EECE 5644 MACHINE LEARNING AND PATTERN RECOGNITION

IPL SCORE PREDICTION

TEAM

1. CHITHARANJAN GANESH KUMAR
2. NIKHIL ANIL PRAKASH
3. TRISHAL SAI SRINIVAS VENGETELA

EECE 5644 MACHINE LEARNING AND PATTERN RECOGNITION

INTRODUCTION : IPL PREDICTION MODEL

1. What is IPL?
2. Overview of the Project
3. Purpose of the Model
4. Technologies used: Python, ML Libraries
5. Data Collection : GitHub, Kaggle

Objective and Scope of the Project

The project includes using the model to predict future match outcomes and player performances, analyzing and interpreting the results to provide actionable insights.

Challenges such as venue variability, weather conditions, and player form are addressed. Future enhancements involve incorporating real-time data for live predictions and exploring advanced techniques like deep learning to improve accuracy.

Objective

The primary objective of this project is to develop a predictive model for IPL cricket matches. This model aims to forecast match outcomes and player performances by analyzing historical data. The goal is to provide accurate predictions that can aid teams in strategic decision-making, enhance fan engagement, and support commercial applications such as betting markets and fantasy leagues.

Scope

The scope of the project includes several key components aimed at developing a reliable predictive model for IPL cricket matches. First, data collection and preprocessing involve gathering historical IPL match data, including team performances, player statistics, and match conditions, and then cleaning and transforming this data for analysis. Next, feature engineering focuses on identifying and extracting key variables that influence match outcomes and player performances.

WHAT IS IPL?

Overview:
- Popularity
- Format
- Economic Impact

Relevance to Machine Learning:
- Data-Rich Environment
- Predictive Analysis
- Applications

Examples:
- Score Predictor
- Match Winner

Benefits:
- Strategic Decision Making
- Enhanced Fan Experience
- Competitive Edge

Challenges and Considerations:
- Data Quality
- Model Selection
- Real-time prediction
- Overfitting

DATA EXPLORATION

Data Source

- Kaggle Datasets- DATA from Season 2008 to 2017 (around 500 matches data i.e 76000 samples)
- CricBuzz, ESPN, etc.

Data Types

- Match Data- scores, Wickets, Overs
- Player Data- Runs, Strike rate
- Contextual Data- Weather, Toss

Data Collection Methodology

- Manual Collection
- API Integration

DATA CLEANING AND TRANSFORMATION

Data Cleaning

- Handling Missing Values
- Removing Duplicates

Data Integration

- Combining Datasets
- Storage

Data Quality Assurance

- Validation
- Visualization

PRE-PROCESSING

One Hot Encoding

Column Rearrangement

Data Splitting (till 2016 for training and beyond 2017 for testing)

Data Column Removal

Verification

Heat Map (Key Takeaway)

1. Runs and Overs: Strong positive correlation (0.88) indicating that more overs lead to more runs scored.

2. Runs in Last 5 Overs and Total Runs: Positive correlation (0.59) suggesting that good performance in the last 5 overs significantly increases the total runs.

3. Wickets and Total Runs: Negative correlation (-0.46) indicating that more wickets result in a lower total score.

4. Overs and Wickets: Moderate positive correlation (0.64), showing that as the number of overs increases, so do the wickets.

5. Wickets in Last 5 Overs and Total Runs: Negative correlation (-0.3), indicating that losing more wickets in the last 5 overs tends to decrease the total runs.

MODEL SELECTION

Understanding Problem Statement
- Type of Problem
- Key Factors

Commonly used models
- Liner Regression
- Ridge Regression
- Decision Tree
- XG Boost
- ADA Boost
- Support Vector Machining

Criteria for Model Selection
- Performance
- Interpretability
- Computational Efficiency
- Handling of Non-linear Relationships
- Feature Importance

Model Tuning and Optimization
- Test-Train Split
- Techniques

Model Evaluation
- Metrics
- Validation

REGRESSION TECHNIQUES

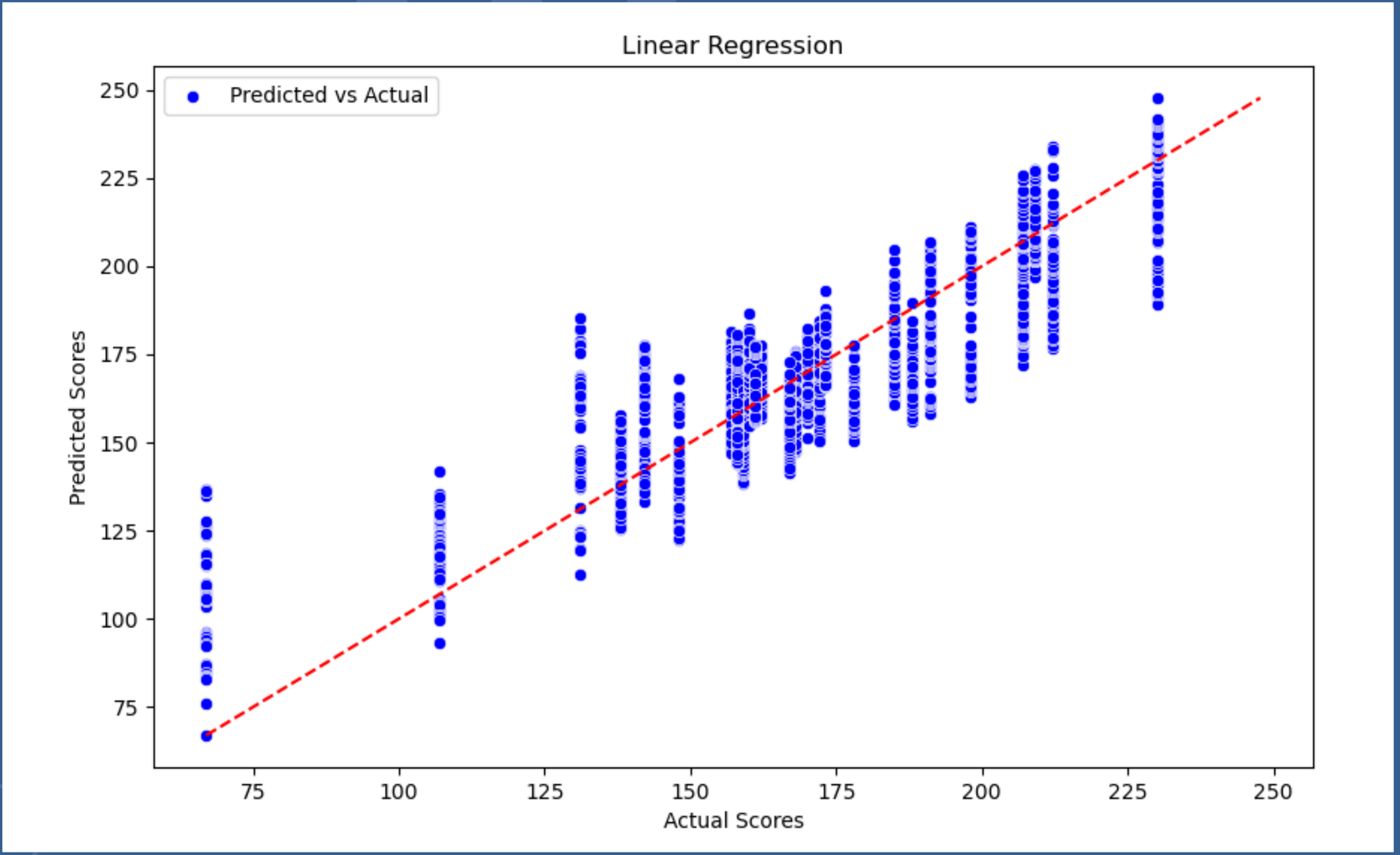
Linear Regression

- Represented by $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \mu$
- Where, y is predicted output, x1 .. Xn are the features, β_0 is the intercept, μ is the error term and β_j represents impact of each independent variable.

