230003026 – Inturi Siva Rishitha

Analysis of IPL Auction Data (2013-2023)

Introduction

The Indian Premier League (IPL) is one of the most prominent T20 cricket leagues, attracting top players from around the world. The auction process plays a crucial role in team formation, where franchises bid for players based on their skills and past performances. In this project, we analyze IPL auction data from 2013 to 2023 to understand trends, pricing patterns, and anomalies.

The dataset initially contained raw auction data, which required thorough cleaning and preprocessing before analysis. The objective of this report is to document the steps taken in data cleaning and manipulation to prepare the dataset for meaningful insights.

Data Cleaning

To ensure the accuracy and reliability of the dataset, we performed several data cleaning steps using Python libraries such as NumPy and Pandas. Below are the key steps undertaken:

Loading the Dataset - The dataset was loaded into a Pandas DataFrame for efficient processing and analysis.

Removing Unnecessary Index Columns - The dataset contained an index column that was redundant, so it was removed to improve readability.

Handling Duplicate Entries - Duplicate rows were identified and dropped. This step resulted in the removal of three duplicate rows, ensuring that each record in the dataset was unique.

Checking for Missing Values - The dataset was examined for missing values using isnull().sum(). Fortunately, no missing values were found, so no further imputation was required.

Data Type Correction - The Winning Bid column was initially stored as an object (string) type.

It contained additional characters (such as currency symbols or text) that were removed to convert the column into a float (numeric) type, making it suitable for mathematical operations.

Standardizing Country Names - The names of countries were reviewed, and discrepancies were corrected to ensure consistency across all records.

Updating Team Names - Over the years, some IPL teams have changed names due to sponsorship or ownership changes. The dataset was updated to reflect the correct names.

Cleaning Player Names - Player names were checked for inconsistencies such as extra spaces, lowercase letters, and formatting errors.

All names were standardized to start with capital letters for uniformity.

Data Manipulation

After cleaning, further data manipulation was performed to enhance the dataset and derive meaningful insights:

Creating New Features - Two new features were introduced to better analyze player valuations:

Price Difference: The difference between the Winning Bid and the Base Price. This helps determine how much a player’s final price exceeded the initial base price.

Increment Ratio: Calculated as (Price Difference / Base Price). This ratio indicates the percentage increase in price relative to the base price.

Previous year Winning bid, Previous year Base price, Previous year Price difference, average of previous 2 years winning bid, avg of previous 2 years base price, avg of previous 2 years price difference.

When adding these extra columns, I encountered NAN values. This was expected when attempting to retrieve previous year values from the initial years. These NAN values were subsequently replaced with the mean of their respective columns.

Outlier Removal - Since auctions often see exceptionally high or low bids, outliers were identified and replaced.

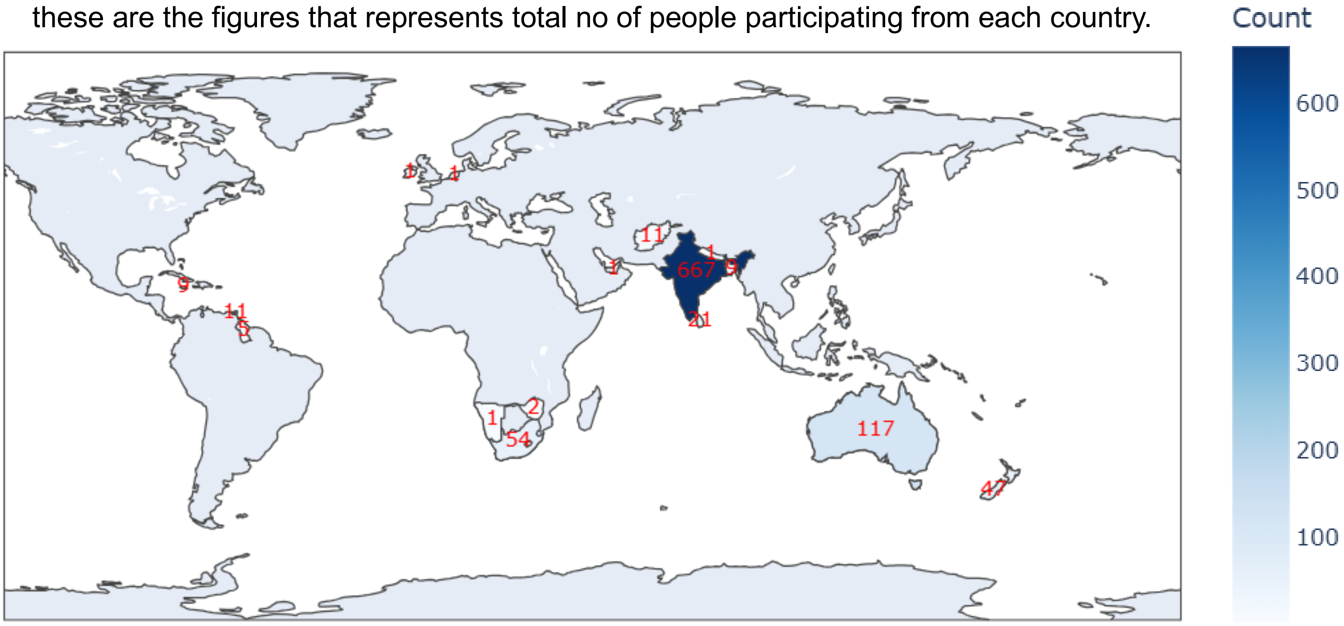
The threshold for outliers was set using mean ± 3 standard deviations for both Base Price and Winning Bid.

