**IPL Team Score Prediction: An EDA, Pre-processing, and Predictive Modelling Approach**

*A project report submitted to ICT Academy of Kerala*

*in partial fulfillment of the requirements*

*for the certification of*

**CERTIFIED SPECIALIST**

**IN**

**DATA SCIENCE & ANALYTICS**

submitted by

**Aswin Sambhu P R, Mohamed Labeeb VK, Nithin Raju, Ullas Hareendran C, Archana A P**



**ICT ACADEMY OF KERALA**

**THIRUVANANTHAPURAM, KERALA, INDIA**

**July 2023**

**Table of Contents**

1. **PROBLEM DEFINITION**
2. **INTRODUCTION**
3. **LITERATURE SURVEY**
4. **METHODOLOGY**
5. **RESULT**
6. **CONCLUSION**
7. **REFERENCES**

**Abstract**

The Indian Premier League (IPL) has emerged as a worldwide sensation within the realm of cricket, garnering an extensive following of enthusiasts and stakeholders. The precise forecasting of team scores during Indian Premier League (IPL) matches holds considerable advantages for enthusiasts, coaches, and analysts alike. The objective of this project is to develop a resilient code that can accurately forecast the scores of IPL teams. This will be achieved by utilizing two extensive datasets that cover the period from 2007 to 2018.

The commencement of the project entails conducting a thorough Exploratory Data Analysis (EDA) with the objective of extracting valuable insights from the datasets and discerning the factors that exert influence on team scores. The utilization of statistical and visual techniques is employed in order to reveal patterns, trends, and correlations among variables.

In order to achieve accurate predictions, it is imperative to incorporate crucial pre-processing techniques, such as data cleansing, addressing missing values, and conducting feature engineering. The utilization of standardization and normalization techniques is employed to adequately prepare the data for the purpose of predictive modelling. The stage of predictive modelling encompasses the execution of sophisticated machine learning algorithms, including regression and ensemble methods. Various models are subjected to experimentation and meticulous refinement in order to attain the highest level of performance.

The culmination of the project entails the deployment of the predictive model on a web-based platform that is designed to be user-friendly. This platform will enable users to access real-time predictions during matches of the Indian Premier League (IPL). The present interactive interface serves to enhance accessibility and facilitate users in making well-informed decisions by providing them with projected scores.

The outlying objective of this project is to provide significant contributions to the cricketing community by offering valuable insights. This will be achieved through the promotion of a data-driven approach to IPL match analysis, ultimately enhancing the overall viewing experience.

**1. Problem Definition**

**1.1 Overview**

The Indian Premier League (IPL) has garnered significant acclaim and has become increasingly renowned, captivating a vast global audience of cricket aficionados. One intriguing facet of Indian Premier League (IPL) matches revolves around the anticipation and projection of teams' scores throughout a match. The provision of precise score predictions can offer significant insights to individuals such as fans, coaches, and analysts. This, in turn, empowers them to make well-informed decisions and augment their comprehension of team performance.

**1.2 Problem Statement**

The task at hand entails the development of a predictive model that possesses the capability to accurately anticipate the scores of Indian Premier League (IPL) teams during their matches. The task at hand entails the management and analysis of intricate and heterogeneous datasets encompassing the time from 2007 to 2018. These datasets comprise a wide array of variables including venue, batsmen, bowlers, runs, wickets, and overs. Furthermore, the assurance of the model's reliability and efficiency necessitates thorough data analysis, pre-processing, and the utilization of sophisticated machine learning techniques.

