An Analysis of Depression

Nicolas Siska

School of Mathematics and Statistics

Introduction

Depression is defined as a mood disorder that is characterized by persistent feelings of sadness and hopelessness. According to the World Health Organisation, around 280 million people live with depression. It causes severe symptoms that affect how you feel, think, and handle daily activities. Many people who suffer from depression report disrupted sleep, lack of concentration, and thoughts of suicide. The cause of depression is complex and can be due to several psychological, biological, and social factors.

Objectives

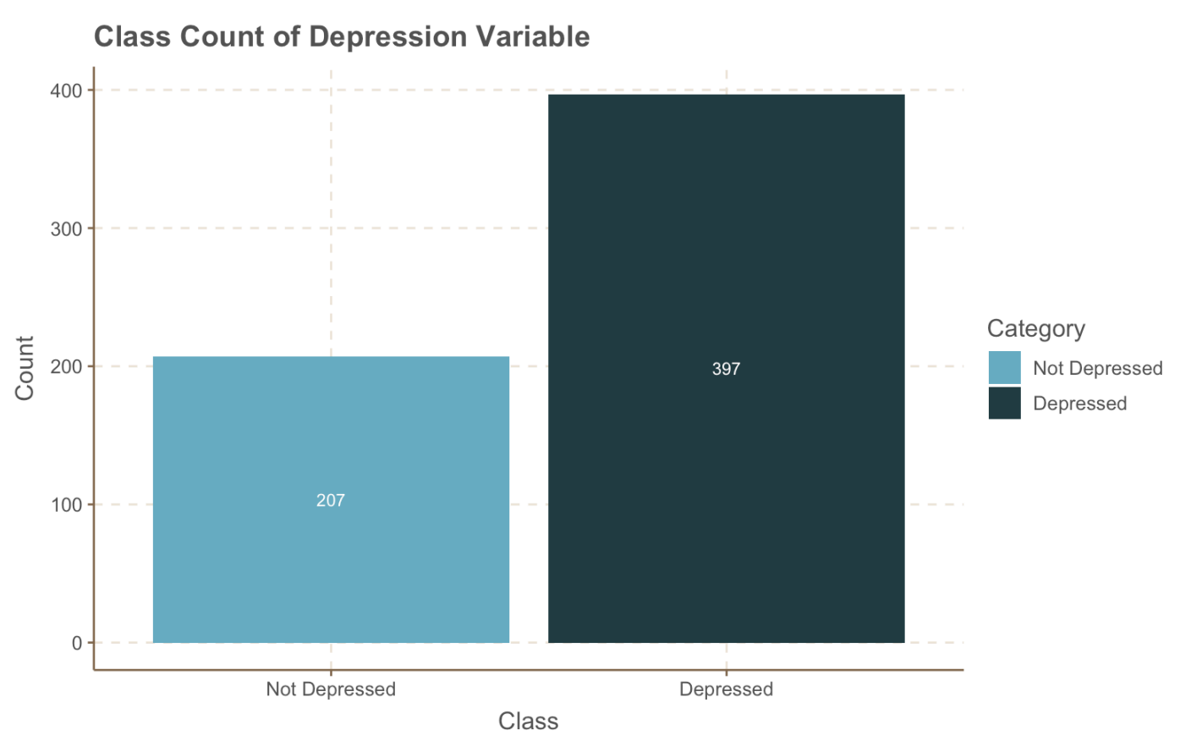
The intent of this study is to analyze what are some main factors that are correlated with depression and whether exercise has a significant effect in treating depression.

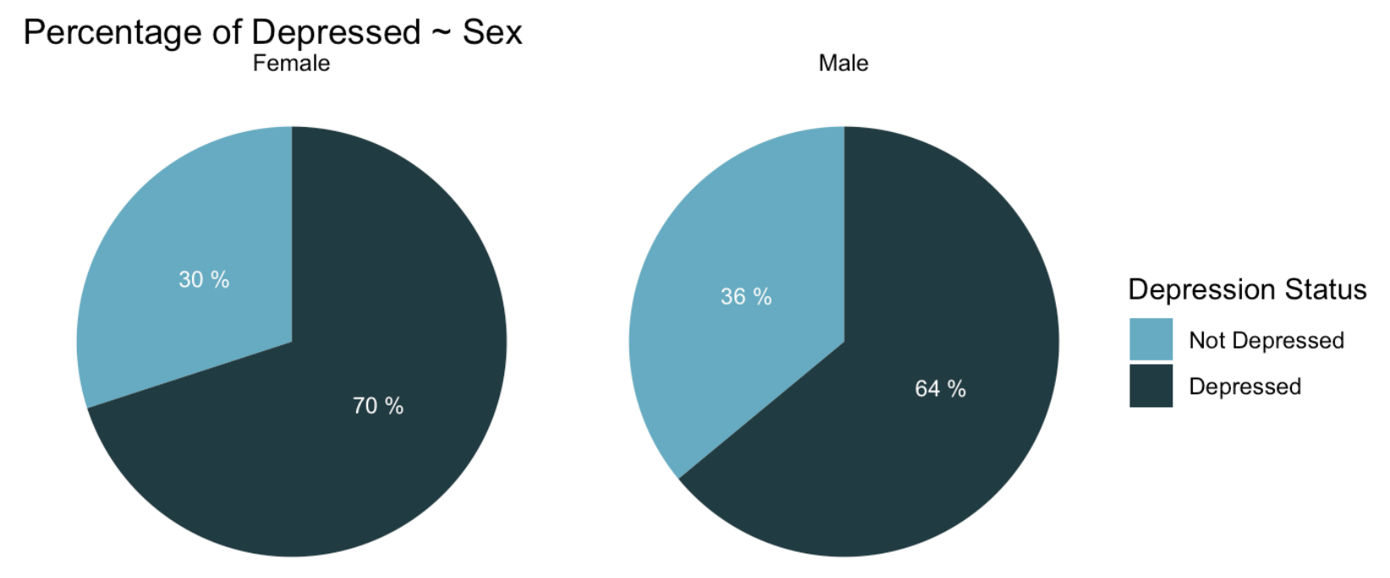
- Analyzing correlation and association using Pearson's chi-squared test and Cramer's V measure.
- Apply statistical machine learning methods to analyze variable importance and significant factors that predict depression.
- Use ANOVA to analyze significant differences in depression scores among different exercise treatments, or use the Kruskal-Wallis H test if the data do not meet the assumptions required for ANOVA.

