Project Report

Project: Predicting Indian Premier League (IPL) Match Winners

Dhiraj H. Gawhare
Roll Number: 22M0062
M. Tech (Aerodynamics)
IIT Bombay.

Project Title: Predicting Indian Premier League (IPL) Match Winners

Project Overview: The project aims to predict the winners of Indian Premier League (IPL) matches based on various features and historical data. The project involves data preprocessing, feature engineering, exploratory data analysis, and building predictive models using machine learning algorithms.

Key Steps:

1. **Data Loading and Exploration:**
    - The project starts by importing necessary libraries such as NumPy, Pandas, Matplotlib, Seaborn, and Plotly for data manipulation, visualization, and modeling.
    - Two datasets, "Trainmatches.csv" and "TrainDeliveries.csv," are loaded into Pandas DataFrames.
    - Initial exploration of the datasets is performed using basic functions like **.head()**, **.info()**, and **.describe()**.
2. **Data Visualization:**

- The number of matches won by each team is visualized using bar plots to give an overview of team performance.
- Histograms and bar plots are used to analyze features like seasons, toss winners, and match winners.
- A pie chart is used to illustrate the correlation between toss winners and match winners.

3. **Data Cleaning:**
    - Null values in the dataset are identified and handled using appropriate methods, including dropping rows with missing values.
4. **Feature Engineering:**
    - A feature table (FT) is created to store various features for each match identified by the match_id.
    - Features like teams (Team A, Team B), season, toss winner, Duckworth-Lewis method application, and cross-validation features are added to the feature table.
    - Batting and bowling averages for each team are calculated and added to the feature table.
    - A team performance DataFrame is created to calculate and store batting and bowling averages for both teams in each match.
    - Super over information is included in the feature table.
    - The player of the match feature is analyzed, but due to discrepancies, it's not included in the feature table.
5. **Preparation of Feature Table:**
    - A function **df_feature()** is created to generate the feature table based on provided datasets.
    - The function calculates team performance, super over results, and other features for each match.
6. **Modeling and Prediction:**
    - The feature table is divided into training and testing sets for building and evaluating models.
    - Three different machine learning models are applied for prediction:
        - Gaussian Naive Bayes
        - Decision Tree Regressor
        - Support Vector Machine (SVM)
    - Accuracy scores and confusion matrices are used to evaluate the performance of each model.

Results and Conclusion:

- The Gaussian Naive Bayes model achieved an accuracy of approximately X% in predicting IPL match winners.
- The Decision Tree Regressor model achieved an accuracy of approximately Y%.
- The Support Vector Machine (SVM) model's accuracy was approximately Z%.
- Confusion matrices provided insight into the models' performance in terms of true positives, true negatives, false positives, and false negatives.

Key Insights:

- Seasonal performance varies for teams, indicating that seasons are an important feature.
- Toss winning is correlated with match winning to some extent.
- Batting and bowling averages influence team performance significantly.

Challenges and Limitations:

- Handling missing data and discrepancies in features like player of the match.
- The dataset's scope is limited to historical matches and may not account for dynamic team changes or emerging strategies.
- The model's predictive power depends on the chosen features and the inherent complexity of cricket matches.

Recommendations for Improvement:

- Include additional features such as player statistics, player injuries, and team rankings.
- Experiment with more advanced machine learning algorithms and ensemble methods.
- Explore time series analysis to capture the temporal aspects of IPL matches.
- Collect more diverse and comprehensive data to improve model accuracy.