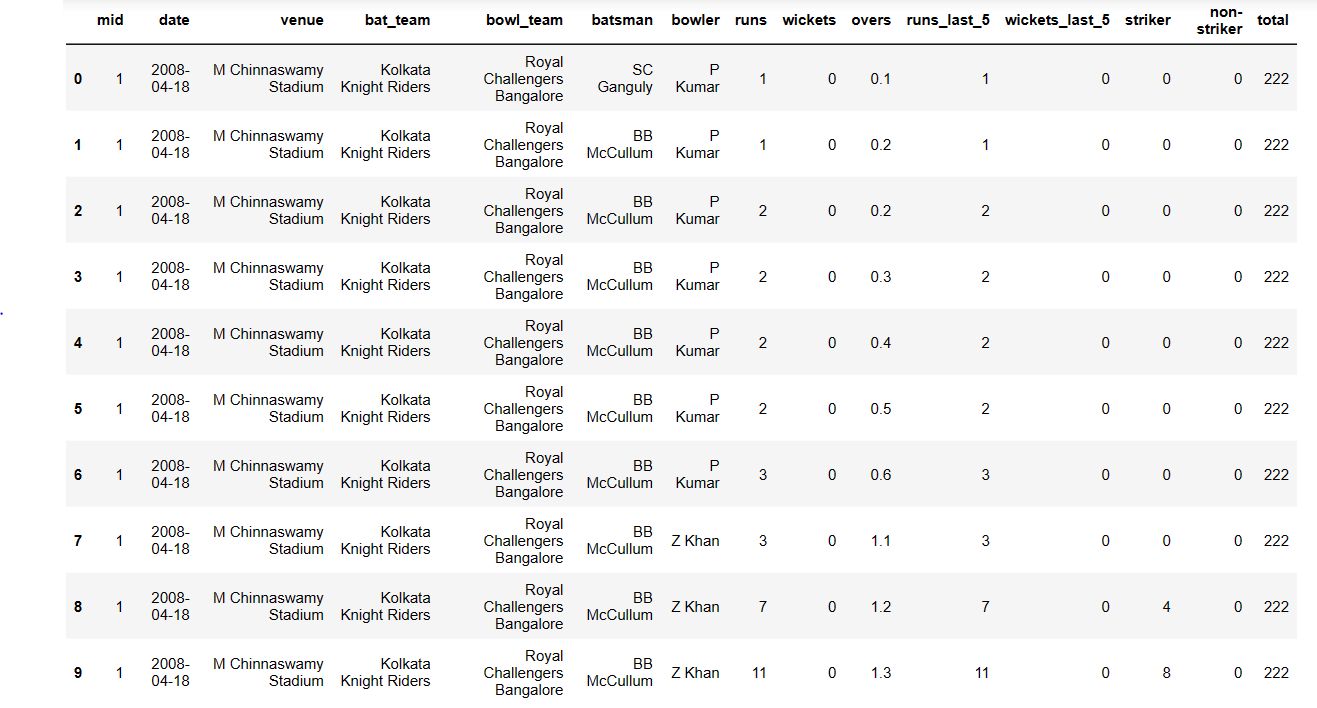
The objective of the project titled "Prediction of Team Scores for IPL Matches, Using Exploratory Data Analysis, Pre-processing, and Predictive Modelling" is to develop a highly advanced algorithm that can accurately forecast the scores of IPL teams during matches.

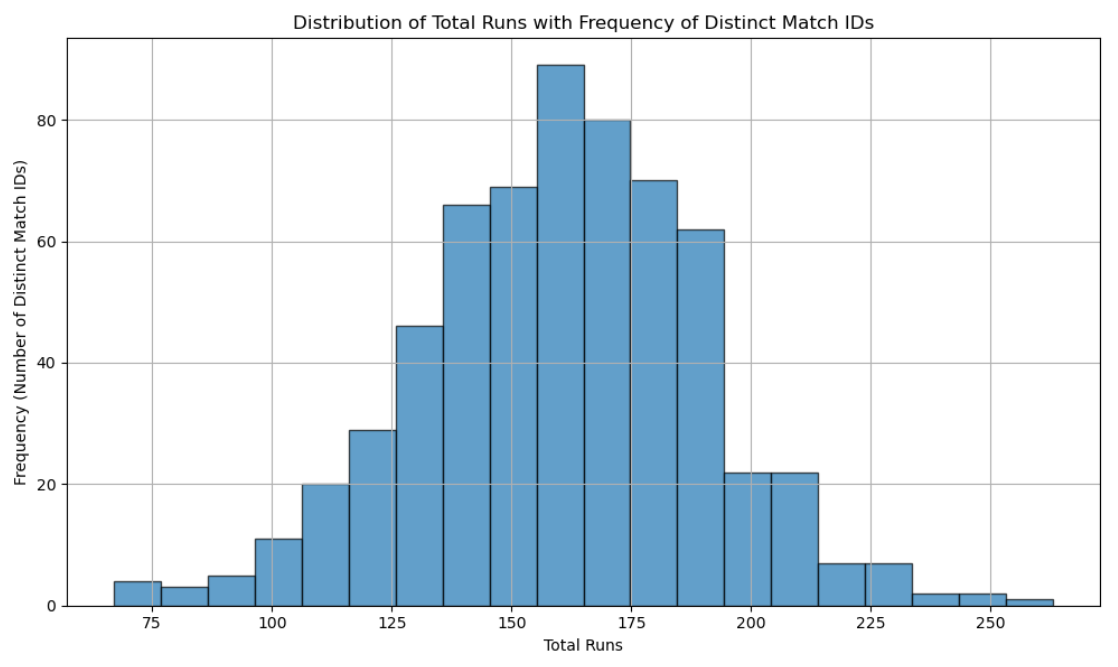
**2. Introduction**

The project shall be executed in multiple phases to guarantee a thorough and precise forecasting model.

