Cummins DATATHON

Team Name : UNDERDOGS_ON_FIRE
Problem Statement Number : 2

Problem Statement & Objectives

Problem Statement : IPL 2024 Score and Winner Prediction

Objective : Develop a predictive model to forecast the score and predict the winner of IPL 2024 matches based on historical game telemetry data.

Tasks :

1. Data Preprocessing: Clean and prepare the game telemetry data for analysis. This may involve handling missing values, identifying outliers, and feature engineering to create new relevant features from the raw data.
2. Exploratory Data analysis for showing different trends on IPL seasons with respect to players, venues
3. Try with different model approaches and evaluate different metrics and use the best model for prediction.
4. Prediction of winner of any match on the basis of certain inputs which will be shared on the event date.
5. Prediction of score of any match on the basis of certain inputs which will be shared on the event date.

Proposed Approach

Data Preprocessing

- Cleanse the collected data by handling missing values, outliers, and inconsistencies to ensure data quality and reliability.
- Perform data validation and verification to ensure accuracy and consistency across datasets.
- Handling all the unnecessary values and insufficient values

Exploratory Data Analysis (EDA) and Feature Engineering

- Perform exploratory data analysis to gain insights into the underlying patterns, trends, and relationships within the dataset.
- Visualize key variables and distributions to understand the characteristics of the data.
- Identify potential correlations and dependencies between features and the target variables (match score and winner) to inform model development.
- Conduct feature engineering to extract relevant features from the raw data that are likely to influence the match score and outcome.
- Create new features or transform existing ones to capture meaningful information, such as player performance metrics, team statistics, venue characteristics, etc.
- Utilize domain knowledge and insights from exploratory data analysis to guide feature selection and transformation

Proposed Approach

Model Development for Score Prediction

- Build machine learning models to predict the score of IPL 2024 matches based on historical game telemetry data.
- Experiment with various regression algorithms such as linear regression, decision trees, random forests, gradient boosting, etc.
- Train the models using appropriate training and validation techniques, considering factors such as overfitting and model complexity.
- Evaluate model performance using relevant evaluation metrics such as mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), etc.
- Fine-tune model hyperparameters to optimize performance and generalization capability.

Model Development for Winner Prediction

- Develop machine learning models to predict the winner of IPL 2024 matches based on historical game telemetry data.
- Frame winner prediction as a classification problem and experiment with classification algorithms such as logistic regression, decision trees, support vector machines (SVM), etc.
- Incorporate features related to team performance, player statistics, venue characteristics, etc., to enhance prediction accuracy.
- Train the models using appropriate training and validation techniques, ensuring robustness and reliability.
- Evaluate model performance using classification evaluation metrics

Flow of the solution

Data Preprocessing

- Extract batting and bowling data from cards (runs, balls, fours, sixes, etc.).
- Compute batting average rate (batting runs / balls faced) and bowling average (bowling runs / wickets) for each player.
- Calculate team batting and bowling average rates by averaging individual player rates.
- Derive other metrics like team bowling economy rate (average economy rate of bowlers).
- Analyze trends in team performance across seasons using these metrics.
- Merged the main data of summary and newly found batting bowling metrices for all the teams
- Removed all the NaN values , anomalies , redundant values , unnecessary columns from the season summary and summary details datasets.
- Analyzed the 2022 and 2023 dataset patterns and merged the dataset for further exploration

batting_strike_rate	economy_rate	bowling_strike_rate	bowling_average

	season	id	name	short_name	description	home_team	away_team	toss_won	decision	1st_inning_score
0	2023	1359475	Gujarat Titans v Chennai Super Kings	GT v CSK	1st Match (N), Indian Premier League at Ahmeda...	GT	CSK	GT	BOWL FIRST	178/7

away_boundaries	away_overs	away_runs	away_wickets	home_boundaries	home_overs	home_runs	home_wickets
0.20	0.04	0.20	0.01	0.17	0.07	0.19	0.03
0.20	0.02	0.20	0.00	0.17	0.06	0.19	0.03
-0.07	0.03	-0.03	0.01	-0.11	0.06	-0.06	0.06