Ridge Regression

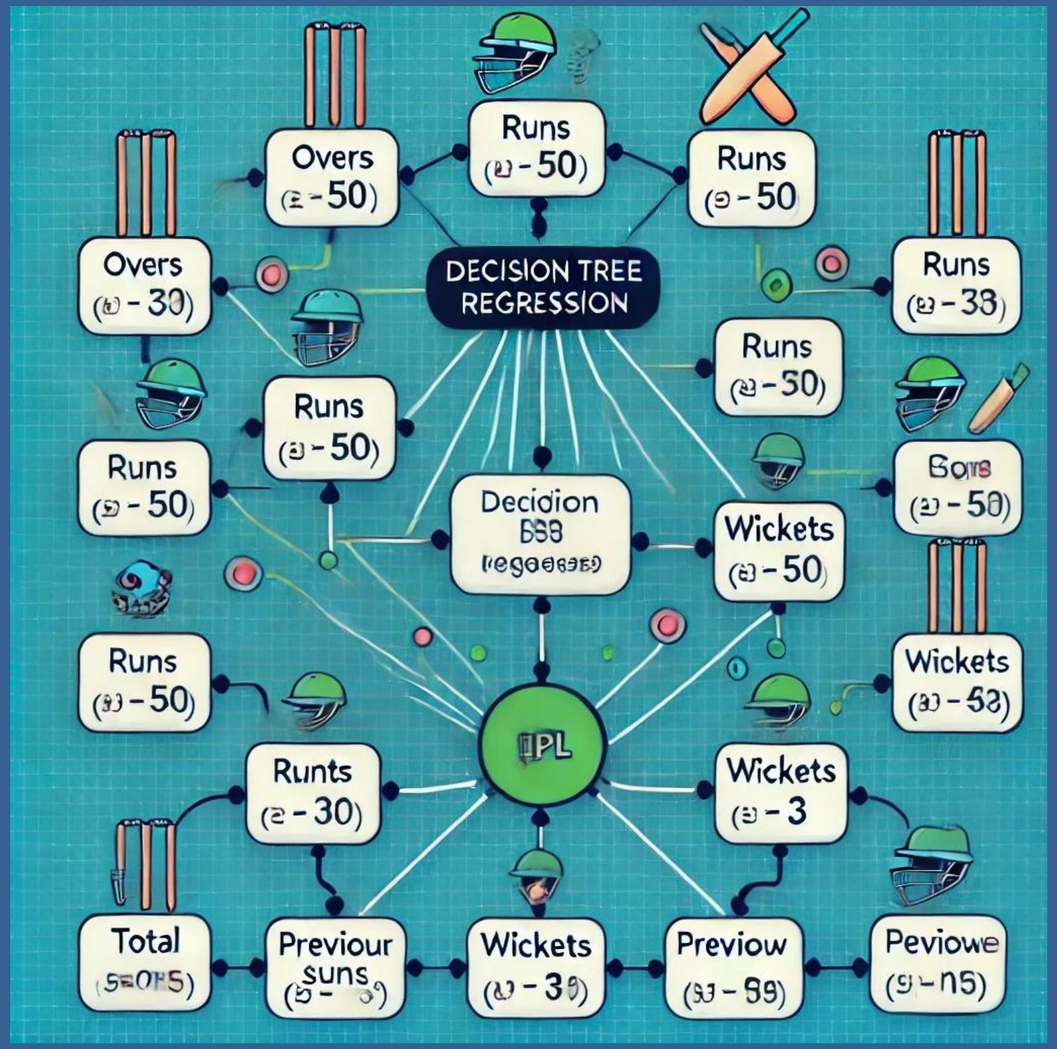
- Also, an extension of Linear Regression by the addition of a regularization term
- Cost function is represented by
  - $\Sigma^m_{i=1}(y_i - y'_i)^2 + \alpha\Sigma^n_{j=1}\beta_j$
  - where β_j represents impact of each independent variable and α is the regularization parameter that controls amount of shrinkage applied to the coefficients

EECE 5644 MACHINE LEARNING AND PATTERN RECOGNITION



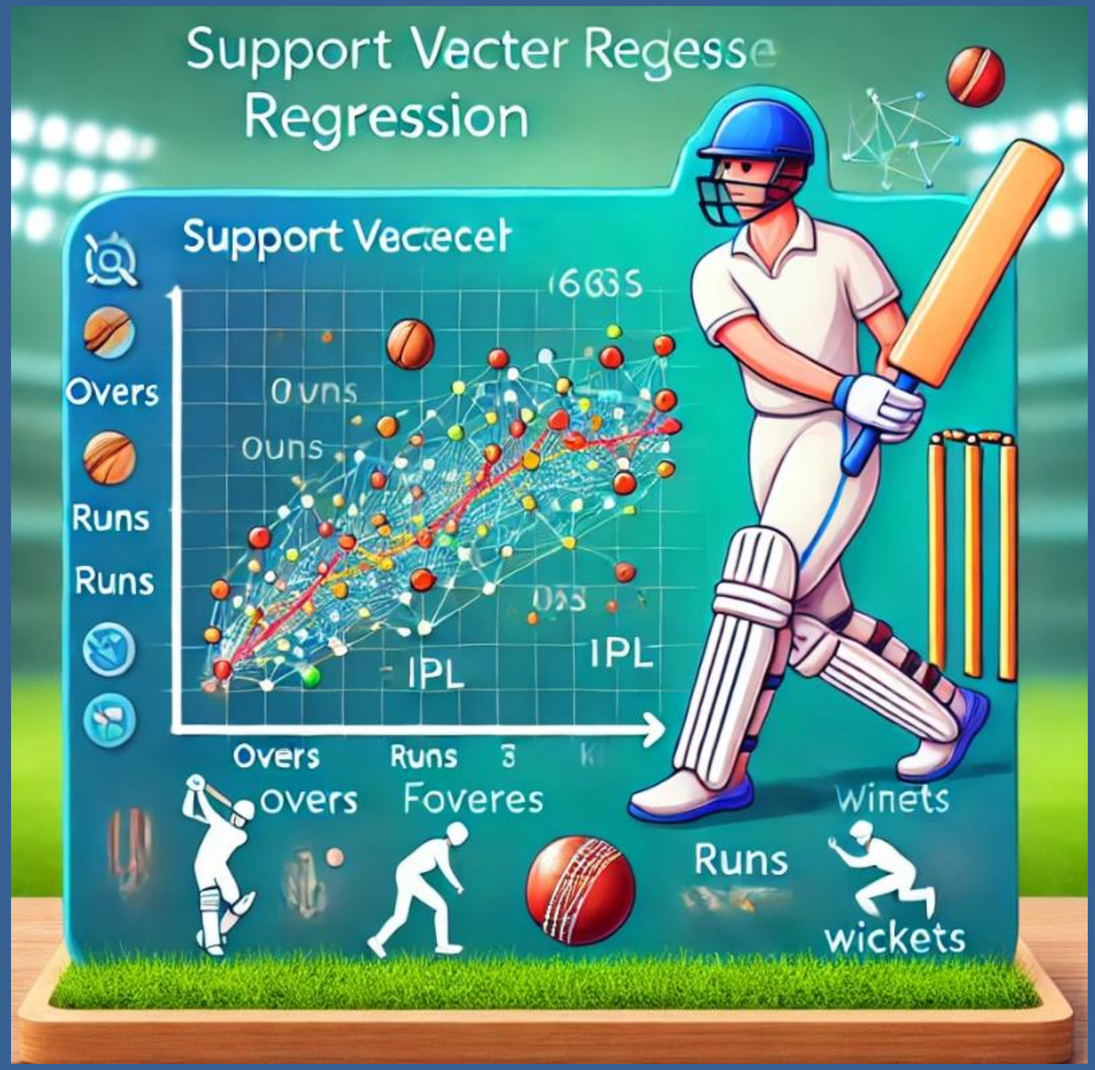
Decision Tree Regression

- A Decision Tree Regressor is a non-parametric supervised learning method used for regression tasks.
- It works by partitioning the data into subsets based on the values of input features, creating a tree-like structure where each internal node represents a "decision" based on a feature, and each leaf node represents the prediction.