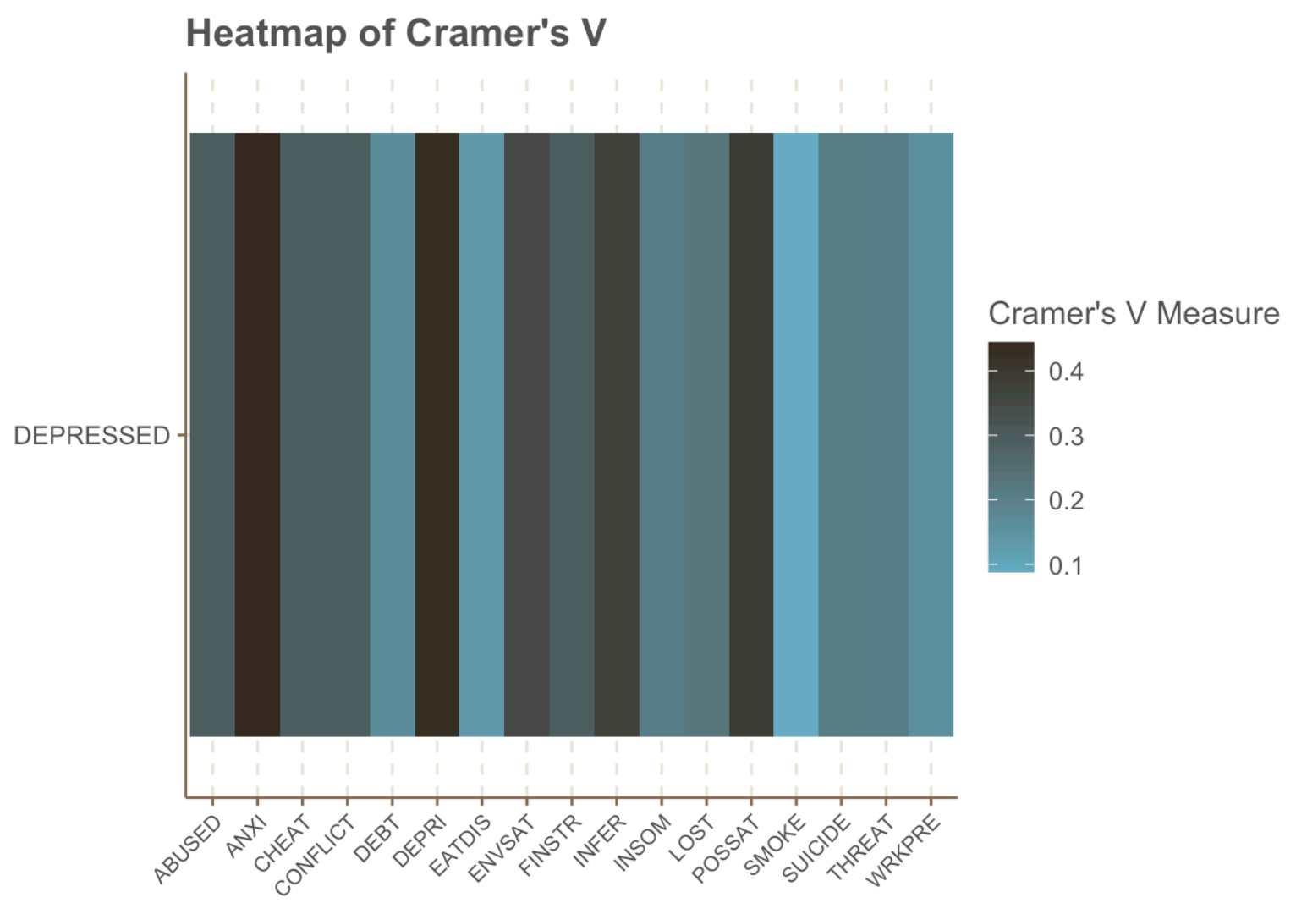
Exploratory Analysis

Two datasets were used in this analysis. The first dataset comes from a study performed in Bangladesh. The second dataset originates from a study with the objective of analyzing the right dosage and modality of exercise treatment for serious depressive disorders.

