Analysis on Crossfit Games championship

2019

Himakar Raju Gunda

Individual Project

**Individual Project Report**

**Project on**

CrossFit Games 2019

**Overview on CrossFit Games**

The CrossFit Games is an annual athletic competition owned and operated by CrossFit, LLC. Athletes compete in a series of events at the Games, which may be various standard CrossFit workouts consisting of metabolic conditioning exercises, weightlifting, and gymnastics movements, as well as a range of activities from other sports such as swimming, road cycling and strongman. The events generally are not revealed before the Games, can include unexpected elements to challenge the athlete’s readiness to compete, and they are designed to test the athlete’s fitness using CrossFit's own criteria. Winners of the CrossFit Games earn cash prizes and the title of “Fittest on Earth”.

**Expectations**

1. Understanding of the data set
2. Relations that exist between the datasets in case there are multiple data set
3. What is the current information extracted from the dataset
4. What can be done :
   1. To improve athlete performance in terms of decisions

**Key points understood**

By Analysing this CrossFit championship can provide a wealth of insights across various-dimensions. Here are some key areas that can be explored:

1. **Athlete Performance :**
   * **Physical Metrics :** Assessing athletes’ strength, endurance, speed, and agility.
   * **Consistency :** Evaluating how consistently athletes perform across different events.
   * **Improvement :** Tracking performance improvements over time.
2. **Event Programming :**
   * **Variety and Balance :** Analysing the diversity and balance of events to ensure a comprehensive test of fitness.
   * **Difficulty :** Assessing the difficulty level of events and how they challenge different fitness domains.
3. **Demographic Trends :**
   * **Age and Gender:** Examining how age and gender impact performance. For example, younger athletes within an age bracket tend to perform better.
   * **Geographic Representation:** Analysing the diversity of participants from different regions.
4. **Injury Rates :**
   * **Common Injuries:** Identifying common injuries and their causes.
   * **Prevention Strategies:** Developing strategies to minimize injury risks.
5. **Statistical Analysis :**
   * **Performance Metrics:** Using data analysis to uncover trends and patterns in performance.
   * **Predictive Modelling:** Creating models to predict future performance based on historical data.
6. **Economic Impact :**
   * **Sponsorship and Revenue:** Analysing the financial aspects, including sponsorship deals and revenue generation.
   * **Prize Distribution:** Reviewing how prize money is distributed among winners.

By examining these aspects, stakeholders can gain a deeper understanding of what contributes to success in CrossFit and how to enhance training, event organization, and athlete support.

**Initial observations :**

1. The data set contains various files relating to each one
2. The data is split into CSV and Excel.
3. The data is uncleaned
4. Few columns are not of same datatype
5. This data is from the 2019 of overall ranking not including other heats or plays.

**Processing the data :**

1. Cleaned the datasets of blank rows using power query editor of excel
2. Changed the column headers to correct headers, as the headers were misinterpreted in the data set of athlete details
3. Re-arranged the columns to match men and women scores and merge the scores based on the competitor ID
4. Removed unusable columns from the data

**Conclusions Based on All Results:**

1. **Model Comparison**:
   * **Linear Regression**: The performance of Linear Regression yielded an MSE of 1747.85, RMSE of 41.81, and an R² of 0.079. The low R² indicates that this model isn't explaining much variance in the target variable and may not be the best choice for this dataset.
   * **Decision Tree**: The Decision Tree's **MSE was higher at 3342.58**, **RMSE of 57.82**, and a **negative R² (-0.76)**, showing poor performance. This model may be overfitting or not generalizing well.
   * **Random Forest**: The Random Forest model gave the best performance with an MSE of 0.82, RMSE of 0.91, and a near-perfect R² of 0.999. This indicates the model fits the data very well and explains almost all the variance in the target variable.
   * **Gradient Boosting**: The Gradient Boosting model performed better than some other models but still showed a lower R² (-0.015) and higher error metrics (MSE of 1926.80 and RMSE of 43.90).
   * **SVR**: The Support Vector Regressor (SVR) had an MSE of 1781.38 and RMSE of 42.21 with an R² of 0.061, but this model also did not perform as well as Random Forest.

**Conclusion**:

The **Random Forest** model performed the best by far, indicating that this ensemble method, which leverages multiple decision trees, can handle the complexity and variability of the data effectively.

1. **Feature Importance**:
   * Features with the highest importance included numerical columns such as 'age', 'weight', 'points', and certain categorical encodings.
   * Some features, such as 'affiliatename' and 'scoredisplay', had low importance and minimal influence on the target variable.
   * You could consider removing or modifying low-importance features to streamline the model and potentially improve performance.
2. **Error Analysis**:
   * The prediction errors for Random Forest are very small, indicating highly accurate predictions. Most errors are within a small range, meaning the model generalizes well to unseen data.
   * Examples: For one sample, the predicted value was 1.48 points higher than the actual value, while for another it was 1.41 points lower.
3. **Handling Missing Values**:
   * Missing values were present and handled using an imputation technique (median/most frequent values). This step ensures the models can be trained without errors and improves model reliability.
4. **Data Preprocessing and Categorical Encoding**:
   * Categorical variables such as 'affiliatename', 'division', and 'gender' were encoded using advanced transformers like OneHotEncoder and TargetEncoder.
   * This allowed these categorical features to contribute to model prediction, improving model accuracy.
5. **Most Successful Lane**:
   * Based on the analysis, the lane with the highest average points is identified as the most successful. This could be due to factors such as favorable conditions or more skilled competitors being assigned to that lane.
6. **Impact of Results and Next Steps**:
   * **Improve Feature Engineering**: Further exploration of feature interactions and engineered features could provide more insights and potentially improve model performance.
   * **Hyperparameter Tuning**: For Random Forest, tuning hyperparameters like ***n\_estimators, max\_depth, and min\_samples\_split*** might yield even better results.
   * **Balance the Dataset**: If there are imbalances in target values across lanes or other features, techniques like SMOTE or stratified sampling might help.
   * **Explore Additional Models**: While Random Forest performed well, you could experiment with boosting methods like XGBoost or CatBoost, which often outperform on more complex datasets.

References

CrossFit 2019 data set