Support Vector Regression

- Type of regression analysis where the goal is to find a hyperplane in a high-dimensional space that best fits the data points.
- SVR is particularly useful when dealing with non-linear relationships between the dependent variable and the independent variables.
- Represented by $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \mu$
- Unlike linear regression, coefficients β_i are determined by the SVR algorithm based on support vectors and margin.

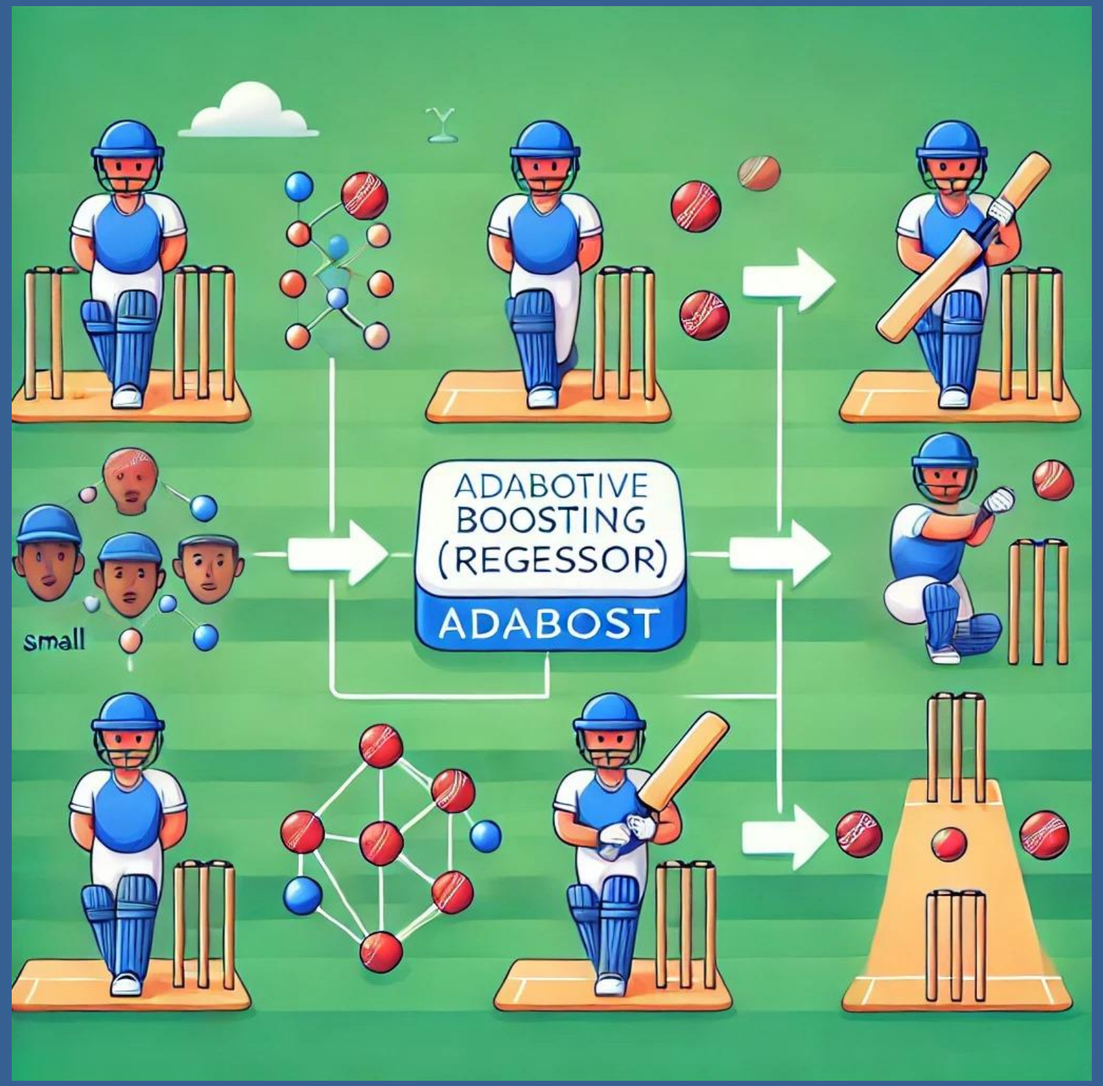


ADA BOOST Regression

- Known as Adaptive Boost, is a powerful ensemble technique that enhances the performance of weak learners by focusing on difficult-to-predict instances and combining their outputs to form a robust predictive model.
- It is widely used due to its simplicity and effectiveness in various regression tasks.

ADA BOOST Regression

- Represented by $F(x) = \sum \alpha_m h_m(x)$
- Where α is calculated based on the performance of the m-th weak learner.

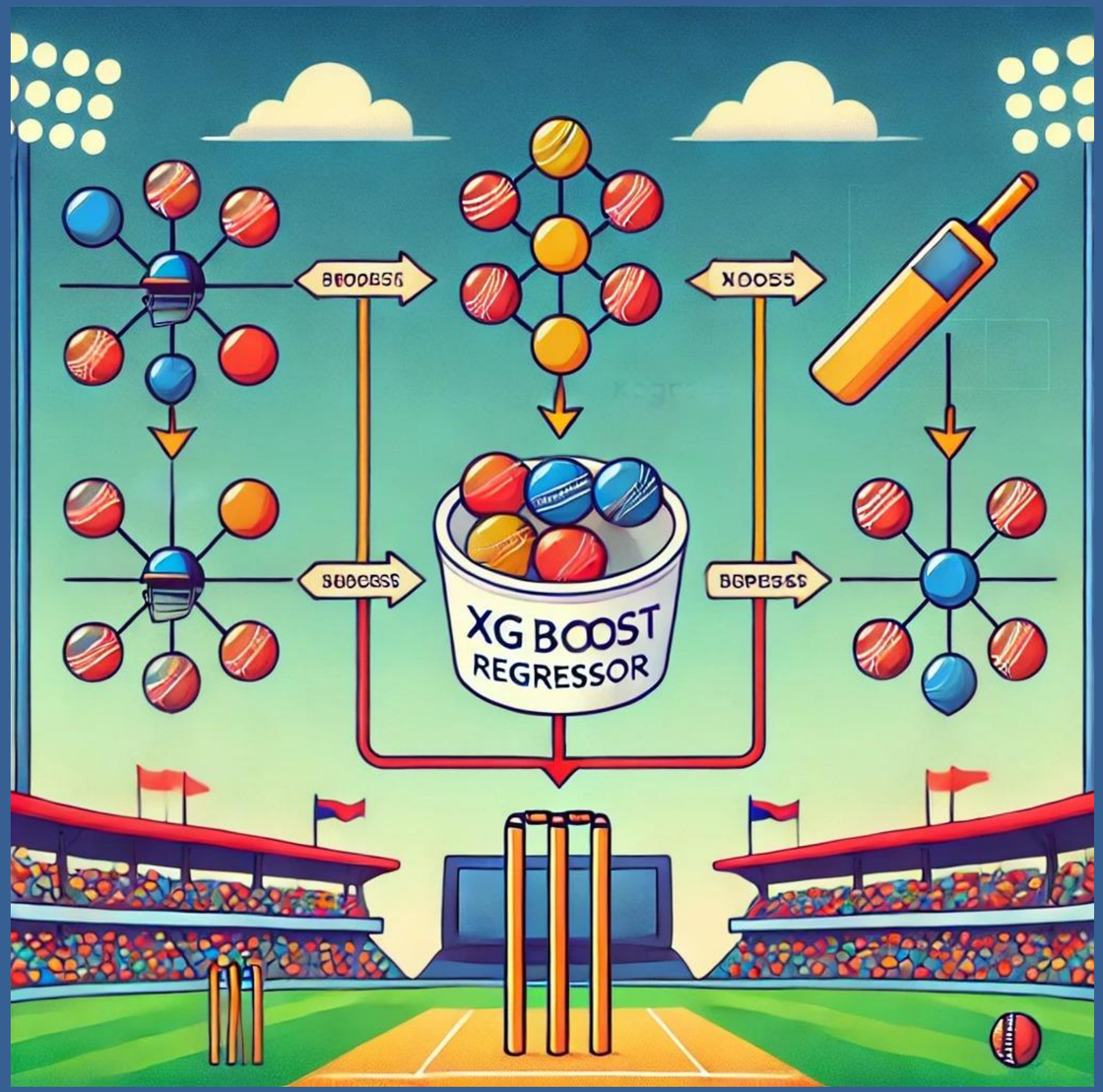


XG Boost Regression

- XGBoost, short for Extreme Gradient Boosting, is an advanced implementation of gradient boosting designed for efficiency.
- XGBoost builds an ensemble of trees sequentially, where each new tree aims to correct the errors of the previous trees.
- XGBoost includes regularization terms in the loss function to prevent overfitting. The regularization terms penalize the complexity of the model (e.g., the number of leaves in the trees) to ensure that the model generalizes well to new data.

