Report for regression exercise

Data preprocessing

General preprocessing

- The data set has fourteen features and the target is "**Calories_Burned**".
- There was no missing data and duplicated rows in the data set.
- The outliers have been visualized by box-plot and values **below** Q1 - 1.5*IQR are replaced with Q1 - 1.5*IQR values **above** Q3 + 1.5*IQR are replaced with Q3 + 1.5*IQR.
- The distribution of categorical features is shown by a pie-plot and distribution of numerical features is shown by a hist-plot
- The relationship between the features and the target for the categorical features is visualized by a bar plot (showing the mean 'calories_burned' for each category.) and for the numerical features the relationship of the feature and the target is shown with a scatter plot.
- StandardScaler of sk-learn library has been used to scale the features.
- Highly correlated features (a correlation of more than 0.85) have been identified and the feature with the most correlation with the target has been kept and the

other has been eliminated due to dimensionality reduction. (e.g. ‘BMI’ and ‘weight’ were highly correlated; BMI has been dropped due to lower correlation with target comparing to correlation of weight and target).

- The final Clean_Data got sixteen features.

Specific preprocessing

- Features like ‘Gender’ and ‘workout-type’ have been one hot encoded.
- ‘Water-intake’ has been divided into bins by the length of 0.5 due to the spikes in the hist-plot.

Cross Validation Evaluation

The test R2 and test MAE for each fold has been shown for the Randomforest and linearregression models in the figure1 and figure 2.

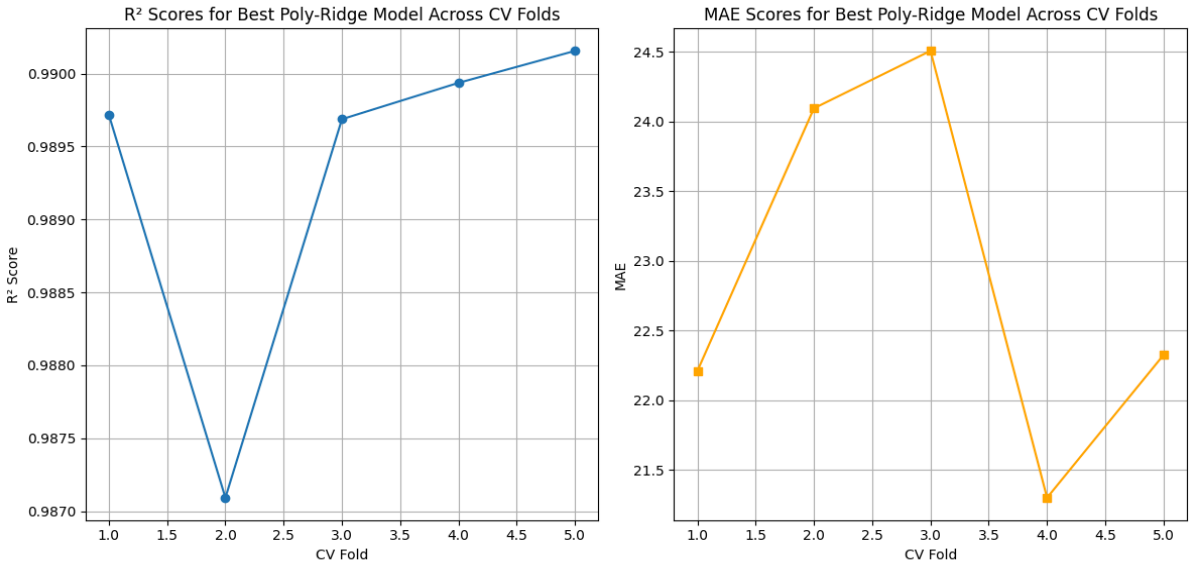

Figure 1 - R2 score and MAE for best randomforest model.

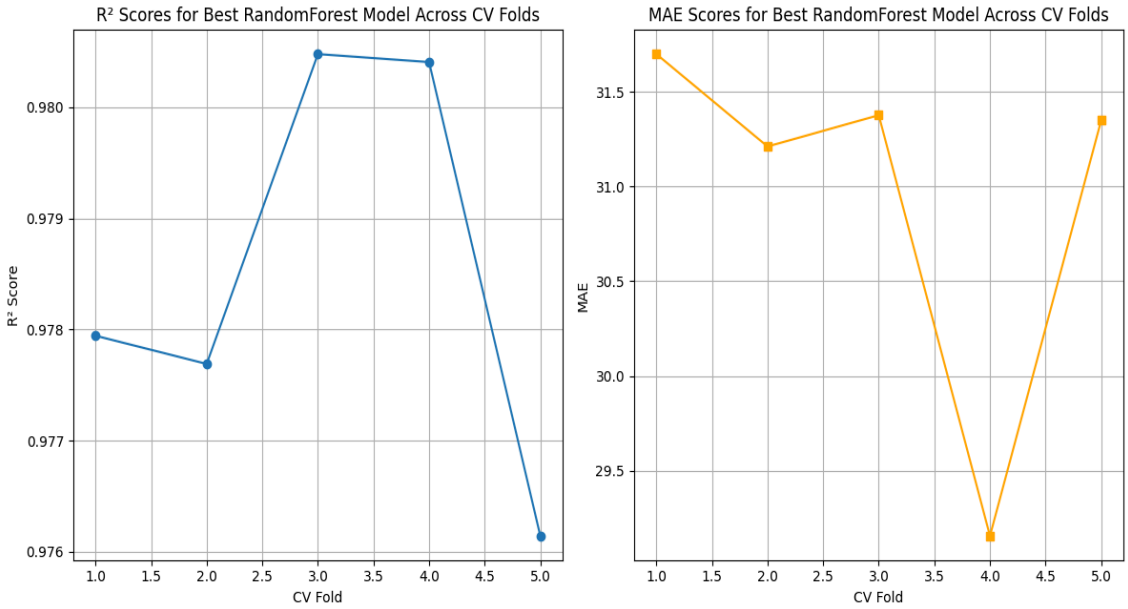

Figure 2 - R2 score and MAE for best linear regression model.

