# Modeling Summary Report

Objective:

The goal of this project is to predict the target variable `y` using a subset of features from `var\_1` to `var\_10`.

1. Data Exploration and Preparation:

- The dataset was loaded and examined using basic statistics, shape, and head samples.

- A correlation matrix was generated to visualize linear relationships between features.

- Feature importance was assessed using a Random Forest Regressor.

- Based on the importance scores, the top 3–15 most influential features were selected.

2. Modeling Approach:

- A stacking ensemble model was developed for regression.

- Base models: RidgeCV and Gradient Boosting Regressor.

- Meta-model: LassoCV.

- All models were implemented using scikit-learn’s pipeline, with StandardScaler used for feature scaling.

- Hyperparameter tuning was achieved through cross-validation (cv=5).

3. Model Evaluation:

- The model performance was measured using the following metrics:

- R² Score on test data.

- Cross-validated R² score (5-fold CV).

- Mean Squared Error (MSE).

- Residual plots and actual vs. predicted plots were included for diagnostic analysis.

4. Interpretability:

- SHAP (SHapley Additive exPlanations) was used to interpret the predictions from the Gradient Boosting model.

- Visualizations such as bar plots, beeswarm, and waterfall plots were produced.

5. Deployment Interface:

- A Streamlit app was developed for:

- Uploading and processing new datasets.

- Displaying feature importances and SHAP insights.

- Making predictions and downloading results.

- Trying manual inputs for real-time prediction.

Conclusion:

The project provides a complete pipeline from data exploration to model deployment, ensuring interpretability and ease of use through a visual dashboard. The stacking model performed robustly and allowed users to engage with the predictions interactively.