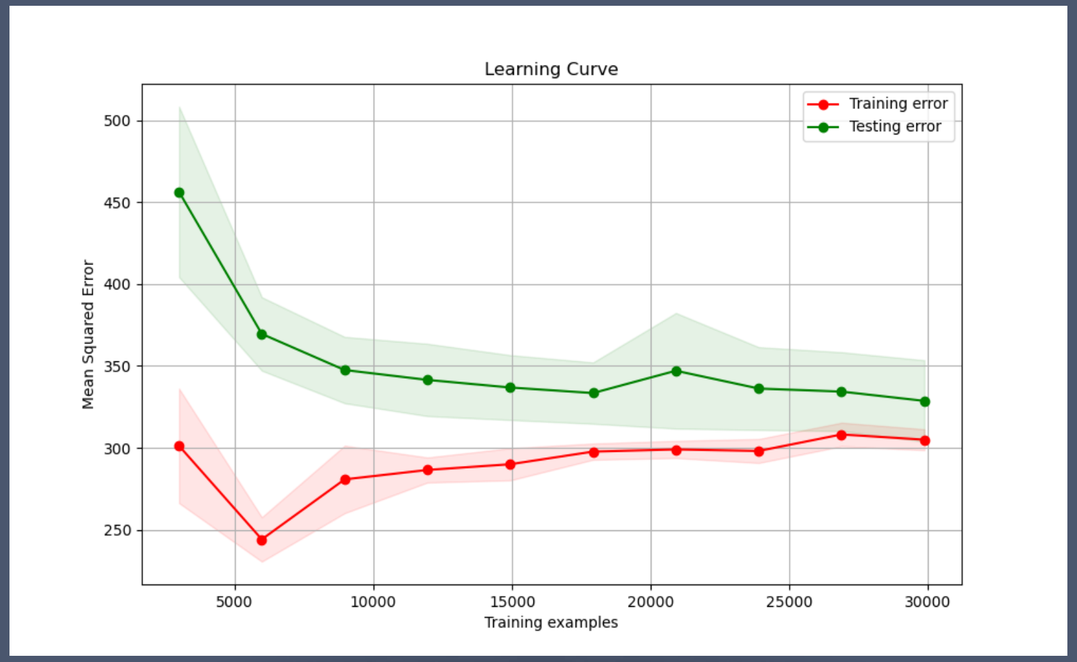
XG Boost Regression

- The objective function in XGBoost consists of two parts: the loss function L, measuring the model's performance, and the regularization term Ω, controlling the model complexity.
- Represented by
  - $\Sigma^{N}_{i=1} L(y_i, y'_i) + \Sigma^{K}_{k=1} \Omega(f_k)$
  - Where y_i is the actual value, y'_i is the predicted output and f_k is the k^{th} tree ensemble



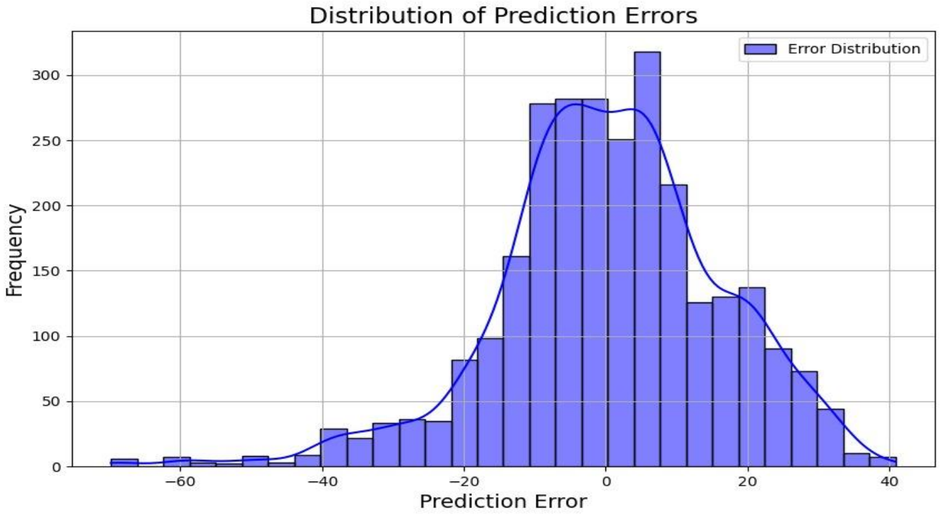
EECE 5644 MACHINE LEARNING AND PATTERN RECOGNITION

Linear Regression



Performance Curve

Analysis

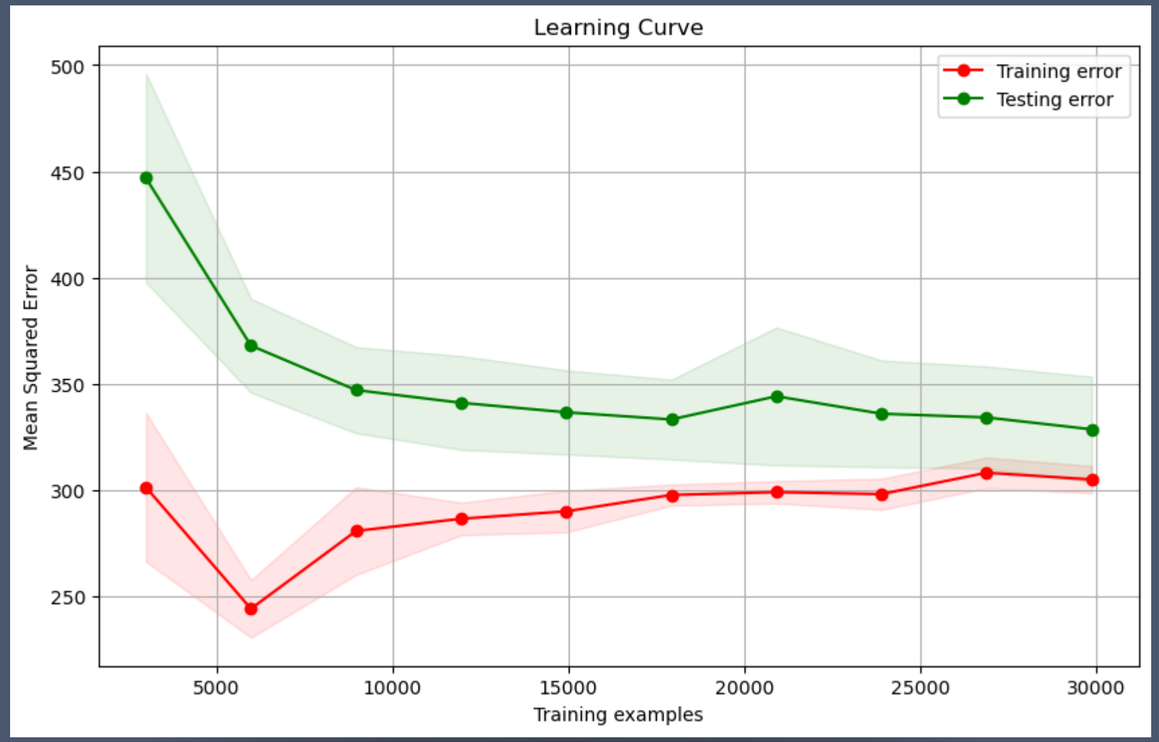


Model Evaluation

Metric	Value
Mean Absolute Error (MAE)	12.11861754619324
Mean Squared Error (MSE)	251.00792310417296
Root Mean Squared Error (RMSE)	15.843229566732061