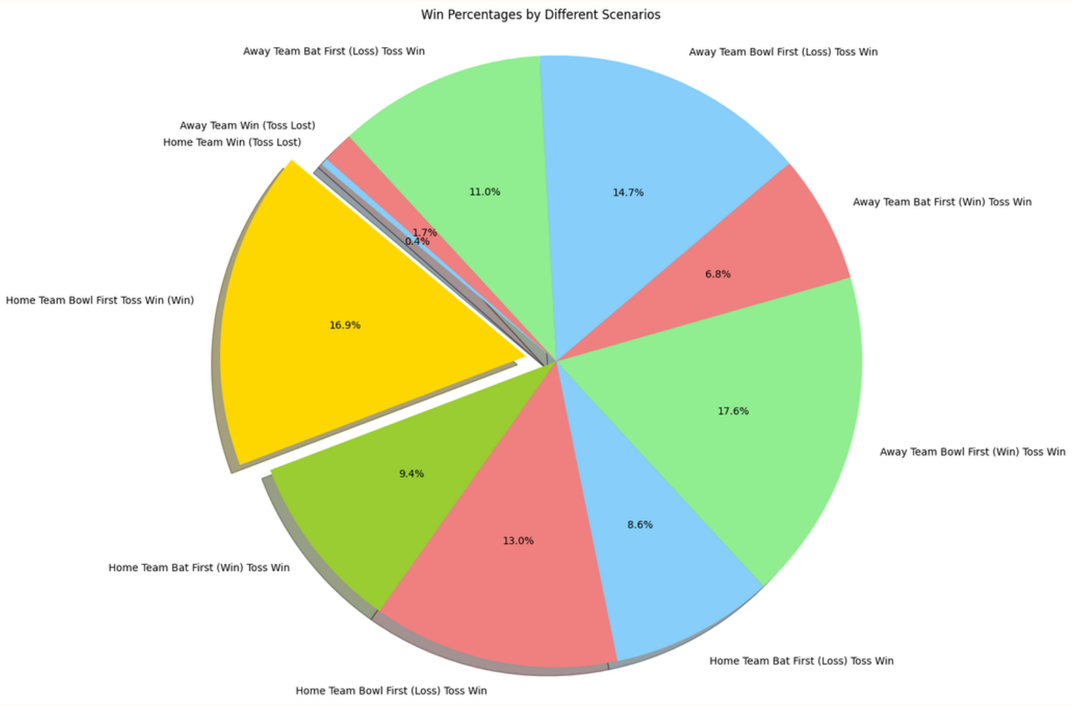
season	id	venue_id	home_overs	home_runs	home_wickets	home_boundaries	away_overs	away_runs	away_wickets	away_boundaries
2023	[illegible]	57851	19	182	5	23	20	178	7	23
2023	[illegible]	57991	20	191	5	26	16	146	7	20
2023	[illegible]	1070094	20	193	6	21	20	143	9	17
2023	[illegible]	58142	20	131	8	13	20	203	5	29
2023	[illegible]	57897	16	172	2	24	20	171	7	22
2023	[illegible]	58008	20	217	7	28	20	205	7	22

	season	venue_name	home_team	away_team	first_innings_average

Flow of the solution

Exploratory Data Analysis Feature Engineering

- We have tried to analyze all the datasets given to identify the best possible features form the dataset
- we are tried to identify multiple possible relations and have produced a lot of graphical data along with relations for various parameters and there correlations.
- We have successfully completed EDA and have produced satisfactory results for the same.