Bangladash Dataset





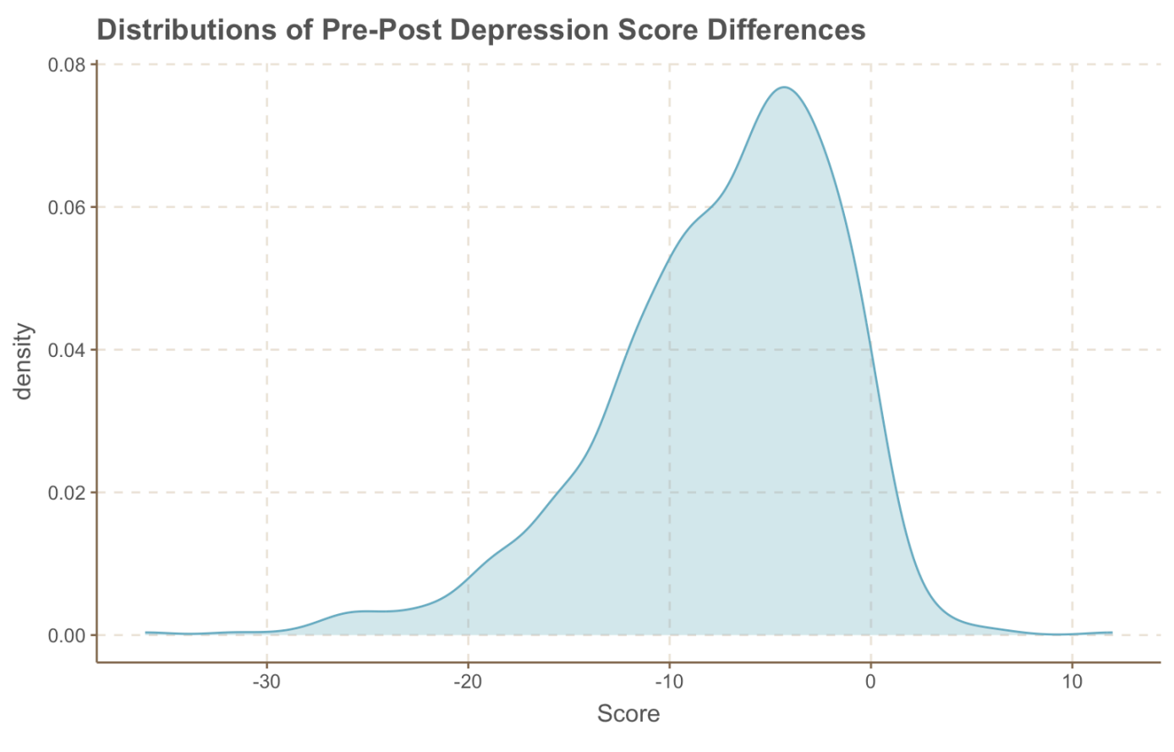


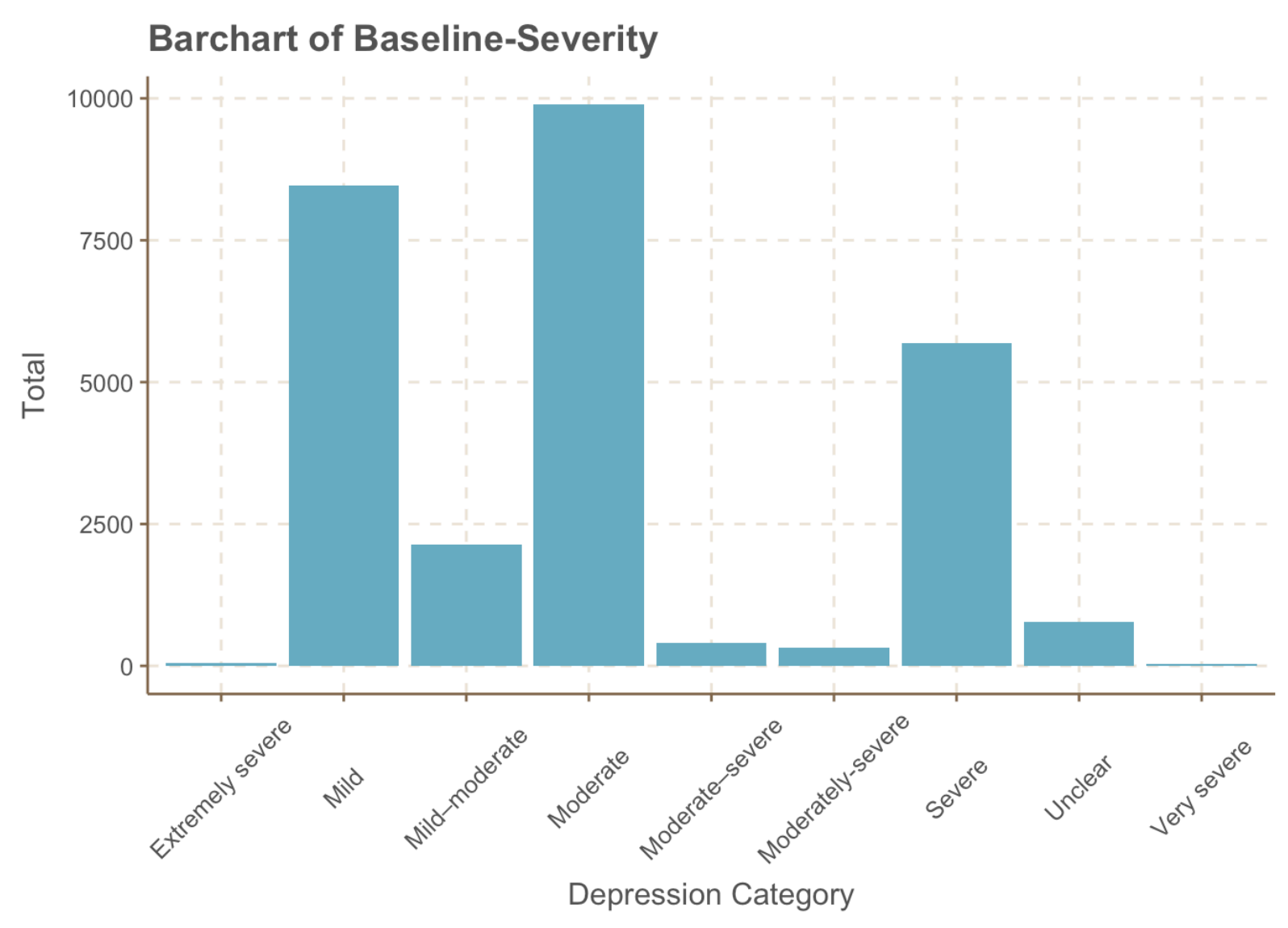
Selected Predictors

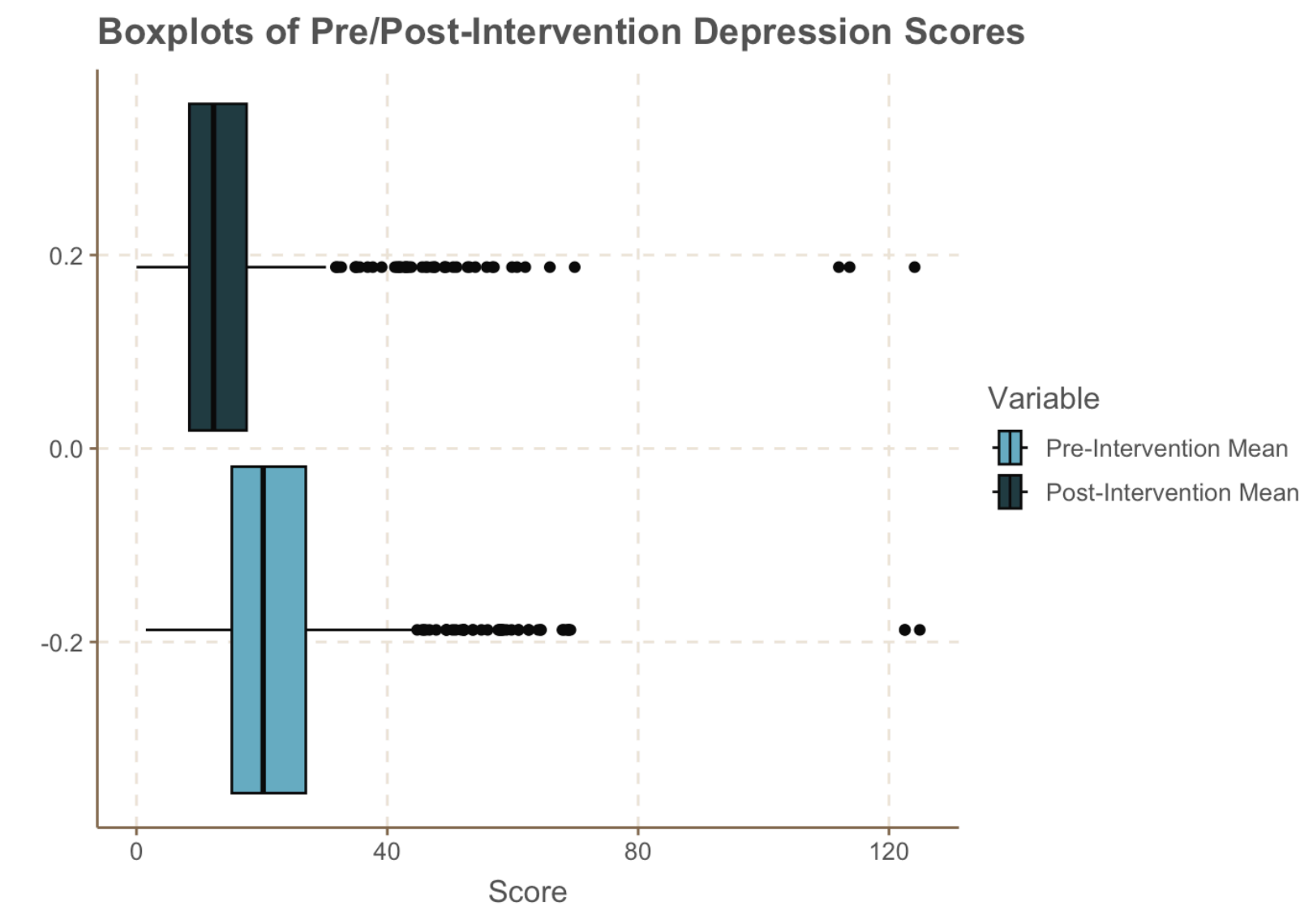
	ENVSAT	POSSAT	FINSTR	INSOM	ANXI	DEPRI	ABUSED
P-Values	3.799530e-17	1.385383e-21	6.900941e-13	3.878315e-07	2.562078e-27	2.048471e-26	4.050444e-13
Cramer's V	0.3465	0.3918	0.2958	0.2107	0.4441	0.4362	0.2996