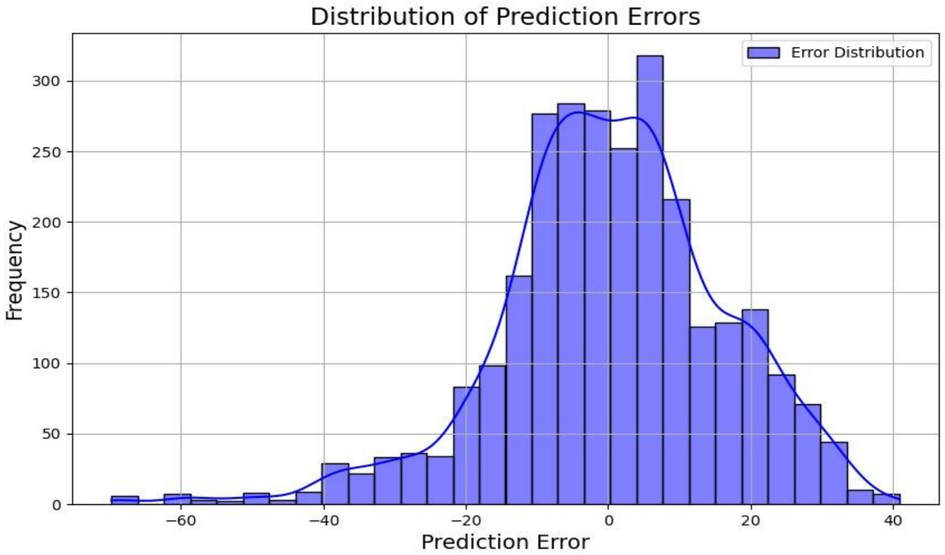
EECE 5644 MACHINE LEARNING AND PATTERN RECOGNITION

Ridge Regression



Performance Curve

Analysis

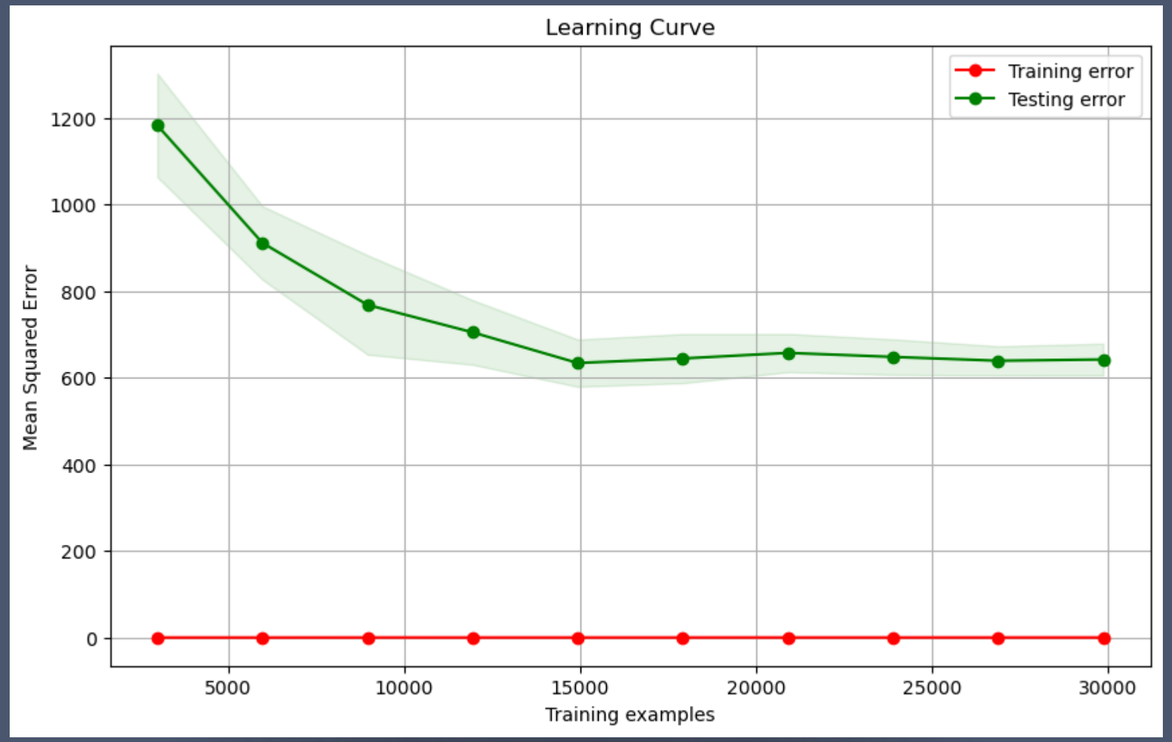


Model Evaluation

Metric	Value
Mean Absolute Error (MAE)	12.118286307248681
Mean Squared Error (MSE)	251.01379277543083
Root Mean Squared Error (RMSE)	15.843414807907758

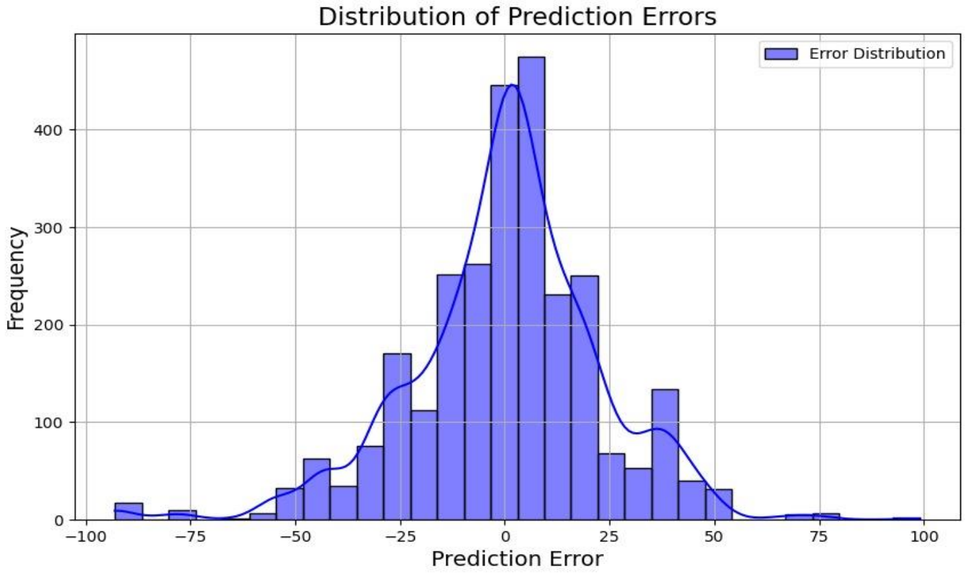
EECE 5644 MACHINE LEARNING AND PATTERN RECOGNITION

