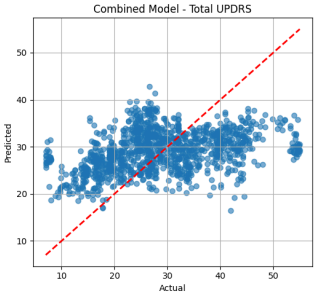
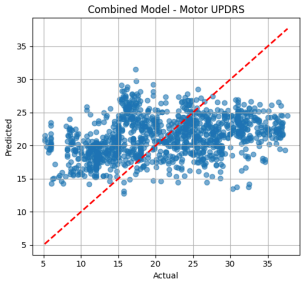
MILESTONE 3:
-HARSHIT SONI (hs5666)

Methodology:

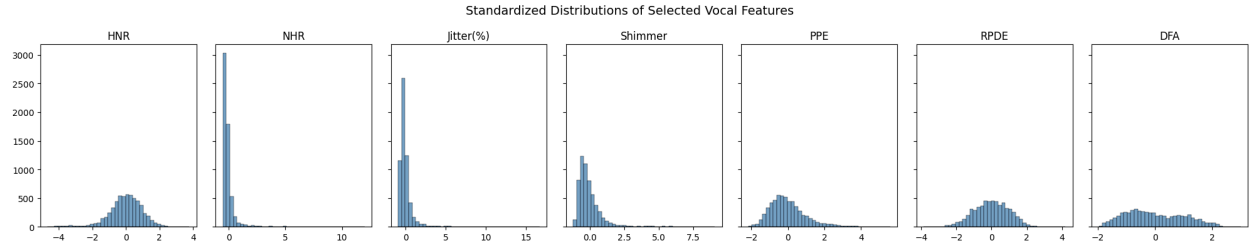




1. Linear Regression on PCA Components: PCA was performed on the full dataset using all vocal features. Top 8 components captured approximately 97.5% of the total variance. Linear regression was performed on the combined dataset using the top 8 PCA components. The model was trained to predict total UPDRS scores and evaluated using RMSE and R^2 metrics on a held-out test set.

2. REGULARIZED REGRESSION: Lasso and Ridge regression were applied on PCA-transformed data using the top 8 components that captured 97.5% of the variance. Models were trained separately for the combined, male, and female datasets. Regularized models were then compared against standard linear regression to evaluate the impact of coefficient shrinkage.

3. TEST FOR NORMALITY: Given the dataset's size (over 5000 entries), traditional statistical tests like the Shapiro-Wilk test were not suitable due to their sensitivity to minor deviations in large samples. Instead, a visual inspection approach was adopted using standardized histograms for key vocal features. This allowed for a more practical and interpretable assessment of distribution shapes, helping to determine whether the normality assumption held for further analysis.



4. NON-LINEAR REGRESSION (GRADIENT BOOSTING AND RANDOM FOREST): Gradient Boosting and Random Forest regression models were applied to the PCA-transformed dataset using the top 8 components that explained 97.5% of the total variance. Separate models were trained for the combined, male, and female subsets. Both methods were used to capture non-linear relationships that linear models may miss. Performance was evaluated using RMSE and R^2.

Group	Metric	Lasso	Linear	Ridge	Random Forest	Gradient Boosting
Combined	RMSE	9.835	9.855	9.855	6.561	8.212
	R^2	0.127	0.124	0.124	0.612	0.391
Male	RMSE	8.260	8.217	8.217	4.792	5.745
	R^2	0.276	0.284	0.284	0.756	0.650
Female	RMSE	10.687	10.668	10.668	7.716	8.770
	R^2	0.141	0.144	0.144	0.552	0.421

5. CAUSAL INFERENCE: Inverse probability weighting (IPW) was used to estimate the causal effect of selected vocal features (HNR, NHR, Jitter(%), Shimmer, PPE, RPDE, DFA) on total UPDRS scores. For each feature, a binary treatment variable was defined by splitting at the median value. Propensity scores were calculated using logistic regression on confounding variables: age, sex, and test time. These scores were used to compute observation weights, and a weighted linear regression was then performed to estimate the average treatment effect (ATE) of each feature. Confidence intervals and p-values were derived to assess the significance and direction of the effects.

Metric	HNR	NHR	Jitter(%)	Shimmer	PPE	RPDE	DFA
ATE	−2.6098	+1.0839	+1.3702	+1.2713	+2.3826	+1.5482	−1.1959
p-value	1.6816e−21	0.0001	7.2205e−07	3.8467e−06	3.764e−18	2.985e−08	2.0711e−05
95% CI	[−3.1450, −2.0747]	[0.5356, 1.6323]	[0.8287, 1.9116]	[0.7323, 1.8103]	[1.8466, 2.9187]	[1.0013, 2.0951]	[−1.7462, −0.6457]

Result:

1. Linear regression on PCA-transformed data (top 8 components) for the combined dataset yielded low predictive performance, even worse than stratified regression models.
2. Lasso and Ridge regularization showed minimal improvement, hinting towards a possible non-linear relationship between features.
3. Histogram analysis revealed non-normal distributions for most vocal features.
4. Non-linear models significantly outperformed linear ones; Random Forest achieved the best performance ($R^2 = 0.612$), followed by Gradient Boosting ($R^2 = 0.391$) in the combined setting.
5. Causal inference using IPW identified significant effects for multiple vocal features positively associated with disease severity, while HNR and DFA showed negative associations.

Analysis:

1. Linear regression methods were limited in capturing complex relationships in the data.
2. Non-normal feature distributions and residuals justify the use of non-linear models.
3. Random Forest's performance confirms non-linear patterns in vocal biomarkers.
4. Causal inference confirmed that certain features have a statistically significant directional effect on UPDRS scores, supporting their potential role as disease indicators.
5. The combined model provides a comprehensive overview but may obscure subgroup-specific patterns that stratified analysis can reveal.

Final Analysis: Analysis of various vocal features revealed that telemetric data exhibits strong non-linear relationships with UPDRS scores. Linear models were limited in capturing these patterns, while non-linear methods and causal inference highlighted key biomarkers linked to disease severity. Additionally, age and sex emerged as crucial factors influencing the progression and manifestation of the disease, especially in gender possibly due to anatomical/biological differences in voice mechanism.