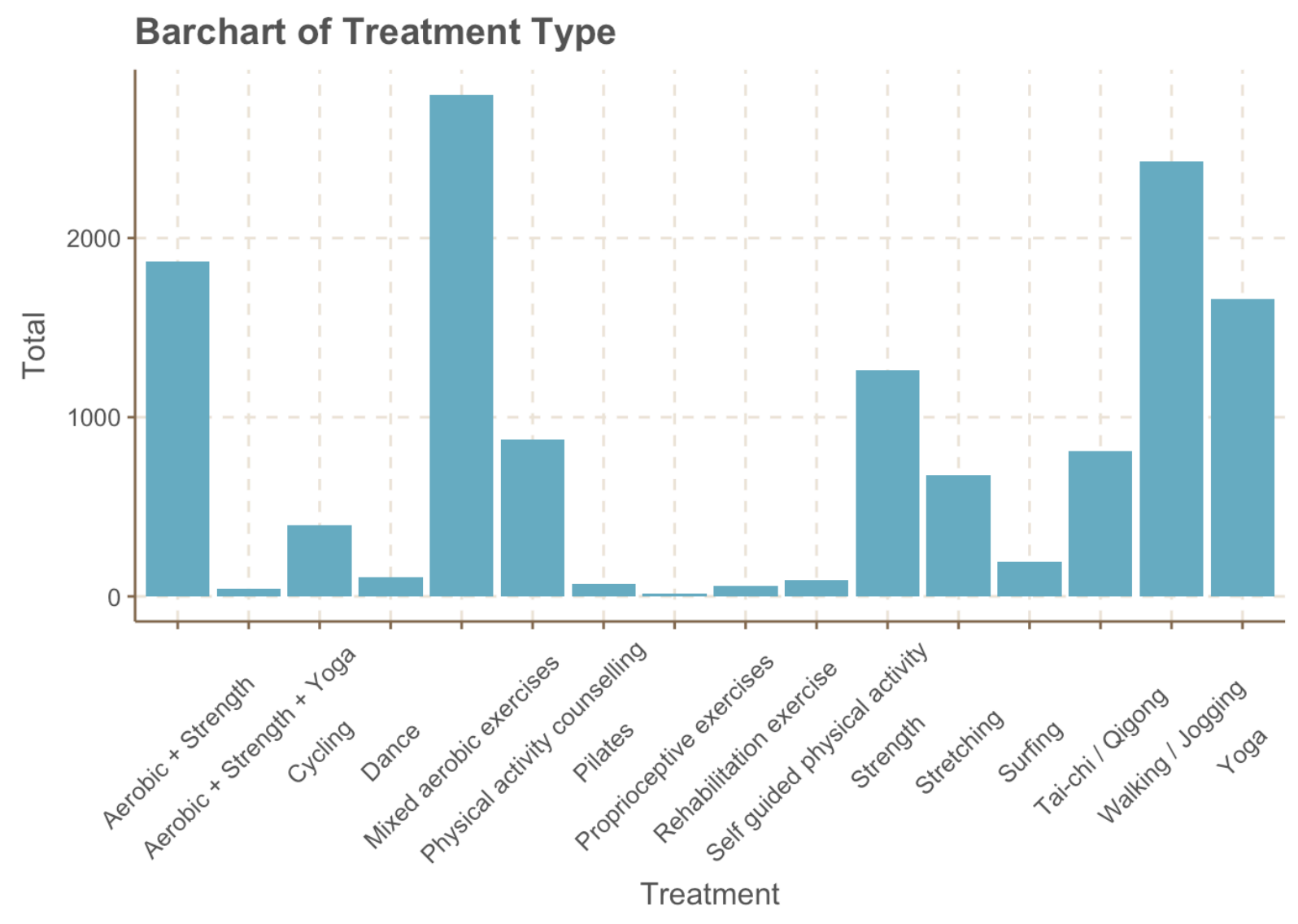
	CHEAT	THREAT	SUICIDE	INFER	CONFLICT	LOST
P-Values	3.680228e-13	5.236748e-07	3.222806e-07	3.278535e-19	6.569594e-13	3.149437e-08
Cramer's V	0.3000	0.2090	0.2140	0.3686	0.2965	0.2291

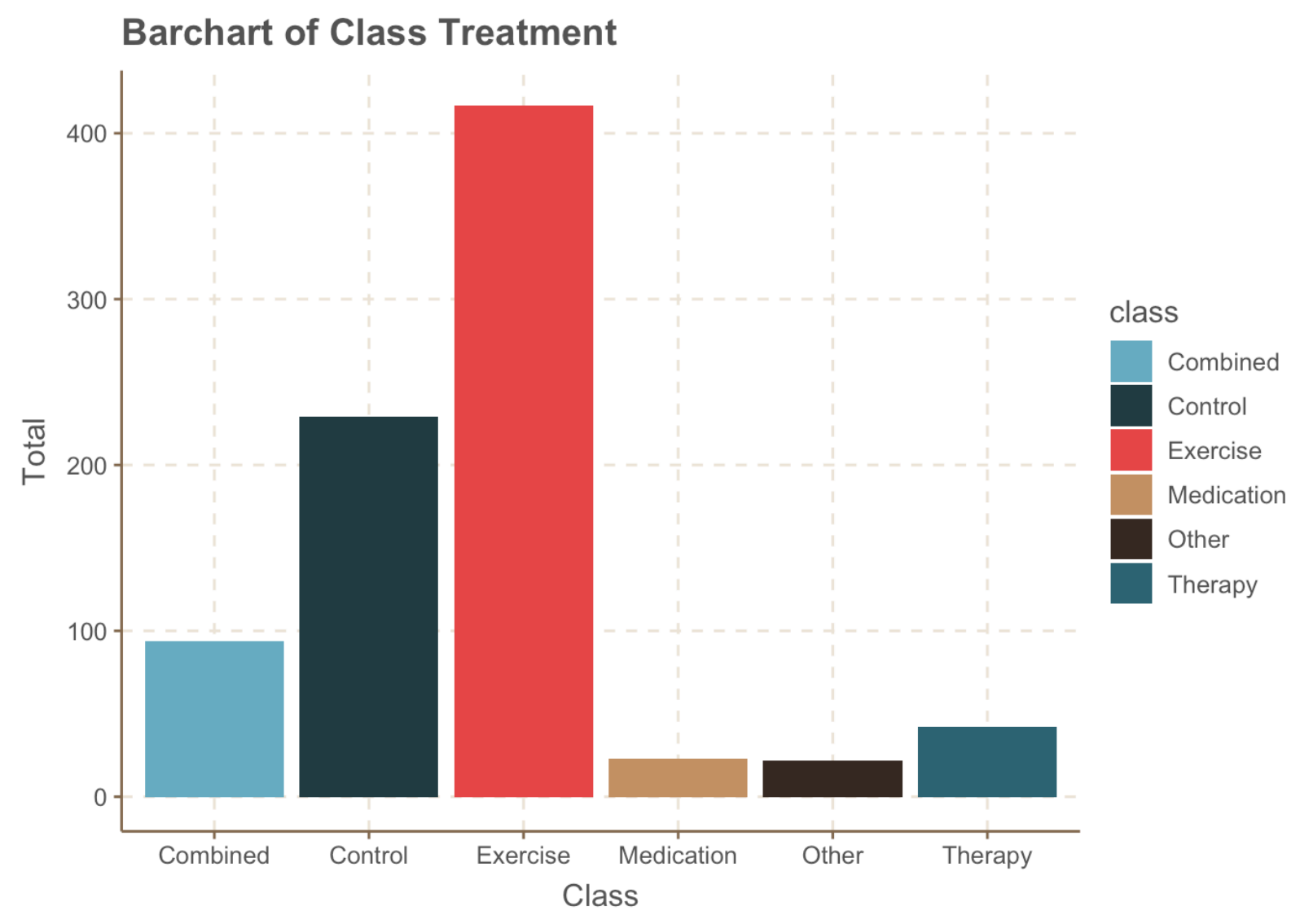
Exercise and Depression Dataset











Methodology

In order to predict the response variable of the first dataset, three machine learning models were used: Random Forest, Logistic Regression, and Gradient Boosting. The following process was implemented.

