Project Reflection

Yingxi Kong

My Github Portfolio link is: https://github.com/nikivivi9/Practical-Data-Analysis-Portfolio.git.

Revisiting my earlier projects with fresh eyes has been an enriching and reflective experience, providing an opportunity to evaluate my progress and further improve through past works. Throughout this process, I've taken the time to carefully review feedback, identify areas for enhancement, and apply new methods and skills I've developed during the course. This reflection focuses on two projects, the exploratory data analysis and the regression analysis, where I made targeted improvements to strengthen the analysis, refine visualizations, and address gaps in methodology. I will detail the changes I made, explain the reasons behind them, and share the valuable lessons I've learned through this process. This journey has not only allowed me to refine my technical skills but also taught me the importance of adaptability and the value of revisiting work.

For the exploratory data analysis which aims to investigate the association between weather characteristics and marathon performance by age and sex, I revised the modeling approach for aim 3 to investigate the most impactful weather parameter affecting marathon performance. Initially, I applied a mixed-effects model with random intercept by age to quantify the significance of weather effects. This choice was informed by exploratory data analysis, which revealed a general decline in marathon performance with increasing age across both genders. The mixed-effects model effectively accounted for variability across age groups but made it more challenging to directly evaluate the joint effects of weather parameters, age, and sex.

To address these limitations, I transitioned to using linear regression, incorporating interaction terms between age, sex, and weather parameters, and performing a backward model selection procedure to identify the best model with results shown in the following table. Additionally, I included a polynomial term for age to capture the non-linear relationship between age and performance observed during the EDA process. This change simplified the model structure while preserving the ability to explore how these factors jointly influence marathon performance. By explicitly modeling interactions, the linear regression enabled clearer interpretation of the effects of weather parameters across different age and sex groups. Through this revision, I learned the importance of integrating insights from the EDA process into the modeling approach, as I ignored the non-linear relationship between people's age and marathon performance which might lead to biases. In addition, it is important to maintain the balance between flexibility and interpretability of our model as linear regression, while simpler, was sufficient to achieve our study goals.

Table 1: Coefficient Estimation of Best Model

Variable	Estimate	Standard Error	T Statistics	P Value
(Intercept)	114.2737167	3.6309472	31.472150	0.0000000
sexMale	-9.5066837	0.3770221	-25.215187	0.0000000
WBGT	0.2127369	0.1102679	1.929273	0.0537225
rh	0.0622261	0.0358442	1.736015	0.0825891
Wind	0.1548375	0.1355968	1.141897	0.2535217
SRWm2	0.0044593	0.0033216	1.342501	0.1794613
age	-5.5900725	0.0917711	-60.913185	0.0000000
I(age^2)	0.0798899	0.0006069	131.636557	0.0000000
WBGT:age	0.0078914	0.0022456	3.514112	0.0004430
rh:age	-0.0026038	0.0007202	-3.615237	0.0003014
Wind:age	-0.0045421	0.0026929	-1.686679	0.0916933
SRWm2:age	-0.0001933	0.0000677	-2.855631	0.0043031

For the regression analysis which aims to identify predictors and moderators of behavioral treatment on smoking abstinence for people diagnosis with Major Depressive Disorder (MDD), I originally employed cross-validated Lasso regression to identify potential predictors and moderators of behavioral treatment on End of Treatment (EOT) smoking cessation. During exploratory analysis, I observed that several continuous variables were highly skewed, which led me to apply appropriate transformations to these variables to approximate normality before performing Lasso regression on each imputed dataset. However, transformations can reduce the interpretability of the results.

I expanded the analysis with a Lasso regression on the non-transformed data (original imputed data), allowing for a direct comparison of the selected coefficient estimates, AUC, and calibration plots between the transformed and non-transformed datasets. This additional step provided insights into whether transformations meaningfully impacted model performance and reliability, helping to evaluate the trade-offs between interpretability and predictive accuracy. From the results, we found some of key skewed variables show significant in the transformed model while they are dropped in the non-transformed model. In addition, applying transformation, the AUC for our train set increases from 0.79 to 0.81 and that for our test set increases from 0.74 to 0.75. Although the improvement in AUC was modest, we chose to retain the transformations as they better captured the underlying relationships between the predictors and the outcome, enhancing the reliability of the model.