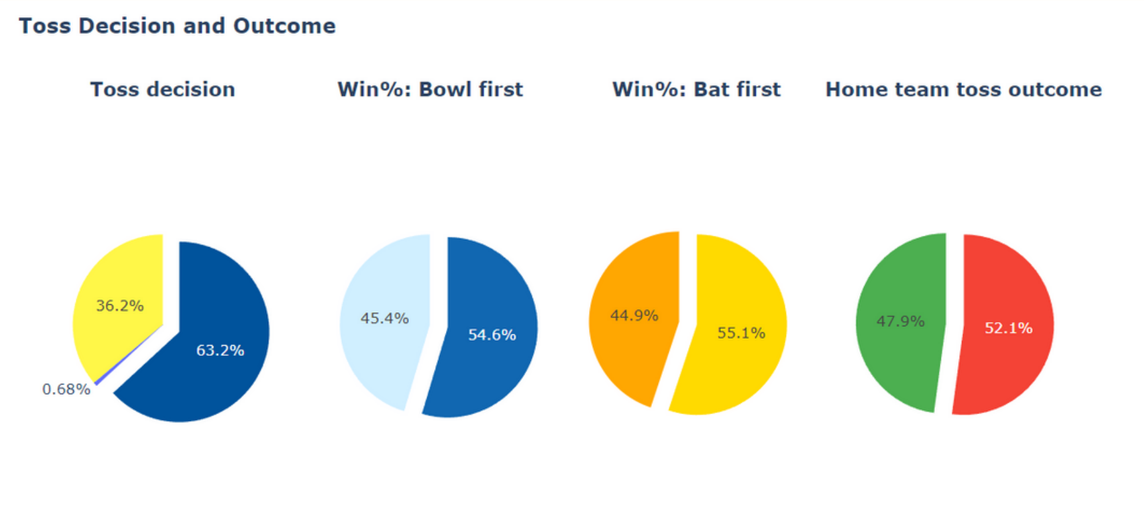


Here we are trying to understand the various possibilites of home team and away team winnning and loosing the game based on various factors like toss and venues along with toss decision. This analysis helped us to train our win prediction model and guided us for feature engineering

home_team	away_team	toss_won	decision	1st_inning_score	2nd_inning_score	winner

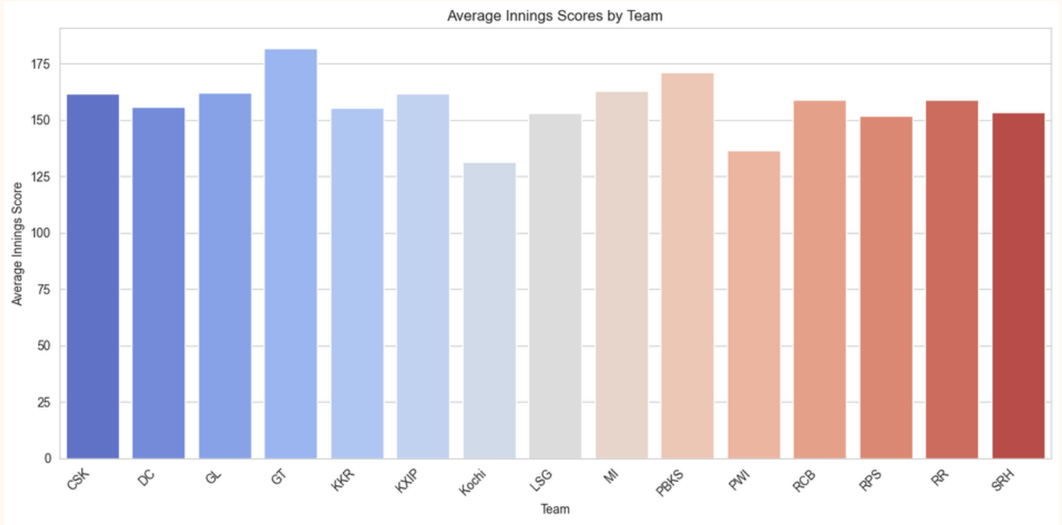
Flow of the solution

Exploratory Data Analysis Feature Engineering



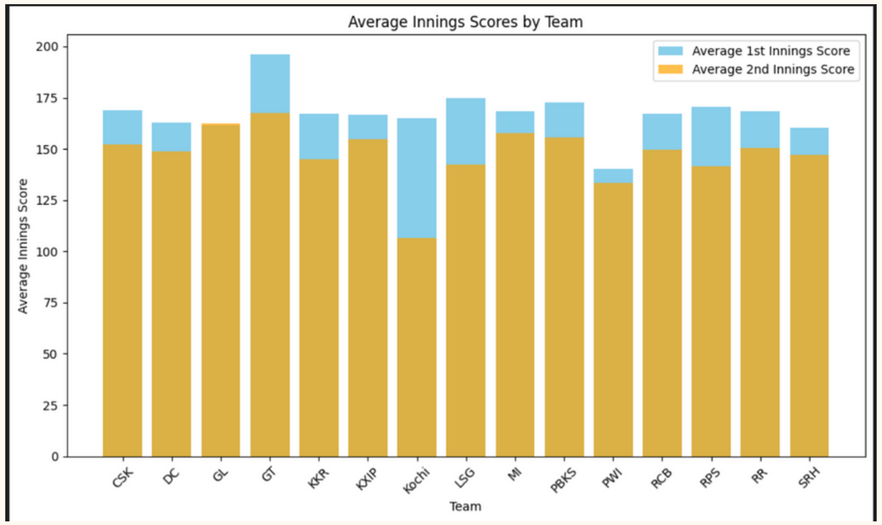
Here we have tried to identify the percentage of winning losing , choosing bat or ball after toss win and all the other parameters for the evaluation of the model