▶ **Data Splitting**

The dataset was split into sixty-five percent training/validating data and thirty-five percent testing data.

▶ **Training Hyperparameter Tuning**

Using the caret package the three models were trained on the training set, and the hyperparameters were tuned on the validation set.

▶ **Evaluation and Model Selection**

The following metrics were used to choose the best model.

Evaluation Metrics

Model	Sensitivity	Specificity	Precision	Recall	F1	Prevalence	Detection Prevalence	Balanced Accuracy
Random Forest	0.722	0.949	0.881	0.722	0.794	0.343	0.281	0.836
Logistic Regression	0.736	0.971	0.930	0.736	0.822	0.343	0.271	0.854
Gradient Boosting	0.708	0.971	0.927	0.708	0.803	0.343	0.262	0.840

The model with the best metrics was chosen and then trained again on the training set and evaluated on the test set. Variable importance was assessed to evaluate which predictors were most influential in predicting depression.

▶ **Exercise Evaluation**

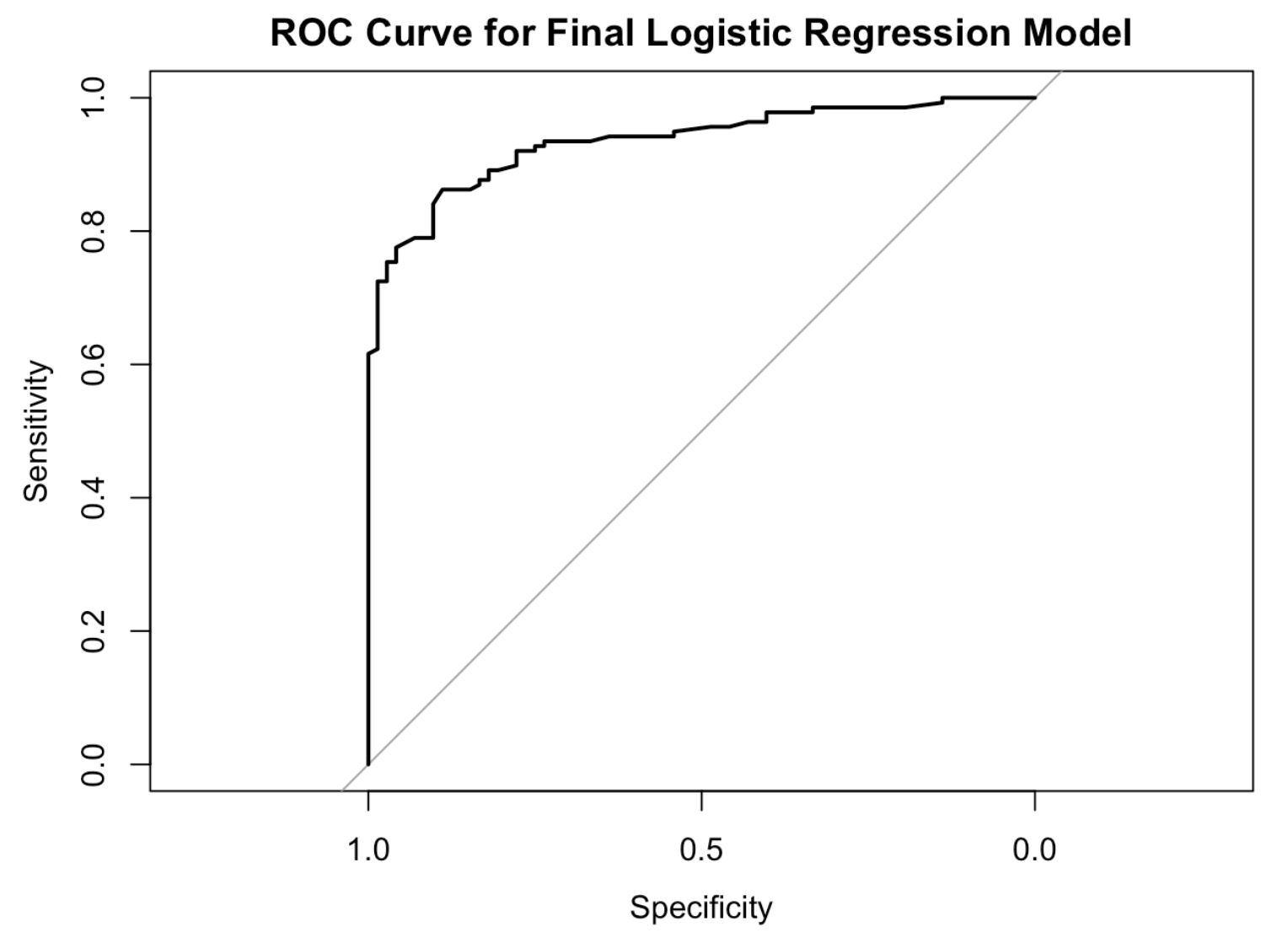
In order to evaluate whether exercise had an effect on depression, the non-parametric Kruskal-Wallis test was conducted to see if there was a significant difference in the depression score between the varying treatments.