Decision Trees



Performance Curve

Analysis

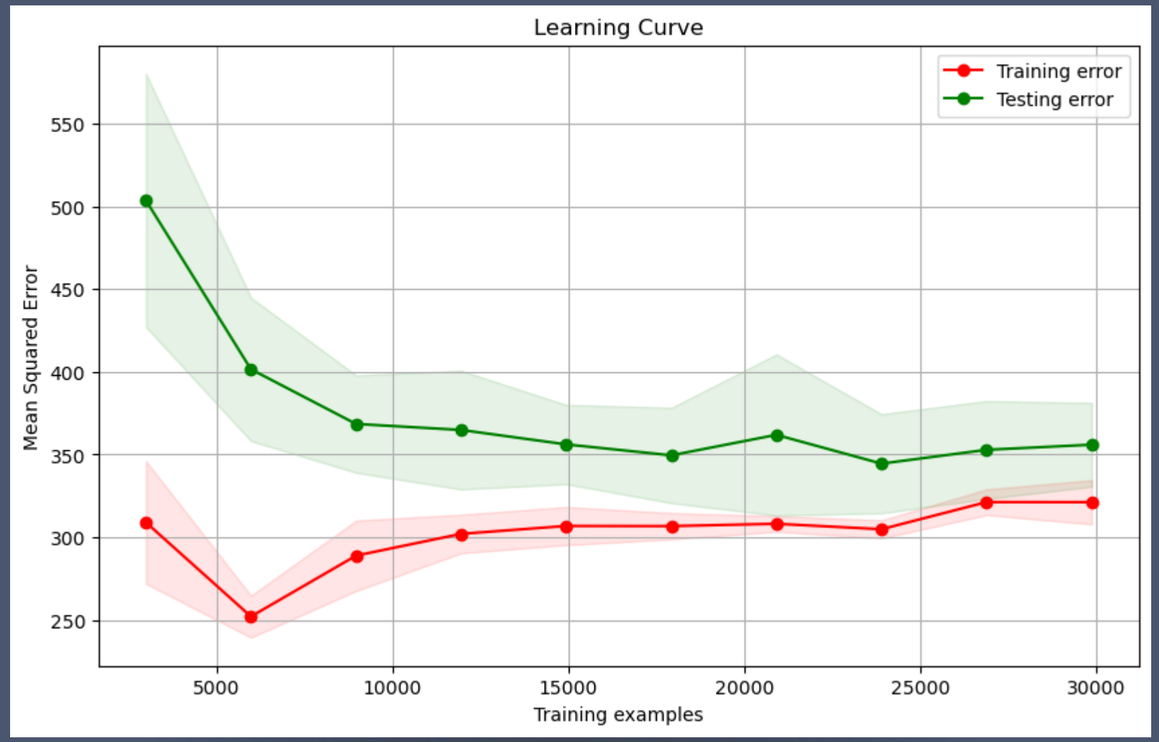


Model Evaluation

Metric	Value
Mean Absolute Error (MAE)	16.764938804895607
Mean Squared Error (MSE)	511.6900647948164
Root Mean Squared Error (RMSE)	22.620567296043138

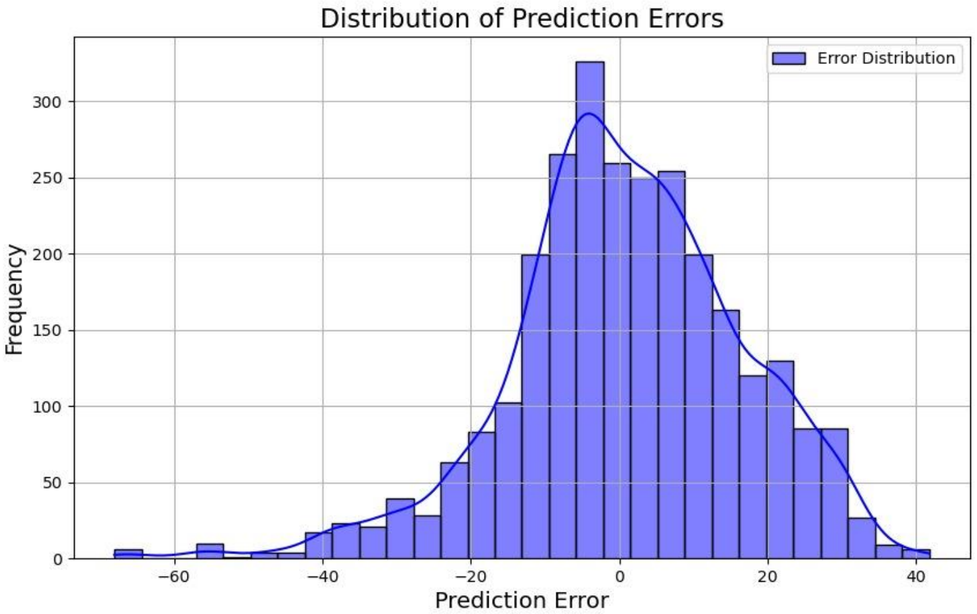
EECE 5644 MACHINE LEARNING AND PATTERN RECOGNITION