Here we have evaluated all the average innings score done by each team and tried to identify the importance of average score in the total score prediction



Flow of the solution

Exploratory Data Analysis Feature Engineering



We have tried to take average in both innings for all the teams and tried to identify the featutres

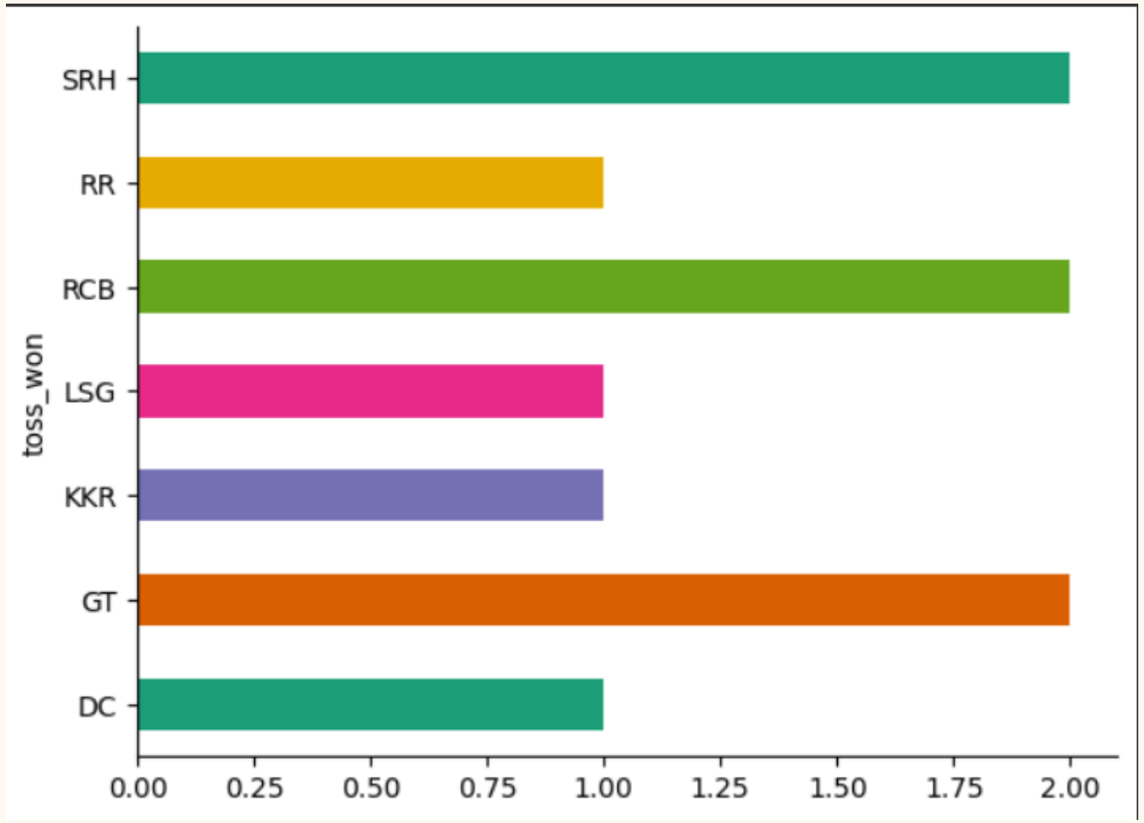
	team	venue_name	total_wins
0	CSK	Arun Jaitley Stadium, Delhi	8
1	CSK	Brabourne Stadium, Mumbai	1
2	CSK	Buffalo Park, East London	1
3	CSK	Diamond Oval, Kimberley	1
4	CSK	Dr DY Patil Sports Academy, Mumbai	3
...	...	...	...
261	SRH	Sheikh Zayed Stadium, Abu Dhabi	4
262	SRH	SuperSport Park, Centurion	1
263	SRH	The Wanderers Stadium, Johannesburg	2
264	SRH	Vidarbha Cricket Association Stadium, Jamtha, ...	2
265	SRH	Wankhede Stadium, Mumbai	4