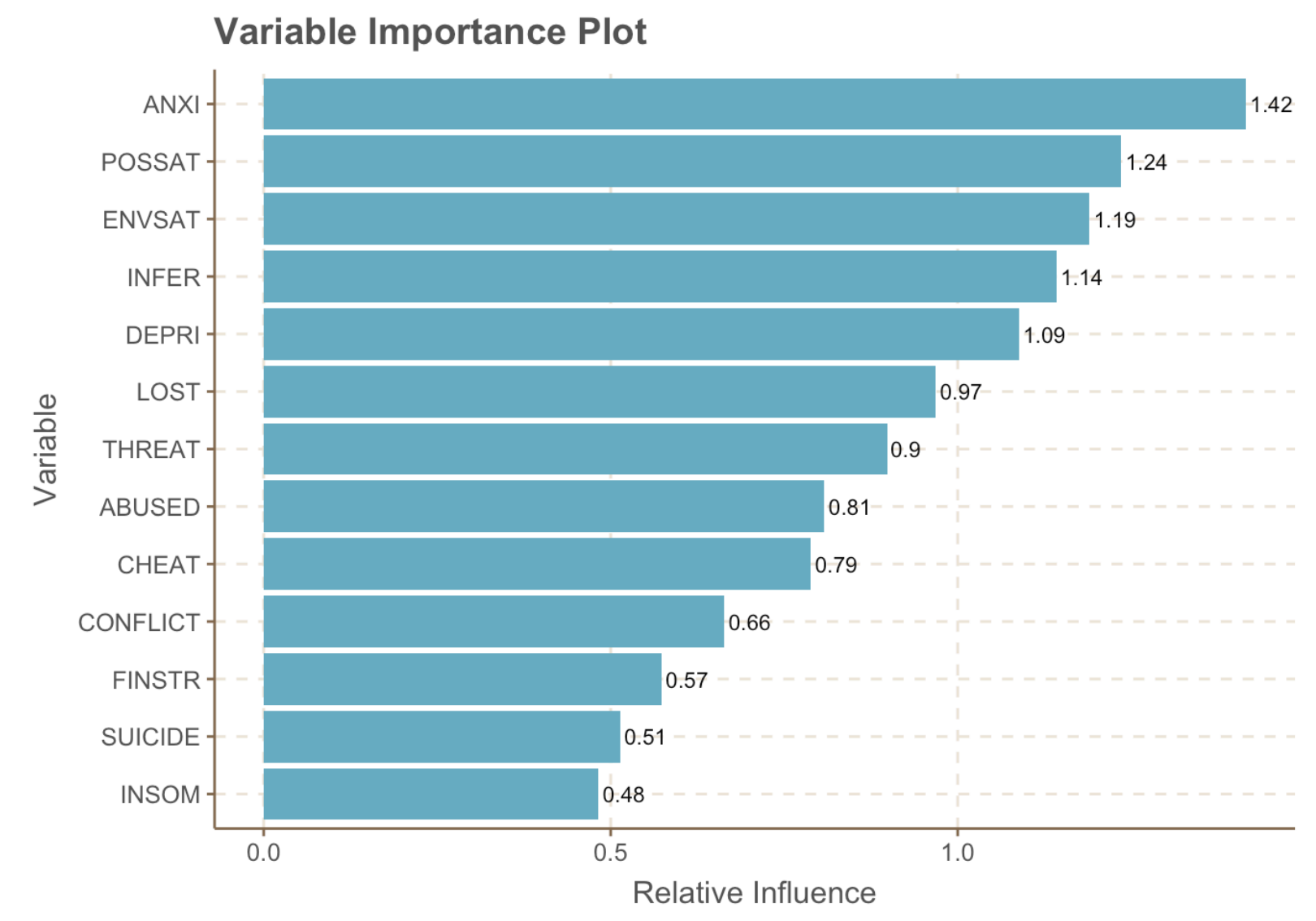
Results

The logistic regression model performed the best compared to the other two models. Below are the performance results on the test set.

Best Logistic Regression Results

Metric	Values
AUC	0.9346316
Sensitivity	0.7500000
Specificity	0.9275362
Pos Pred Value	0.8437500
Neg Pred Value	0.8767123
Precision	0.8437500
Recall	0.7500000
F1	0.7941176
Prevalence	0.3428571
Detection Rate	0.2571429
Detection Prevalence	0.3047619
Balanced Accuracy	0.8387681



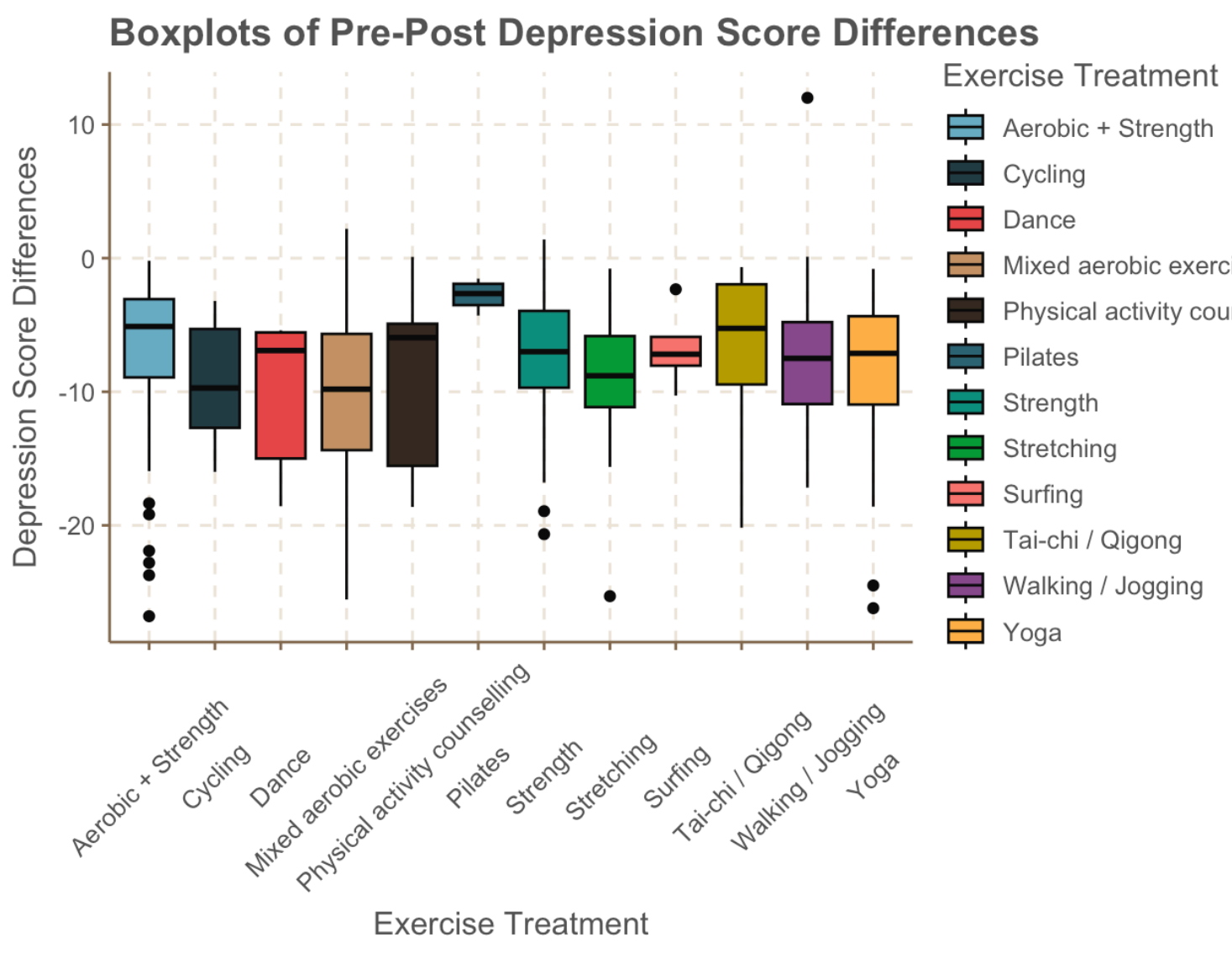


ANXI: Whether a person recently feels anxiety.
POSSAT: Whether a person is satisfied with their position or academic achievements.
ENVSAT: Whether the participant is satisfied with their living environment or not.
INFER: Whether a person suffers from inferiority complex.
DEPRI : Whether a person feels that they have been deprived of something they deserve.