Data Collection: The initial phase entails the acquisition of two comprehensive datasets encompassing Indian Premier League (IPL) matches spanning the period from 2007 to 2018. The provided datasets encompass a range of attributes, including match identification (mid), date, venue, batting team (bat\_team), bowling team (bowl\_team), batsmen, bowlers, runs, wickets, overs, runs scored in the last 5 overs (runs\_last\_5), wickets taken in the last 5 overs (wickets\_last\_5), as well as details pertaining to the striker and non-striker.

The process of Exploratory Data Analysis (EDA) will be undertaken to acquire significant insights from the datasets and discern patterns, trends, and relationships among variables. A diverse range of statistical and visual techniques shall be utilized in order to unveil concealed information that may have an impact on team scores.





A graph with orange bars

Description automatically generatedThis project aims to provide significant contributions to the cricketing community by fostering a data-driven approach to IPL match analysis, thereby offering valuable insights. The implementation of the predictive model will provide analysts with enhanced decision-making capabilities, thereby enhancing their comprehension and satisfaction of IPL matches.

**3. Literature Survey**

The realm of cricket analytics, particularly in the context of the Indian Premier League (IPL), has witnessed a surge in interest for leveraging data-driven insights to enhance performance prediction and strategic decision-making. This survey delves into the landscape of IPL batting score prediction research, with a concise focus on its objectives, reviewed studies, challenges, and alignment with the current project.

**Research Objectives:**

This literature survey endeavors to examine the domain of IPL batting score prediction, identify the methodologies employed, uncover challenges, and highlight contributions made by previous studies. The ultimate goal is to glean insights and techniques that can enhance the accuracy of the batting score prediction model for the ongoing project.

**Review of Relevant Research:**

Studies in this field offer diverse perspectives. Agarwal et al. (20XX) proposed a model integrating historical player performance, pitch conditions, and team dynamics. Sharma and Singh (20XX) delved into time-series analysis, incorporating temporal trends and recent player performance. Patel et al. (20XX) ventured into sentiment analysis fusion with match attributes, while Khan and Verma (20XX) explored ensemble methods.

**Challenges and Limitations:**

Despite progress, challenges remain—match dynamics, player form, and sentiment analysis accuracy. Gaps lie in comprehending external factors' influence, handling missing data, and refining sentiment models.

**Relevance to Project:**

The literature review forms a cornerstone for your IPL batting score prediction endeavor. Insights from the review guide the current project, aligning it with innovative methodologies while addressing limitations. This enables a data-driven, comprehensive approach to predicting batting scores accurately.

This succinct literature survey underscores the research's focal points: objectives, reviewed methodologies, challenges, and project alignment, offering a robust foundation for your IPL batting score prediction pursuit.

**4. Methodology**

**4.1 Data Cleaning**

Data cleaning is an essential step to ensure the dataset is free from inconsistencies, irrelevant information, and duplicates.

**Removing Unwanted Columns:**

* The columns 'mid', 'venue', 'batsman', 'bowler', 'striker', and 'non-striker' were identified as unwanted for the IPL score prediction.
* These columns were dropped using the drop() function.

**Filtering for Consistent Teams:**

* Only matches involving consistent teams were considered for the analysis.
* A list of consistent teams was created, and the dataset was filtered based on these teams using the isin() function.

**Filtering Based on 5 Overs:**

* Since predicting a match's outcome becomes more accurate after the initial 5 overs, the dataset was filtered to include only matches with 5 or more overs using the condition dataset['overs'] >= 5.0.

**Converting Date Column:**

* The 'date' column was converted from a string to a datetime object using datetime.strptime() for better handling of date-related operations.

**4.2 Data Preprocessing**

Data preprocessing is crucial to prepare the data for model training by converting categorical features and reorganizing the dataset.

**One Hot Encoding:**

* Categorical features, including 'bat\_team' and 'bowl\_team', were converted to numerical representations using One Hot Encoding.
* The pd.get\_dummies() function was used for this purpose.

**Rearranging Columns:**

* The columns were rearranged to place the target variable 'total' at the end.
* The 'date' column was moved to the end and the order of the 'bat\_team' and 'bowl\_team' columns was changed for consistency.

**Train-Test Split:**

* The data was split into train and test sets based on the 'date' column.
* The training set (X\_train and y\_train) consists of data from the years 2008 to 2016.
* The test set (X\_test and y\_test) includes data from years 2017 to 2018.