Ada Boost



Performance Curve

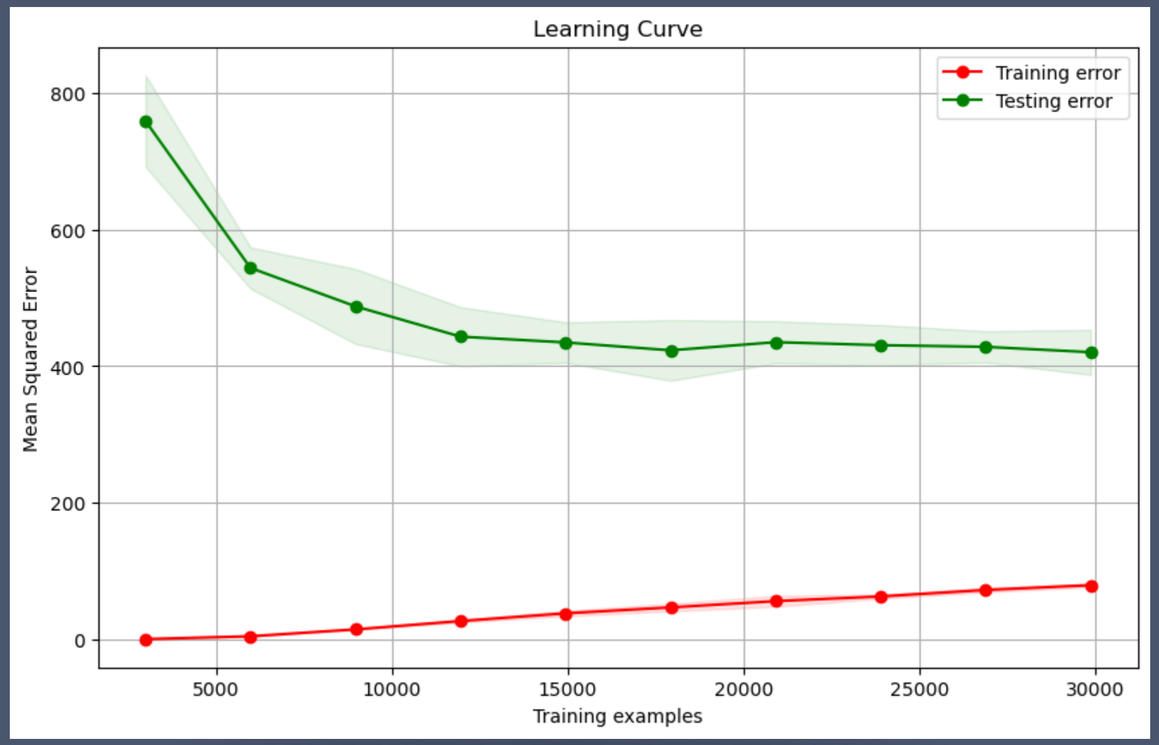
Analysis



Model Evaluation

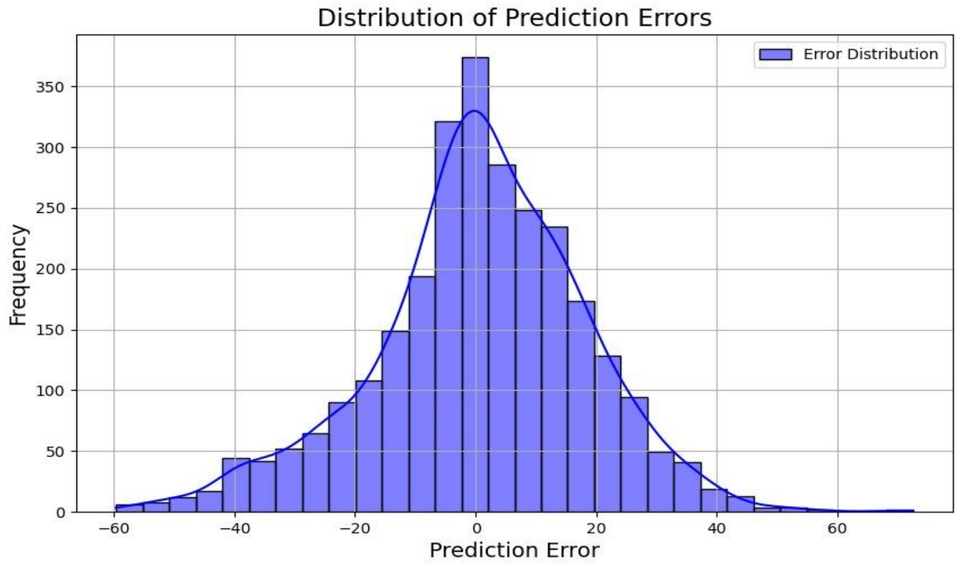
Metric	Value
Mean Absolute Error (MAE)	12.147296146769941
Mean Squared Error (MSE)	246.3139483338546
Root Mean Squared Error (RMSE)	15.694392257550293

XG Boost



Performance Curve

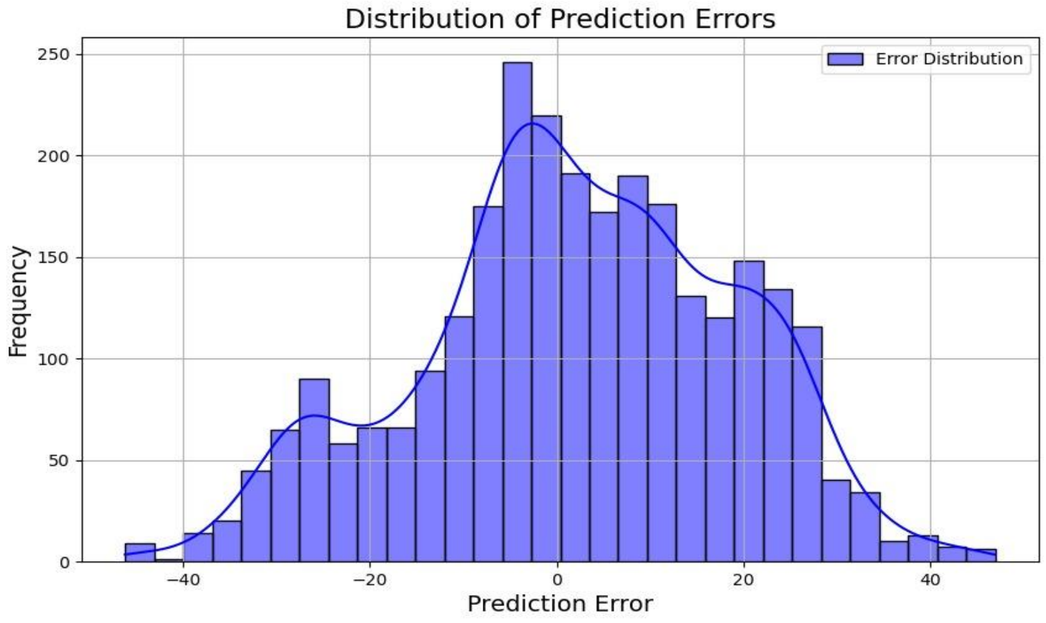
Analysis



Model Evaluation

Metric	Value
Mean Absolute Error (MAE)	13.516547755769382
Mean Squared Error (MSE)	313.2122425617864
Root Mean Squared Error (RMSE)	17.69780331160279

Support Vector Regression

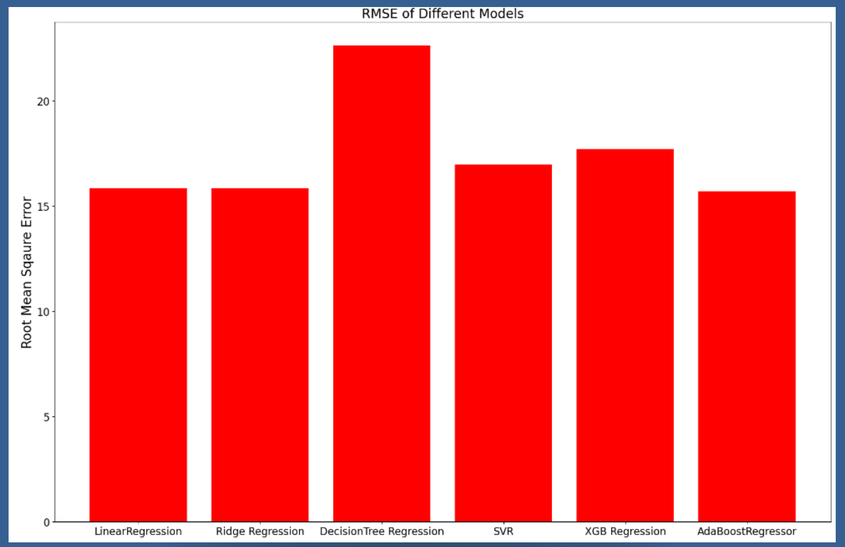


Analysis

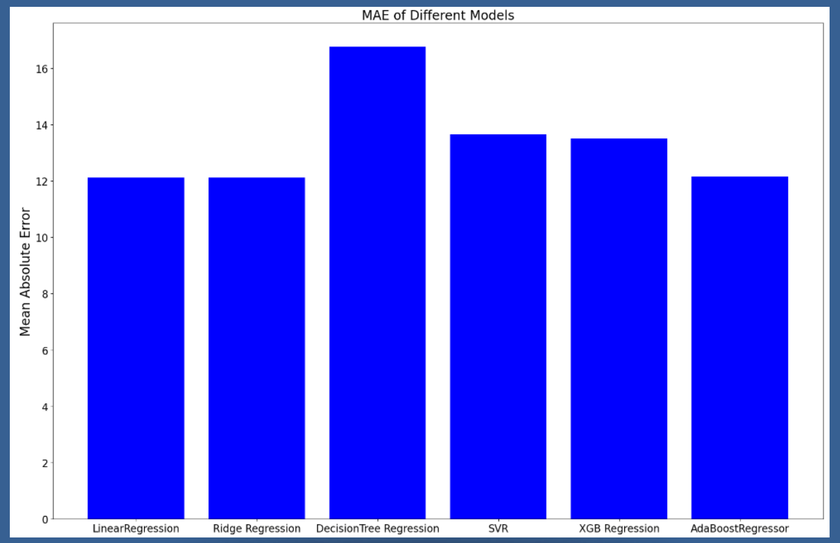
Model Evaluation

Metric	Value
Mean Absolute Error (MAE)	13.661376416962995
Mean Squared Error (MSE)	287.7359928427503
Root Mean Squared Error (RMSE)	16.96278257959909

