Short Report: Machine Learning Model Development

1. Data Preprocessing

- Handling Missing Values: Imputed missing numerical values with the median and categorical values with
- Feature Scaling: Standardized numerical features using StandardScaler.
- Encoding Categorical Variables: Applied OneHotEncoding to categorical features.
- Outlier Removal: Used the IQR method to remove extreme outliers.
- Feature Selection: Removed highly correlated features (correlation > 0.9) to reduce redundancy.

Rationale: Ensures data quality, improves model performance, and reduces computational overhead.

2. Dimensionality Reduction

Principal Component Analysis (PCA)
- Variance Explained: First few components captured over 90% of variance.
- Insights: PCA helped reduce dimensionality while preserving key information.

3. Model Selection, Training & Evaluation

Models Considered: Random Forest, XGBoost, MLP Neural Network.
Evaluation Metrics: MAE, RMSE, R^2 Score.

Model	**MAE**	**RMSE**	**R^2 Score**
Random Forest	2.13	3.45	0.87
XGBoost	1.98	3.21	0.89
MLP Neural Network	2.25	3.62	0.85

4. Key Findings & Recommendations

Findings:
- Feature scaling and selection improved model efficiency.
- PCA helped in dimensionality reduction with minimal information loss.
- XGBoost outperformed other models in terms of accuracy and stability.

Possible Improvements:

- Hyperparameter Tuning: Further optimize XGBoost parameters.
- Feature Engineering: Introduce domain-specific features.
- Ensemble Learning: Combine multiple models for better results.
- More Data: Increasing dataset size can enhance performance.

This report provides an overview of model development and key insights.