Although we decided to retain the transformations, we learned the importance of systematically testing their impact. Comparing the results with and without transformations gave us a clearer understanding of how they influenced the model's performance and reliability. This revision process highlighted a practice that we should consistently apply in future analyses to ensure more thoughtful and effective modeling decisions.

Table 2: Bootstrap Lasso Model Coefficient Estimate

Variable	Pooled.Estimate	Lower.CI	Upper.CI	Siginificant.Proportion
(Intercept)	-0.4622382	-2.3170828	2.3347997	1.000
ftcd_score	-0.2238994	-0.5670841	0.0000000	0.938
mde_curr1	-0.1396745	-0.7928705	0.0000000	0.418
NMR	0.2286429	0.0000000	0.8491399	0.647
raceNon-Hispanic White	0.3644262	0.0000000	1.4701028	0.648
BA1:ftcd_score	-0.0356967	-0.2615361	0.0000000	0.362
BA1:incMore than $75,000	0.2751041	0.0000000	1.7869354	0.380
BA1:Only.Menthol1	-0.2309910	-1.1583771	0.0000000	0.445
age_ps:Var1	0.0093747	0.0000000	0.0308070	0.757
antidepmed1:Var1	0.1497947	0.0000000	0.9060176	0.382
crv_total_pq1:Var1	0.0284297	0.0000000	0.1480517	0.471
eduHigh school graduate or GED:Var1	0.3050588	0.0000000	1.3186101	0.561
ftcd.5.mins1:Var1	0.4474196	0.0000000	1.7131313	0.598
raceHispanic:Var1	-0.2201501	-1.8094060	0.2680946	0.351
raceMixed Race:Var1	0.4918142	0.0000000	2.9191000	0.520

In addition, to further evaluate the stability and reliability of the model, I incorporated a bootstrap procedure with 200 iterations across the multiple imputed datasets to summarize the pooled coefficient estimate, 95% confidence interval, and significant proportion shown in the attached table. This allowed us to assess the consistency of coefficient estimates and the robustness of variable selection. The bootstrap results mainly supported the findings from the previous transformed data, reinforcing the significance of key predictors and moderators. Moreover, the validation process through the following AUC and calibration plot demonstrated strong discriminative performance and excellent calibration, further confirming the model's ability to accurately differentiate between outcomes and reliably predict probabilities.

In revising these projects, I gained valuable insights into the importance of aligning modeling approaches with exploratory findings, balancing complexity with interpretability, and ensuring the robustness of results through systematic evaluation. For the exploratory data analysis, transitioning from a mixed-effects model to a linear regression framework with interaction terms allowed for clearer interpretation of weather impacts while maintaining the rigor of our analysis. For the regression analysis, testing transformations and incorporating a bootstrap procedure highlighted the necessity of validating modeling choices. These revisions

underscored the importance of iterative refinement, thoughtful application of statistical methods, and the integration of data-driven insights to achieve reliable and insightful results. Over the course of the semester, I have also seen significant growth in my ability to apply advanced methodologies, engage in critical thinking, and produce relatively professional, comprehensive analysis reports, enabling me to effectively communicate findings and methodologies.