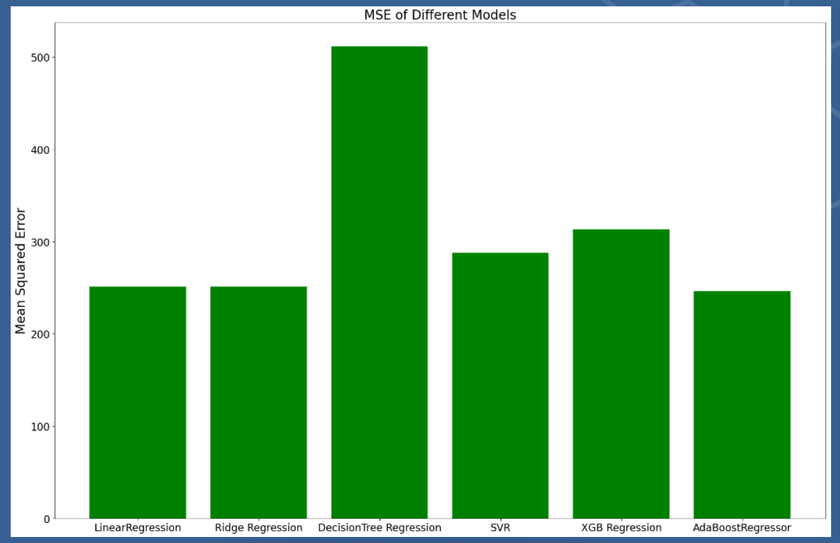
Error Comparision Curve



Root Mean Squared Error



Mean Absolute Error

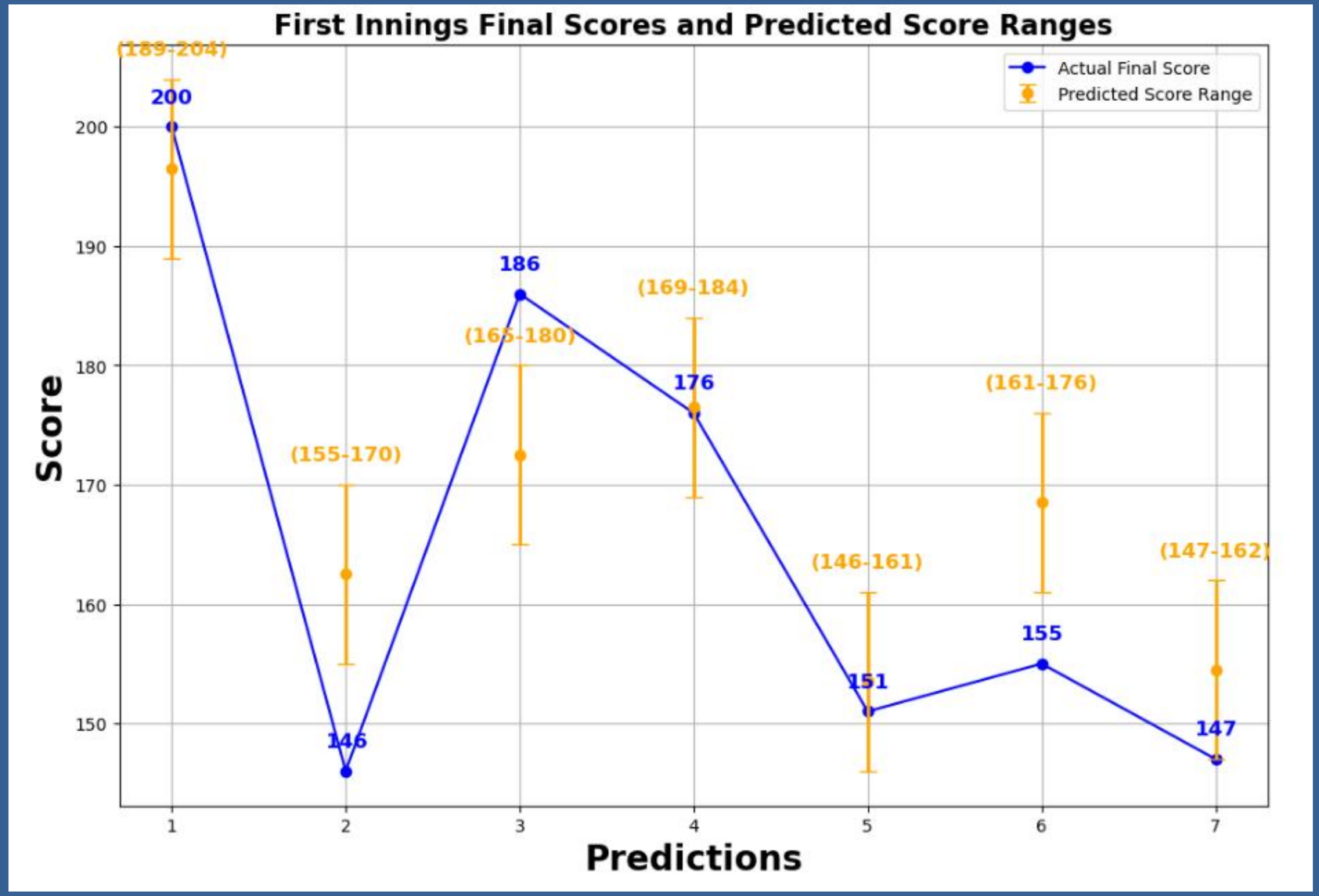


Mean Squared Error

Prediction

Batting Team	Bowling Team	Overs	Runs	Wickets	Runs in Prev 5 Overs	Wickets in Prev 5 Overs	First Innings Final Score	Final Predicted Score (Range)
Kolkata Knight Riders	Delhi Daredevils	9.2	79	2	60	1	200/9	189 to 204
Sunrisers Hyderabad	Royal Challengers Bangalore	10.5	67	3	29	1	146/10	155 to 170
Mumbai Indians	Kings XI Punjab	14.1	136	4	50	0	186/8	165 to 180
Mumbai Indians	Kings XI Punjab	12.3	113	2	55	0	176/7	169 to 184
Rajasthan Royals	Chennai Super Kings	13.3	92	5	27	2	151/7	146 to 161
Delhi Daredevils	Sunrisers Hyderabad	11.5	98	3	41	1	155/7	161 to 176
Delhi Daredevils	Chennai Super Kings	10.2	68	3	29	1	147/9	147 to 162

Prediction


First Innings Final Scores and Predicted Score Ranges
Actual Final Score
Predicted Score Range
Score
Predictions
200
186
176
146
155
151
147
(189-204)
(155-170)
(165-180)
(169-184)
(146-161)
(161-176)
(147-162)

MAJOR OUTSTANDING CHALLENGES

1. Venue Variability:

- Different cricket stadiums have unique pitch conditions that can significantly affect the scoring patterns.
- Home advantage can play a role, as teams might perform better at their home venues.

2. Weather Conditions:

- Weather elements like rain, humidity, and temperature can impact both player performance and pitch conditions.
- Rain interruptions can lead to reduced overs and influence the match outcome unpredictably.

3. Player Form and Fitness:

- The current form and fitness levels of batsmen and bowlers can vary greatly from match to match.
- Injuries or lack of form can lead to unexpected performances, making predictions more difficult.

4. Team Composition and Strategy:

- The choice of playing XI, influenced by team strategy, can change the dynamics of the game.
- Decisions on batting orders, bowling changes, and player roles are crucial but hard to predict accurately.

5. In-Game Variables:

- Factors like toss outcome, early wickets, or unexpected high scores in the powerplay can alter the course of the match.
- The psychological pressure of crucial matches or tournaments can affect player performance unpredictably.

Conclusion

The IPL Score Predictor project has successfully demonstrated the application of machine learning in predicting cricket match scores with practical accuracy.

Linear Regression emerged as the preferred model due to its balance of simplicity and performance on our dataset.

Insights gained provide valuable strategic implications for cricket management, fantasy sports platforms, and betting markets.

REFERENCES

1. Kaggle Datasets:
   - IPL Matches Data: Kaggle IPL Dataset
2. Online Resources:
   - Scikit-Learn Documentation: Scikit-Learn User Guide
   - TensorFlow Documentation: TensorFlow Guide
3. Software Libraries:
   - Python: Python Official Website
   - Pandas Documentation: Pandas User Guide
   - NumPy Documentation: NumPy User Guide
   - Matplotlib Documentation: Matplotlib User Guide
   - Seaborn Documentation: Seaborn User Guide
4. Web Resources:
   - IPL Official Website: IPL T20
   - ESPN Cricinfo: ESPN Cricinfo

THANK YOU


IPL
KINGS XI PUNJAB
SUNRISERS
Knight Riders
ROYAL CHALLENGERS
RAJASTHAN ROYALS
DELHI CAPITALS
MUMBAI INDIANS