Phir Ek Mauka Hack: Ind vs Pak Cricket Hackathon

Solution Framework

Rank 2nd Approach

Name of Author: - Aniket Bhausaheb Barphe

Private Leaderboard: - 2nd Rank

Hackathon Name: - Phir Ek Mauka Hack: Ind vs Pak Cricket Hackathon

Hackathon Platform: - AnalyticsVidhya(AV)

Hackathon Period: - 29-09-2023 09:00 AM to 14-10-2023 02:00 PM

AV-Username: - barphe

LinkedIn Profile: - https://www.linkedin.com/in/aniiketbarphe/

Find the full code implementation here: -

https://github.com/aniiketbarphe/Phir_Ek_Mauka_Hack-Ind_vs_Pak_Cricket_Hackathon_AnalyticsVidhya-Sep2023/blob/main/Phir_Ek_Mauka_Hack-Ind_vs_Pak_Cricket_Hackathon_AnalyticsVidhya_Oct23.ipynb

1) Problem Statement: -

The goal of hackathon is to create an advanced machine learning solution that leverages data science models and techniques. This solution aims to provide accurate predictions for individual player performances, specifically predicting the runs scored and wickets taken.

This prediction will be based on a comprehensive historical dataset that covers both player statistics and team performance. The ultimate objective is to forecast how each player will perform in the highly anticipated ICC World Cup 2023 match scheduled for October 14, 2023.

2) About Data: -

The dataset provided for this hackathon by the AV team includes comprehensive statistics for the 30 players selected to participate in the ICC World Cup 2023, representing both the Indian and Australian teams. This dataset encompasses detailed batting and bowling performance metrics for each One Day International (ODI) match played by the cricketers throughout their careers. In total, the dataset comprises 10 distinctive features and spans a collection of 2,021 meticulously recorded records.

In addition to the AV data, I have extracted comprehensive player information from the "ESPNCricInfo" website. This supplementary dataset encompasses 21 additional features for each player, providing valuable insights and enriching this analysis. You can find this open-source data integrated into the full code mentioned above.

3) Approach: -

3.1) Preprocessing the data: - The objective of this phase was to optimize the dataset for machine learning model utilization. Key steps involved in this process included:

- **Statistical Insights:** The initial step involved gaining a comprehensive understanding of the dataset through statistical analysis. This encompassed examining key aspects like data shape, identifying missing values, and assessing unique value counts.

- **Data Cleaning:**
  - ✓ Eliminated non-numeric characters, including asterisks (*), from the 'Run_Scored' feature.
  - ✓ Substituted 'DNB' and 'TDNB' with blank spaces to ensure uniformity and facilitate further analysis within the 'Run_Scored' feature.
  - ✓ Replaced hyphens "-" with "Zero" to standardize missing values.
  - ✓ Extracted information from the "opposition" column to derive "Opponent_Team_Name" and "Match_Venue".
  - ✓ Eliminated the redundant 'v' prefix in "Country_Name" within the "Opponent_Team_Name" column.

- **Data Integration:** Enriched the existing AnalyticsVidhya dataset by seamlessly incorporating an external dataset.

- **Feature Engineering:** Derived new features, such as "Catches per innings" and "Stumpings per Innings," to enhance the dataset's predictive power.
- **Imputation:** Addressed null values within the dataset to ensure completeness and data integrity.

3.2) Modeling: - Since the problem involves time series data, I chose to implement various strategies using the Seasonal Autoregressive Integrated Moving Average (SARIMA) model.

3.3) Strategies Explored: - Throughout the model-building process, a range of strategies were investigated and implemented to strengthen our predictive capabilities. The following key strategies were tested and applied,

- **Exploratory Analysis:** Conducted exploratory analysis to understand the dataset. Explored features and their correlations with the target variables (run_scored and wickets).
- **Single Model Approach:** Initially, built a single model to predict both run_scored and wickets using all available features.
- **Feature Selection:** Analyzed the feature correlations and dropped irrelevant features to improve model performance.
- **Single Model with Selected Features:** Re-tested a single model using the selected relevant features for predicting both run_scored and wickets.
- **Separate Models for Run and Wickets:** Built separate models with different parameters to predict "runs_scored" and "wickets" individually, using the selected relevant features.
- **Parameter Optimization:** Identified and fine-tuned three different sets of parameters for the models targeting both run_scored and wickets.
- **Ensemble Technique:** Employed ensemble techniques to predict "runs_scored" and "wickets" separately using the selected features. This ensemble approach improved prediction accuracy.
- **Final Submission:** Combined the outputs from the ensemble models into a single file following the format provided by AV. This file was marked as the final submission on the platform.

By following this structured approach, I successfully optimize my models and achieve more accurate predictions for both run_scored and wickets, ultimately improving my final submission for the competition.

For any additional questions or clarifications, please feel free to reach out to me via email at aniiketbarphe@yahoo.com or through my LinkedIn profile, as indicated above.