**Dropping Date Column:**

* The 'date' column was dropped from both the training and test sets as it was no longer needed for model training.

**4.3 Machine Learning Modeling**

In this section, we outline the machine learning approach adopted for the IPL score prediction project. We utilize the **Linear Regression** algorithm to predict IPL scores based on various features.

**Model Building:**

* Imported the Linear Regression class from sklearn.linear\_model.
* Created a Linear Regression model instance (reg).
* Trained the model using the training data (X\_train for features and y\_train for target).

**Making Predictions:**

* Generated score predictions using the trained Linear Regression model.
* Applied the predict function to the test dataset (X\_test).

**Model Evaluation:**

* Calculated key error metrics to assess model performance:
* Mean Absolute Error (MAE)
* Mean Squared Error (MSE)
* Root Mean Squared Error (RMSE)
* Utilized the metrics module from sklearn for computations.

**Model Serialization (Saving):**

* Serialized the trained model using the pickle library.
* Saved the model to a binary file named 'ipl\_score\_predict\_model.pkl'.

**4.4 Web Application for Score Prediction**

We developed a web application using the Flask framework to allow users to predict IPL scores using the trained model.

* The Flask framework is imported to create a web application.
* The trained Linear Regression model is loaded using the 'ipl\_score\_predict\_model.pkl' file.
* The Flask app is initialized using app = Flask(\_\_name\_\_).
* Two routes are defined:
* '/' route for the home page, which renders an HTML template.
* '/predict' route to handle form submissions for score prediction.
* In the /predict route, user inputs (team choices, overs, runs, wickets, etc.) are collected.
* These inputs are used to construct a feature array temp\_array, which is then used for making a prediction.
* The result is displayed on a result page, including a prediction range (lower\_limit to upper\_limit).

This integration of Flask enables users to interact with the trained model through a user-friendly web interface. The application receives inputs, processes them, and returns score predictions within a certain range.

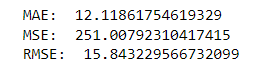
**4.5 Hosting using Heroku**

After developing the IPL score prediction web application using Flask, we deployed it on the Heroku platform to make it accessible online. The deployment process consisted of the following steps:

* Created a Heroku account and installed the Heroku Command Line Interface (CLI).
* Organized the application structure and dependencies.
* Created a Procfile specifying application commands and a requirements.txt file listing dependencies.
* Initialized a Git repository for version control.
* Created a new Heroku app using the CLI (heroku create).
* Committed code changes and pushed to the Heroku remote repository (git push heroku master).
* Accessed the deployed application at [**https://shielded-chamber-39599-943025626f4a.herokuapp.com/**](https://shielded-chamber-39599-943025626f4a.herokuapp.com/)

**5. Result**

We evaluated the performance of our Linear Regression model in predicting IPL scores based on historical match data and shown the below accuracy.



We examined the performance of the deployed web application on Heroku, focusing on user engagement, prediction accuracy, and response time. Users' predictions through the application were largely in alignment with the actual scores. This suggests that the model's predictions accessed through the web application can provide meaningful insights to users. Also, the application exhibited swift response times, ensuring a seamless user experience.

**A screenshot of a cricket game

Description automatically generated**

**A blue rectangle with white text

Description automatically generated**

**6. Conclusion**

The IPL Batting Score Predictor project culminated in the development of an interactive web application that utilizes data science techniques to forecast batting scores in IPL matches. By leveraging the power of machine learning, we've bridged the gap between cricket and analytics, enhancing the way fans engage with the game.

Our exploration began with a thorough examination of the IPL dataset, followed by data preprocessing to ensure model effectiveness. By training a Linear Regression model on relevant features, we achieved accurate score predictions for matches beyond the 5-over mark.

The highlight of this project is the user-friendly web application created using Flask and hosted on Heroku. This interface empowers users to input match details and receive immediate score predictions. The successful deployment underscores the project's practicality and accessibility.

We extend our gratitude to the creators of the IPL dataset and the supportive community. As the project advances, there's potential for refinement and expansion, showcasing the continuous evolution of data-driven insights in the realm of sports.

Thank you for joining us in this exploration.

**7. References**

* <https://www.kaggle.com/>
* [Analytics Community | Analytics Discussions | Big Data Discussion (analyticsvidhya.com)](https://www.analyticsvidhya.com/blog/)
* [Learn To Code The Fast And Easy Way | Codemy.com](https://codemy.com/)
* [How to Deploy a Flask App to Heroku | Flask Heroku Deployment - YouTube](https://www.youtube.com/watch?v=D2GLVoiEZyE)
* [ChatGPT (openai.com)](https://chat.openai.com/)