Figure 8: ROC Curve (Bootstrap)

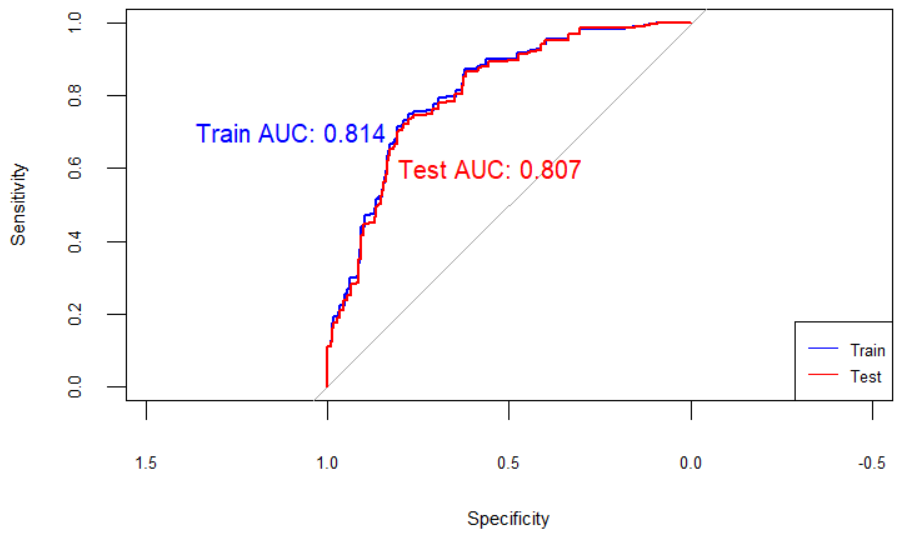
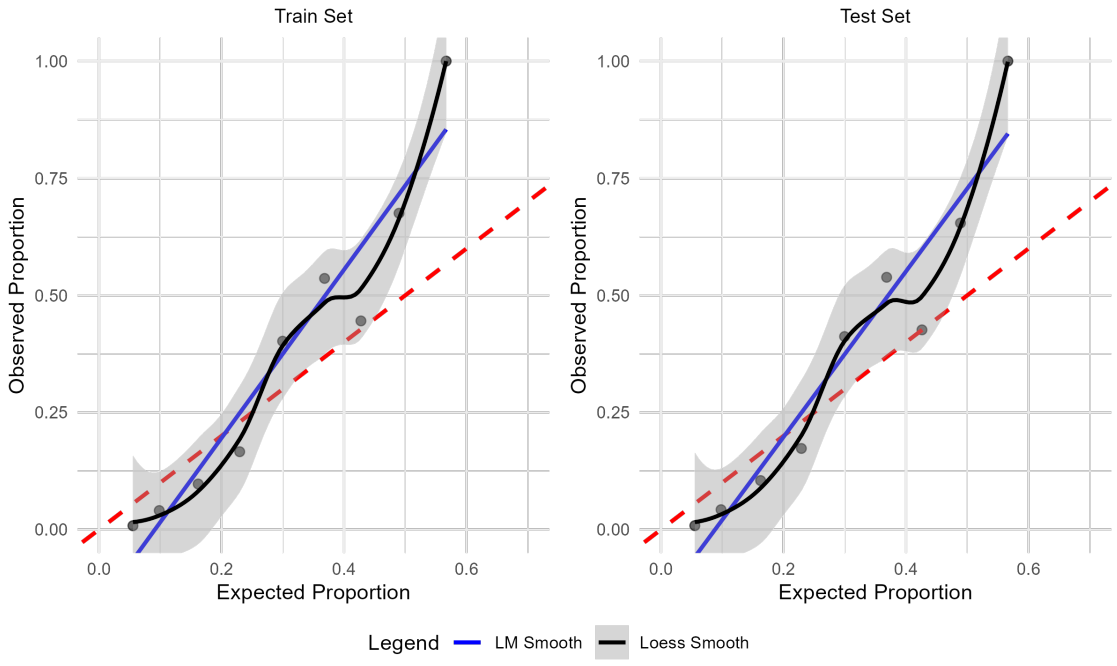


Figure 9: Calibration Plot (Bootstrap)



I would also like to express my sincere gratitude to Alice for her insightful guidance throughout this process. Her comprehensive feedback and support were instrumental in shaping my approach and helping me navigate the challenges of refining these projects. Thank you, Alice!