Logistic Regression Coefficients

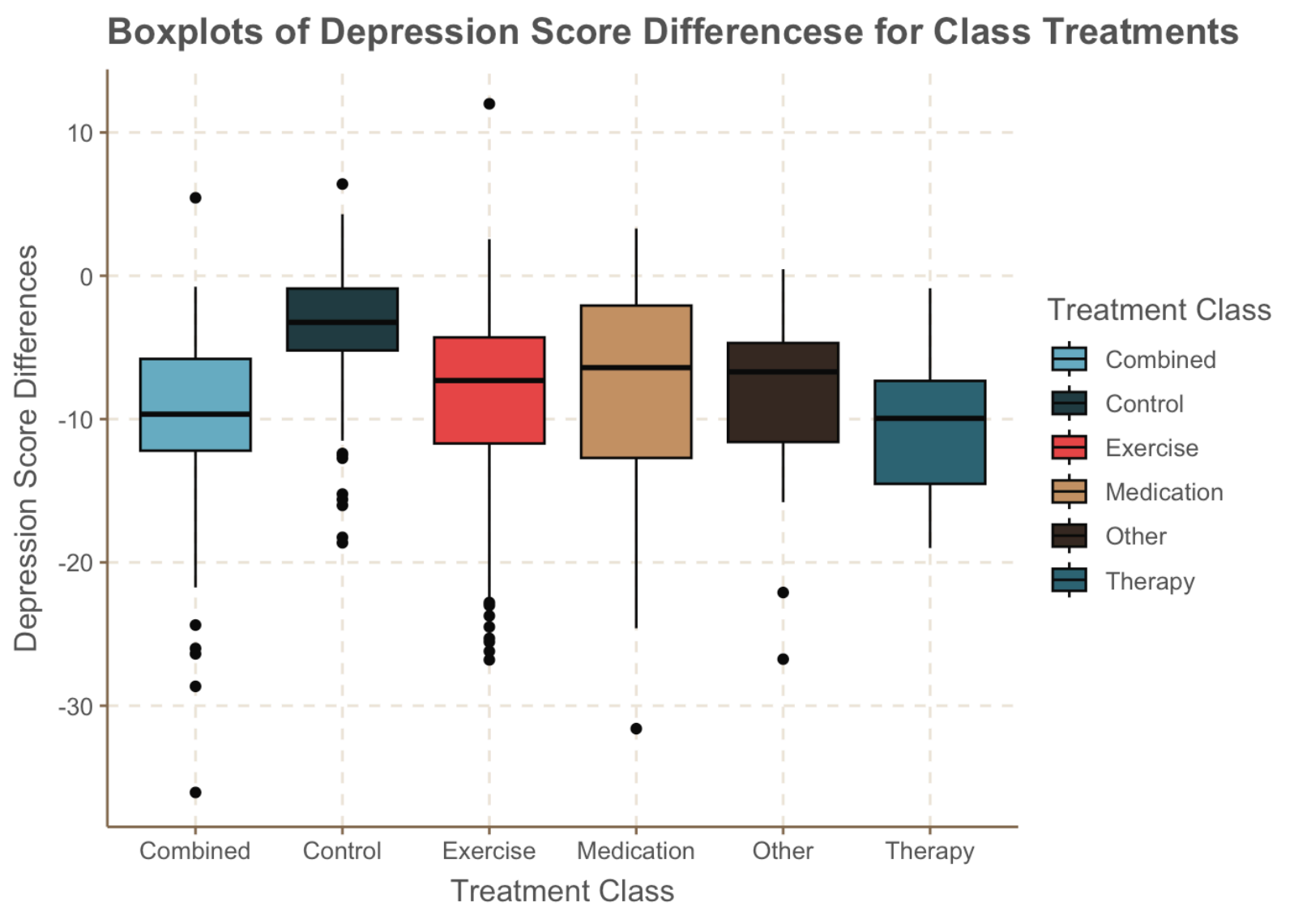
ANXI	POSSAT	ENVSAT	INFER	DEPRI
1.415251	-1.235814	-1.189952	1.142114	1.087965

Excersice and Depression Results



Kruskal-Wallis Test Treatment

Test	Statistic	df	p_value
Kruskal-Wallis	27.79351	11	0.0034815

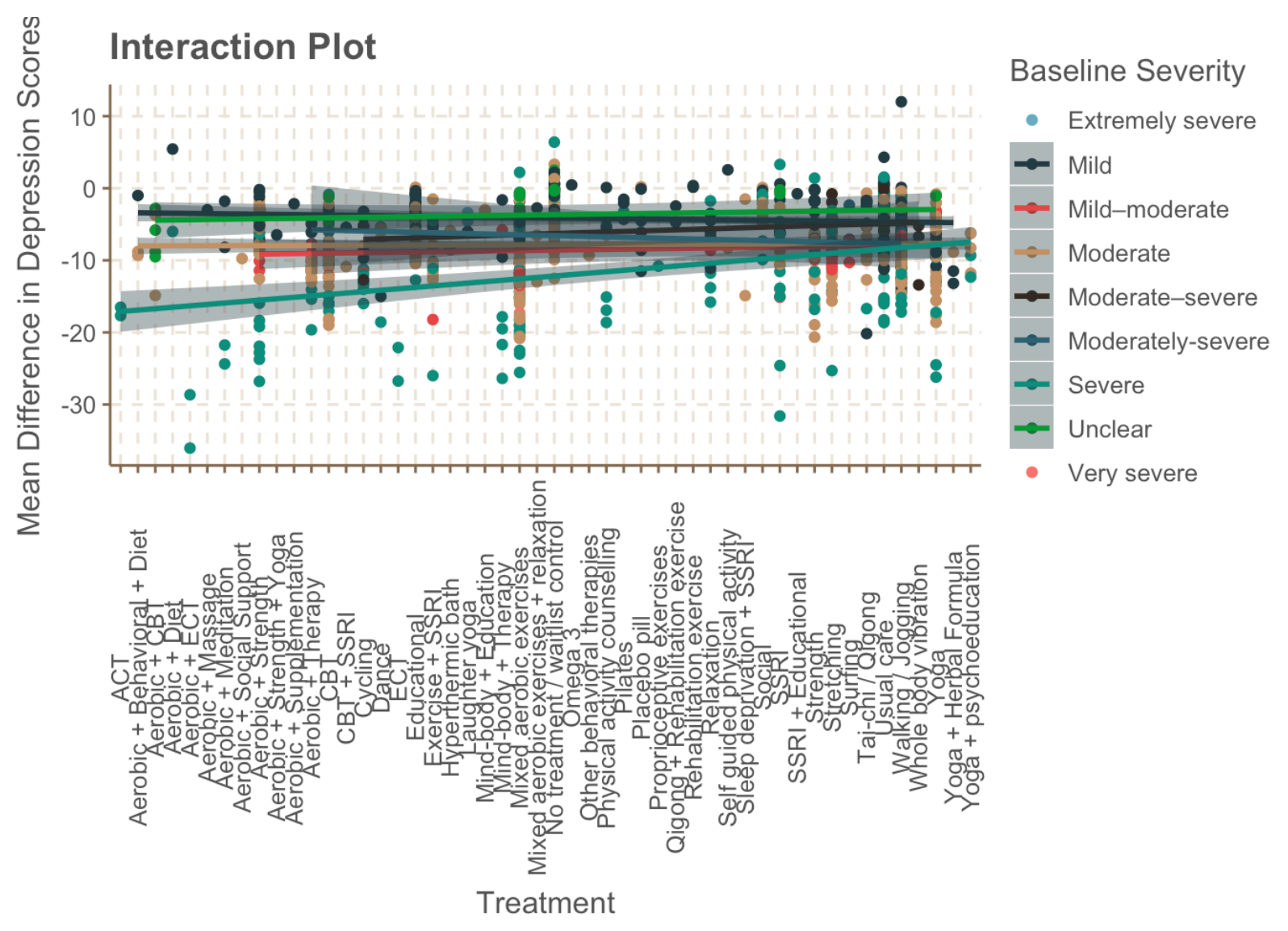


Pairwise Wilcox Test Treatment Class

	Combined	Control	Exercise	Medication	Other
Control	0.0000000	NA	NA	NA	NA
Exercise	0.3084429	0.0000000	NA	NA	NA
Medication	1.0000000	0.0465334	1.0000000	NA	NA
Other	1.0000000	0.0003875	1.0000000	1	NA
Therapy	1.0000000	0.0000000	0.2053394	1	1

