameters
- To reiterate: binary classification refers to whether a person has consumed in the last month (1) or not (0); while the multi-class refers to what was presented on slide 2
- Accuracy is the evaluation metric I used because my problem is a classification problem.
- I was able to optimize my model by reducing it to a binary classification problem, as well as using stochastic gradient descent (as opposed to normal gradient descent that we used in class)

Approaches

Graphs and Comparisons



Approaches

Graphs and Comparisons

My Best LR	My Best DT	My Best RF	My Best ANN	Paper's Best Overall (DT)
82.8%	71.1%	82.5%	81.9%	>75%

Table. Comparison to E. Fehrman, A. K. Muhammad, E. M. Mirkes, V. Egan and A. N. Gorban, "The Five Factor Model of personality and evaluation of drug consumption risk.," arXiv, 2015

- It must be noted, however, that the papers model predict consumption for a variety of different drugs, but tends to use less features
- The paper achieved its best results when using 5-7 features, where as I use all 12 features
- My accuracy is only best achieved when turning the problem into a binary classification problem. This is due to the fact that the classes are very skewed (unbalanced). This is because alcohol consumption is very common, so there will be very little responses 0-1 and much more 5-6.

Shortcomings and Discussion

- One of the main shortcomings of my approach is that I did not have a method for selecting my features, rather I just used all the features. This is something that the paper did really well, hence they were able to maximize their results by carefully selecting the most important features
- In the future, I think that I should extend my model to the other types of drugs as well. This would mean that my labels are less skewed because the number of people who have consumed illegal drugs is not skewed compared to number of people who have consumed alcohol
- My model could be helpful for a judge wanting to predict if a suspect let out on bail will consume alcohol or not. With this information, the judged could set the appropriate bond price, as well as put in approriate measures to stop the suspect from consuming alcohol (think someone who has been charged with driving will under the influence)
    - However, a discussion of algorithmic fairness is crucial

Shortcomings and Discussion

Algorithmic fairness

- For a feature to be ’fair’, it must satisfy independence, separation and sufficiency principle. I won’t go into much detail over the latter two, but I will propose a way to check independence for a few features in my data.
    - This is more so a sanity check as the results of this cannot be used to conclude that my model is fair
- I want to check that $P(\text{Model is the same}|\text{We exclude black people}) = P(\text{Model is the same}|\text{With black people})$.
- I propose ’loosely’ testing this by seeing if the model achieves the same accuracy when I exclude black people from the model. I want to see if there is potentially any bias
    -
    - However, this does not help us conclude whether the model is biased or not since we are not actually calcuating the probability
- Besides ethnicity, personality traits and lived experience (like education) could also contribute to bias

Shortcomings and Discussion

Checking Fairness

- I remove all black people from the training data and retrain my models. My goal is to see if this will change the result when testing the models or not. Hopefully it will not, which would mean that our model is not biased against black people.
- The table below shows the results

LR Before	LR After	DT Before	DT After	RF Before	RF After	ANN Before	ANN After
82.8%	73.1%	71.1%	71.4%	82.5%	81.9%	81.9%	78.7%

Table. Binary classification model accuracy before and after removing black people from the training data

- There does not seem to be a significant difference, so maybe our model is not too biased
- However, we cannot conclude anything from this, and it acts only as a sanity test.
- If we wanted to test fairness, we would have to actually computer probability, as well as test for separation and sufficiency.

Conclusion

- I was able to train a model to predict whether or not a person had consumed alcohol in the last month based on personality traits of that person. The model achieved around 83% accuracy for Logistic Regression and around 82% accuracy for a neural network approach.
  - This is higher than the paper, but my model was a lot more specific than the paper (E. Fehrman, A. K. Muhammad, E. M. Mirkes, V. Egan and A. N. Gorban, "The Five Factor Model of personality and evaluation of drug consumption risk.,")
- The significance of this model I have already discussed: it could be aided to help medical and legal professionals when it comes to decision making
- However, algorithmic fairness remains a concern
  - A quick sanity check shows that the model is a priori exhibting little bias, but we still cannot conclude that it is unbiased

References

- E. Fehrman, A. K. Muhammad, E. M. Mirkes, V. Egan and A. N. Gorban, "The Five Factor Model of personality and evaluation of drug consumption risk.," arXiv [Web Link], 2015
- https://www.kaggle.com/code/obeykhadija/drug-consumptions-edaGeneral-Exploratory-Data-Analysis-(EDA)