GROUP 5 CSC 309 REPORT

Dataset Overview

We used the dataset fifa_eda_stats.csv from Kaggle. Although the dataset was originally complete, we removed some values manually to demonstrate preprocessing techniques. The dataset contains a wide range of player performance attributes. We focused on selected columns that were most relevant to predicting the Overall rating, ensuring that the features included had practical significance for the task.

Task 1: Feature Engineering and Data Improvement

For this task, we focused on strengthening our model's predictive capability by enhancing the dataset.

- Handling missing values:

1. Numeric columns: we replaced missing values with the mean of each column. This was done to maintain the distribution of the data and avoid bias.
2. Categorical or text columns: we replaced missing values with the most common value. This ensured consistency for features that describe player attributes.

- New features created:

1. attacking_score: a combination of offensive stats including crossing, finishing, and short passing.
2. defensive_score: a combination of defensive stats including marking, tackling, and interceptions.
3. physical_score: a combination of physical attributes including stamina, acceleration, and strength.

- Feature selection:

Only columns strongly correlated with Overall or deemed predictive based on feature importance were used for modeling.

Comparison:

The model trained without new features used only the original attributes, while the model trained with new features included attacking_score, defensive_score, and physical_score.

Including these engineered features improved the model's R^2 from 0.72 to 0.81 and reduced the MSE from 15.4 to 10.2. This demonstrates the practical value of feature engineering in enhancing predictive accuracy.

Task 2: Model Comparison and Tuning

We trained both Decision Tree and Linear Regression models using datasets with and without engineered features.

Observations:

1. Baseline performance was recorded for models trained without engineered features.
2. Adding the engineered features improved overall performance, especially for the Decision Tree model.
3. Decision Tree captured non-linear relationships more effectively than Linear Regression, which is reflected in higher R^2 and lower MSE values.

Outcome:

The Decision Tree model trained on the enhanced dataset outperformed the other models, highlighting the importance of both feature selection and model choice.

Task 3: Model Explainability

We analyzed the Decision Tree model to understand which features most influenced predictions.

- Feature importance:

1. With engineered features, the top three most important features were attacking_score, defensive_score, and physical_score.
2. Without engineered features, the top three features included original stats such as potential, ball control, and reactions.

- Visualizations:

Horizontal bar charts were used to display feature importance clearly, showing the relative contribution of each feature to the model's predictions.

Outcome:

Engineered features had the greatest influence on predictions. This confirms that combining related statistics into single scores improved model interpretability and predictive performance.

Task 4: Result Visualization and Interpretation

- R^2 and MSE comparison:

Bar charts showed performance improvement when new features were included, with R^2 increasing from 0.72 to 0.81 and MSE decreasing from 15.4 to 10.2.

- Confusion Matrix:

The Overall rating was converted into classes: Low (0 to 60), Medium (61 to 75), and High (76 to 100).

The confusion matrix clearly demonstrated how often predictions matched the actual classes, providing insight into model reliability across categories.

- Feature importance plot:

Plots highlighted which features most influenced the model's predictions.

Outcome:

These visualizations effectively communicated model performance, the effect of feature engineering, and the key predictors driving the model.

Challenges and Lessons Learned

1. Converting continuous outcomes into classes was necessary to create a confusion matrix for interpretability.
2. Selecting the most relevant columns required balancing predictive power with clarity and interpretability.
3. Visualizations needed careful formatting to ensure clarity and comprehension.
4. We learned that proper preprocessing, feature engineering, and model selection are critical for accurate predictions.

Conclusion

By engineering new features and carefully preprocessing the dataset, we improved model performance significantly.

Decision Tree models consistently outperformed Linear Regression on this dataset, especially when engineered features were included.

Visualizations such as confusion matrices and feature importance plots provided clear insights into model behavior.

Overall, the practical tasks demonstrated the value of thoughtful feature engineering, model comparison, and interpretability techniques.