- Both models are reasonably consistent, but the Polynomial-Ridge model is **more stable across CV folds**, especially in R^2 and MAE values.

Feature importance plot

The top five important features are shown in the figure 3 the bars are annotated with the correlation of each feature with the target.

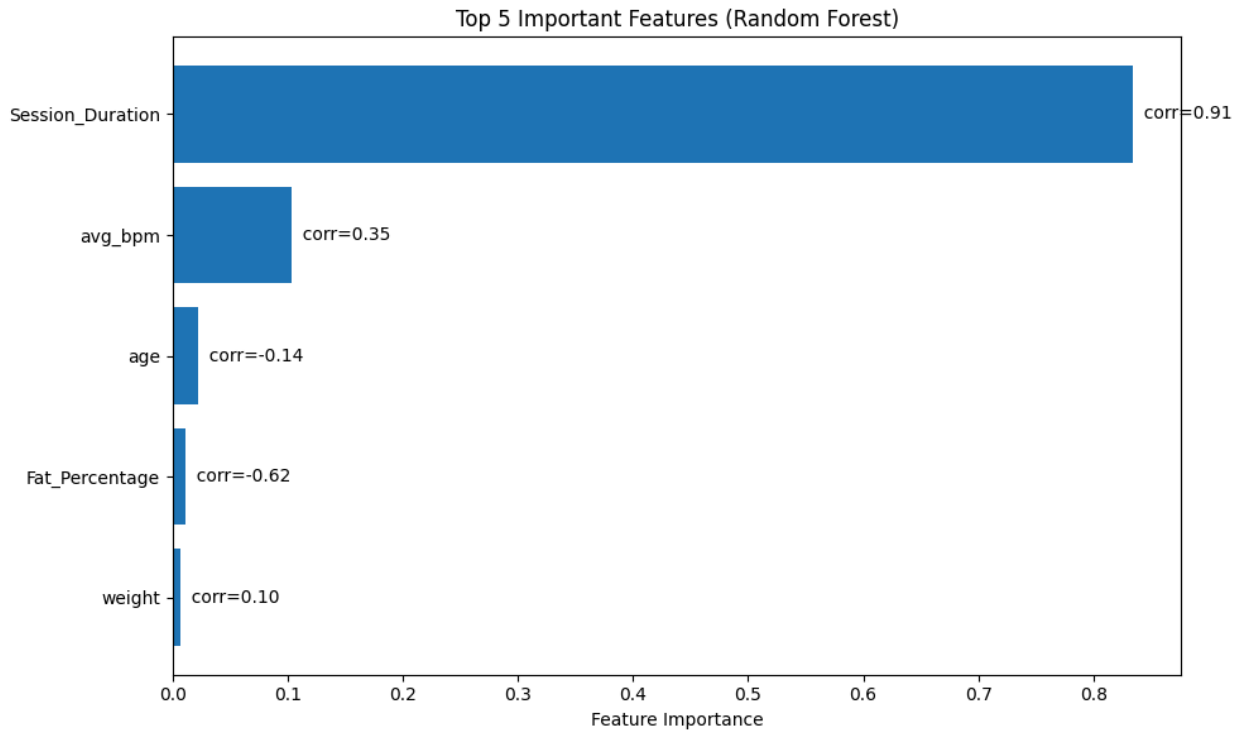

Figure 3 - top five important features.

Model performance

The performance of randomforest and linear regression models is shown in the table 1.

Table 1 - the performance of each model

Model	R2 Score	MAE
Random Forest	0.9661	39.2506
Linear Regression	0.9893	22.8864

Stats Mode

- Dependent variable: **Calories_Burned**
- Number of observations: 973
- Model fit is excellent with **R-squared = 0.979**, indicating the model explains about 97.9% of the variance in calories burned.
- The overall model is highly significant (**F-statistic = 2833, p < 0.001**).

Key coefficients:

- **Intercept:** ~862.7 (significant)
- **Age:** Negative effect (-41.5), highly significant ($p < 0.001$)
- **Gender:** Positive effect (~86.2), highly significant ($p < 0.001$)
- **Weight:** Negative effect (-12.3), not statistically significant ($p = 0.13$)
- **Height:** Positive effect (~8.24), marginal significance ($p = 0.068$)

- **Average BPM:** Strong positive effect (~89.5), highly significant ($p < 0.001$)
- **Resting BPM:** Small positive effect (~2.8), significant ($p = 0.026$)
- **Session Duration:** Strong positive effect (~244.8), highly significant ($p < 0.001$)
- Other workout types (HIIT, Strength) showed no significant effects.