Significant Interactions

	estimate	std. Error	t-value	p-value
(Intercept)	-18.028333	8.914423	-2.022378	0.0435221
trtAerobic + Diet	11.090000	5.348654	2.073419	0.0385033
trtAerobic + ECT	-15.260000	4.367158	-3.494264	0.0005057
trtAerobic + Massage	17.073333	8.362470	2.041662	0.0415644
trtAerobic + Supplementation	17.923333	8.362470	2.143306	0.0324370
trtMind-body + Education	13.526667	5.348654	2.528985	0.0116607
trtOmega 3	20.533333	8.362470	2.455415	0.0143175
trtPilates	13.840000	5.348654	2.587567	0.0098686
trtPlacebo pill	16.793333	6.428282	2.612414	0.0091859
trtQigong + Rehabilitation exercise	16.488333	7.771414	2.121665	0.0342212
trtRehabilitation exercise	20.328333	7.771414	2.615783	0.0090967
trtSSRI + Educational	19.313333	8.362470	2.309525	0.0212084
trtStrength	8.482857	3.501516	2.422624	0.0156649
trtStretching	7.822222	3.413964	2.291243	0.0222494
trtUsual care	10.858800	3.209192	3.383655	0.0007557
trtAerobic + Meditation:baseline_severityMild	21.043333	7.771414	2.707787	0.0069409
trtAerobic + Strength:baseline_severityMild	15.589567	6.664752	2.339107	0.0196139
trtMind-body + Therapy:baseline_severityMild	18.743333	7.458341	2.513070	0.0121950
trtPhysical activity counselling:baseline_severityMild	16.700000	7.351021	2.271793	0.0234056
trtTai-chi / Qigong:baseline_severityMild	17.875655	8.382808	2.132418	0.0333244
trtWalking / Jogging:baseline_severityMild	15.568917	7.471975	2.083641	0.0375597
trtExercise + SSRI:baseline_severityMild–moderate	-16.659871	5.928498	-2.810134	0.0050922
trtSSRI:baseline_severityMild–moderate	-13.491500	6.550736	-2.059540	0.0398167
trtStrength:baseline_severityMild–moderate	-9.854762	4.619800	-2.133158	0.0332635
trtStretching:baseline_severityMild–moderate	-10.384127	4.190276	-2.478149	0.0134444
trtUsual care:baseline_severityMild–moderate	-13.617371	4.749687	-2.867004	0.0042699
trtAerobic + Strength:baseline_severityModerate	10.616710	3.918766	2.709197	0.0069117
trtPhysical activity counselling:baseline_severityModerate	10.555000	5.348654	1.973394	0.0488488
trtNo treatment / waitlist control:baseline_severitySevere	10.605956	3.574086	2.967459	0.0031065
trtWalking / Jogging:baseline_severitySevere	9.179329	3.947629	2.325276	0.0203459



Challenges

Finding publicly available datasets on depression can be a difficult task. In most cases, the data are collected in such a way as to analyze a specific aspect of depression and not to provide a general overview of the factors of depression.Some expected frequencies in the contingency tables were small, and therefore the p-values from the chi-squared test are not exact but approximations. In the second dataset, there is a significantly high amount of missing values in many columns. The dataset seems to have multiple columns with different types of strings used to symbolize missing values. This had to be addressed, especially in the age column. The column mean_diff did not follow a normal distribution, and therefore a non-parametric test had to be used to test whether there was a significant difference in medians between the treatment and class groups.

Conclusion

- The logistic regression model performed the best in predicting depression. The tuned hyperparameters are
  - **alpha**: 0.2 (Elastic Net)
  - **Lambda**: 0.04132012
- The final logistic regression model had a balanced accuracy of 0.8387681 a sensitivity score of 0.7500, a specificity score of 0.9275362 and F1 score of 0.7941176.
- The top five most influential predictors on the response variable depression are: ANXI, POSSAT, ENVSAT, INFER, and DEPRI.
- The non-parametric Kruskal-Wallis test yielded a p-value of 0.003482, indicating that we can reject the null hypothesis. This suggests there is evidence of a significant difference between the median depression scores across the treatment groups.
- The pairwise comparison shows that there is a significant difference in the depression score for the different classes of treatments and the control group however it does not indicate that there is a significant difference between the treatments themselves.
- A Generalized linear Model was used to examine various treatment effects on the difference between pre and post intervention depression scores. Several treatments demonstrate significant reduction in depression scores.
  - **trtAerobic + ECT**: Estimate = -15.26, indicating a substantial decrease in depression severity.
  - **trtExercise + SSRI: baseline_severityMild–moderate**: Estimate = -16.66, suggesting a strong reduction in depression scores for individuals with mild to moderate baseline severity. +**trtStretching: baseline_severityMild–moderate**: Estimate = -10.38, indicating an improvement for individuals with mild to moderate baseline severity.
  - Treatments such as **trtExercise + SSRI** and **trtStretching** for individuals with baseline severity of mild to moderate show significant negative effects, indicating that they lead to the largest decreases in depression symptoms.

References

[1] Noetel M, Sanders T, Gallardo-GÃ³mez D, Taylor P, del Pozo Cruz B, van den Hoek D et al. Effect of exercise for depression: systematic review and network meta-analysis of randomised controlled trials BMJ 2024; 384 :e075847 doi:10.1136/bmj-2023-075847 [2]Md. Sabab Zulfiker, Nasrin Kabir, Al Amin Biswas, Tahmina Nazneen, Mohammad Shorif Uddin,An in-depth analysis of machine learning approaches to predict depression,Current Research in Behavioral Sciences,Volume 2,2021,100044,ISSN 2666-5182,https://doi.org/10.1016/j.crbeha.2021.100044. (https://www.sciencedirect.com/science/article/pii/S2666518221000310) [3]Belgiu, M., & Drăguţ, L. (2016). Random forest in remote sensing: A review of applications and future directions. ISPRS Journal of Photogrammetry and Remote Sensing, 114, 24-31. https://doi.org/10.1016/j.isprsjprs.2016.01.011 [4]Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. Frontiers in Neurorobotics, 7. https://doi.org/10.3389/fnbot.2013.00021 [5]Menard, S. (2002).Applied logistic regression analysis(No. 106). Sage. [6] Müller, Marlene. (2004). Generalized Linear Models. 10.1007/978-3-642-21551-3_24.