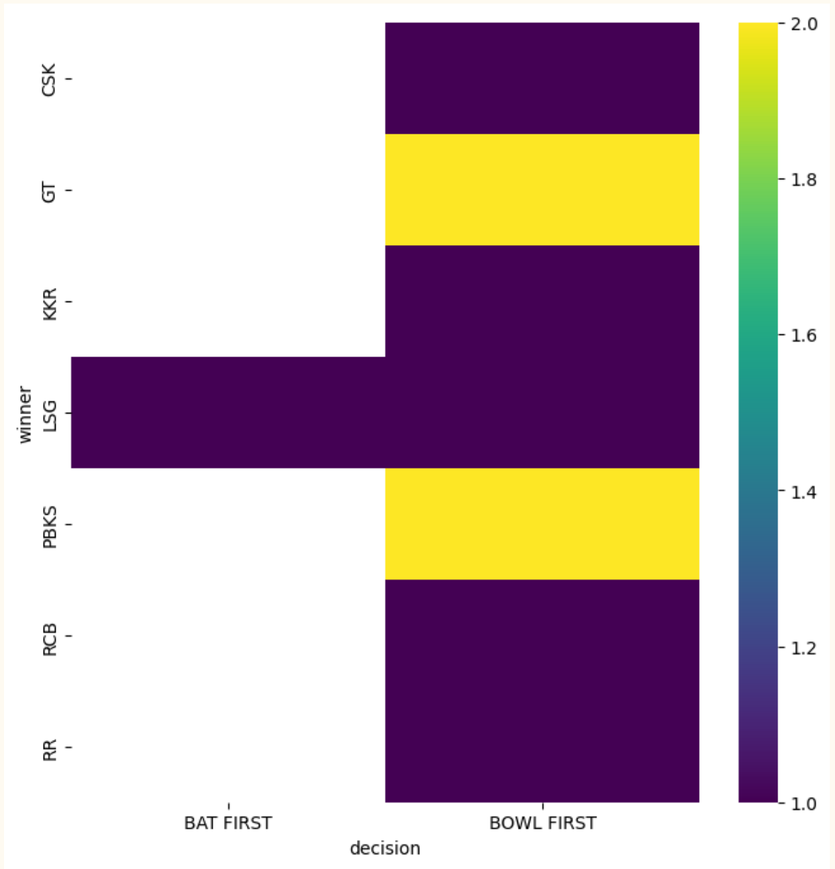
266 rows × 3 columns

Here we are trying the find the total number of wins for the each team on each venue which will help us create a better model feature engineering

Flow of the solution

Exploratory Data Analysis Feature Engineering





Flow of the solution

Model Development for Score Prediction

- We have used LazyPredict Library to predict the match score values using Regression Models.
- We have received lowest RMSE values for multiple models considering the modified parameters.
- Considering boundaries for the score prediction we get Gradient Boosting Regressor as the best one.
- Without considering boundaries for the score prediction we get Elastic Net CV Regressor as the best one.

