MDD Model Improvement Report

This document summarizes the limitations of the initial MDD severity prediction notebook and proposes improvements to strengthen the analysis and modeling pipeline. The original objective was to predict MADRS1 values using clinical and demographic variables. Although the model runs successfully, several methodological and structural issues affect its predictive performance and generalizability.

1. Data limitations

The dataset becomes very small after cleaning. The initial dataset contains 55 rows, but row-wise deletion and data type conversion reduce the size to approximately 20 usable samples. This significantly affects the stability of evaluation metrics and increases the risk of overfitting.

Recommendation

Use targeted imputation strategies and avoid dropping entire rows. Consider:

- mean or median imputation for continuous variables
- mode imputation for categorical variables
- retaining rows with partial missingness when possible

This preserves more training samples.

2. Feature–target relationship concerns

The current model uses MADRS2 as a predictor for MADRS1. In many longitudinal clinical designs, MADRS2 represents follow-up outcomes and should not be used to predict baseline scores. Although the model technically allows this, the interpretation becomes ambiguous.

Recommendation

Consider redesigning the task:

- Predict MADRS2 from baseline information
- Predict change scores such as MADRS2 − MADRS1
- Convert MADRS1 into a binary or ordinal severity label and apply classification models

These alternatives align more closely with clinical interpretation.

3. Model evaluation challenges

A single train–test split on a very small dataset produces unstable metrics. Negative R^2 values in the baseline linear regression indicate that the model does not generalize well.

Recommendation

Use cross-validation to obtain more stable performance estimates.

Example

Five-fold cross validation using:

- Linear regression
- Random forest
- Gradient boosting regressors

Average performance across folds provides more meaningful evaluation than a single split.

4. Improving model expressiveness

The Random Forest model captures nonlinear patterns but may still struggle due to limited feature engineering.

Recommendation

Experiment with:

- polynomial features
- interaction terms
- additional tree-based models (XGBoost, LightGBM)
- standardized or normalized numeric variables

These techniques can help extract additional signal from small datasets.

5. Improving interpretability

The existing visualizations are helpful but can be expanded.

Recommendation

Add:

- partial dependence plots
- SHAP value explanations
- correlation heatmaps with clearer labels
- EDA separating different demographic groups

These provide clearer insights into relationships between predictors and MADRS1.

6. Notebook structure improvements

The current notebook contains several unrelated code cells and repeated imports. Cleaning the notebook improves readability and professionalism.

Recommendation

Organize the notebook into sections:

1. Load packages
2. Load and inspect data
3. Clean and preprocess
4. Exploratory data analysis
5. Baseline model
6. Random forest model
7. Feature importance
8. Conclusion

Remove irrelevant demonstration cells (e.g., basic numpy operations).

Conclusion

The current notebook provides a complete pipeline from data loading to modeling. However, improvements in data retention, model design, evaluation stability and notebook organization can substantially elevate the quality of the analysis. Future work may also involve reframing the task to predict follow-up outcomes or using advanced model interpretation techniques to better understand clinical predictors.