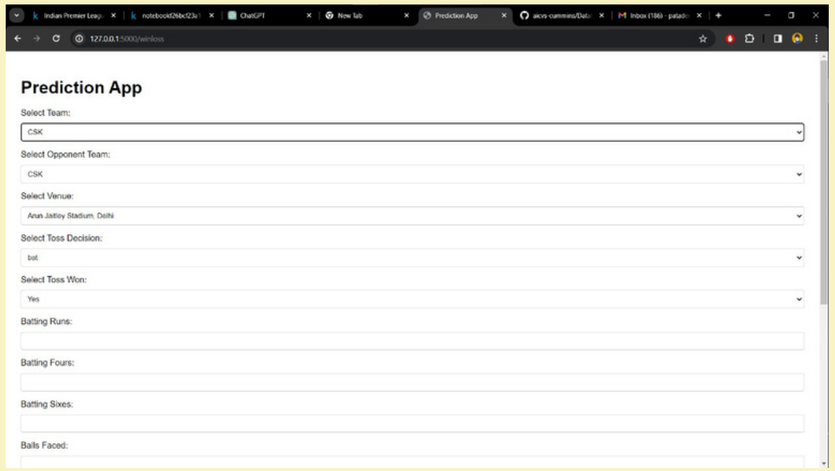
Model	Adjusted R-Squared	R-Squared	RMSE	Time Taken
ElasticNetCV	0.29	0.33	22.27	0.04
RidgeCV	0.29	0.33	22.28	0.01
BayesianRidge	0.29	0.33	22.29	0.01
LassoCV	0.29	0.33	22.31	0.03
LassoLarsCV	0.29	0.33	22.31	0.02
LarsCV	0.29	0.33	22.31	0.01
LassoLarsIC	0.29	0.33	22.32	0.01
Ridge	0.29	0.33	22.32	0.01

Model	Adjusted R-Squared	R-Squared	RMSE	Time Taken
GradientBoostingRegressor	0.91	0.92	9.38	0.10
Lasso	0.90	0.91	9.82	0.01
LassoLars	0.90	0.91	9.82	0.01
ElasticNetCV	0.90	0.91	9.87	0.04
LarsCV	0.90	0.91	9.87	0.01
LassoLarsCV	0.90	0.91	9.87	0.01
LassoCV	0.90	0.91	9.88	0.04
RANSACRegressor	0.90	0.91	9.89	0.07
LassoLarsIC	0.90	0.91	9.90	0.02

Flow of the solution

Model Development for Winner Prediction

- We have used LazyPredict Library to predict the compare multiple regression models to predict the confidence score.
- The best algorithm to predict the confidence score is Extra Tree Regressor.
- Extra Trees Regressor is well-suited for confidence prediction in sports due to its ability to handle noisy data, feature randomness, and perform well with limited tuning, making it robust and efficient for predicting uncertain outcomes accurately.
- We have utilised Flask API to create an UI for the winner prediction.




```
1  test
✓  0.0s
```

Model	Adjusted R-Squared	R-Squared	RMSE	Time Taken
LGBMRegressor	0.82	0.83	0.21	0.08
HistGradientBoostingRegressor	0.81	0.81	0.22	0.31
ExtraTreesRegressor	0.81	0.81	0.22	0.20
RandomForestRegressor	0.80	0.81	0.22	0.38
BaggingRegressor	0.80	0.80	0.22	0.05
MLPRegressor	0.79	0.80	0.22	0.79
XGBRegressor	0.77	0.77	0.24	0.29
GradientBoostingRegressor	0.76	0.77	0.24	0.24

Tech Stack

01.

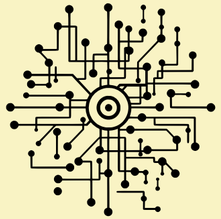
Python

02.

Flask

03.

ML Algorithnms



Conclusion & Future Scope

- **Integration of Real-time Data:** Incorporate real-time match data during live matches to continuously update and refine the predictive models. This can enhance the accuracy and reliability of predictions by considering the latest match conditions and player performances.

- **Player-specific Analysis:** Develop models to predict individual player performances based on historical data and match conditions. This can help teams make informed decisions regarding player selection, batting order, bowling strategies, etc.

- **Dynamic Model Updating:** Implement mechanisms to automatically update predictive models with new data and insights. This ensures that the models remain relevant and effective as the IPL season progresses and new patterns emerge.

- **Sentiment Analysis:** Integrate sentiment analysis of social media data to gauge public opinions and sentiments related to IPL matches. This additional input can provide valuable insights into fan expectations, team dynamics, and potential match outcomes.

- Incorporation of External Factors: Expand the scope of predictive models to incorporate external factors such as weather conditions, pitch characteristics, team dynamics, player injuries, etc. This holistic approach can lead to more comprehensive and accurate predictions.