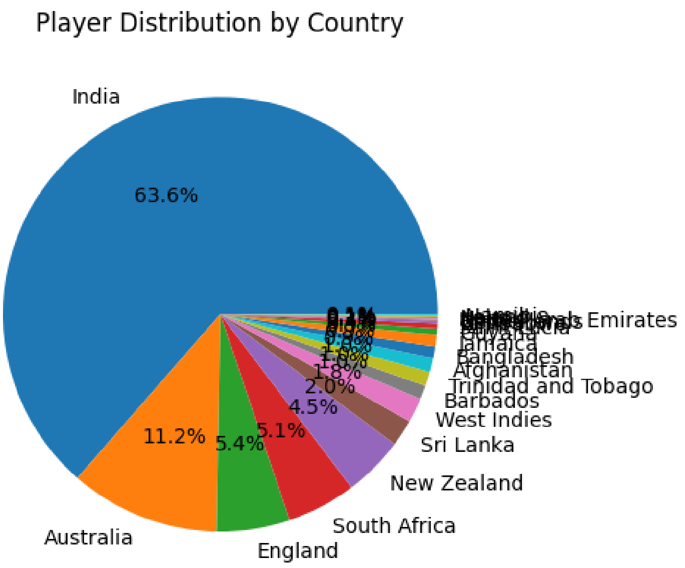
Players whose values exceeded this range were considered extreme cases and got replaced with median values.

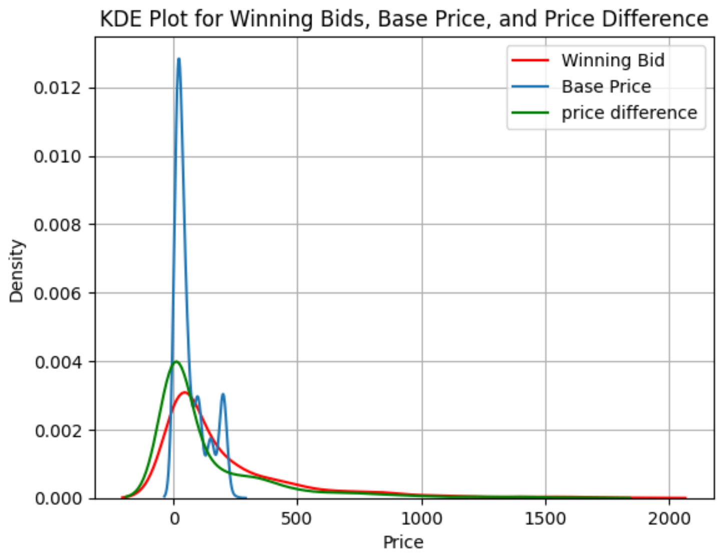
Data visualisation

After manipulating the data, Data visualisation is done.

these are the figures that represents total no of people participating from each country.

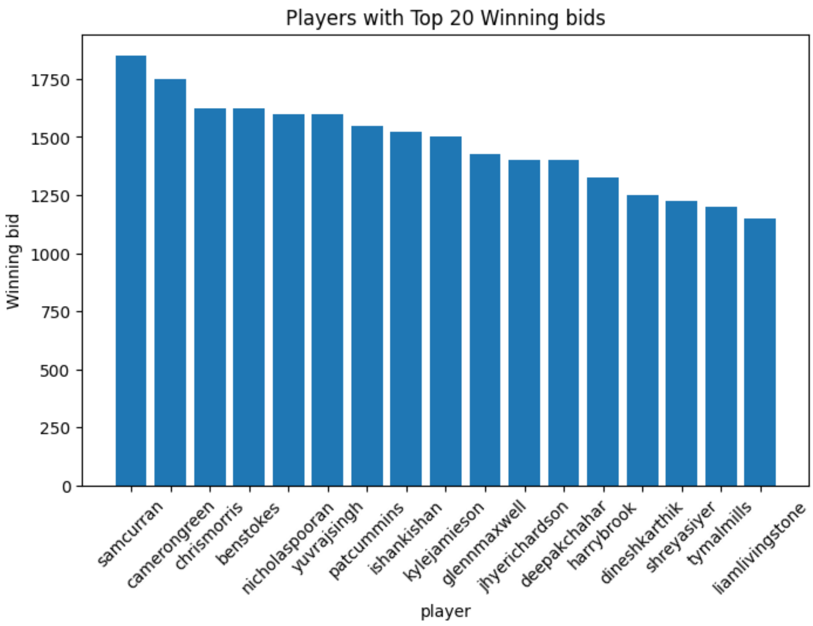


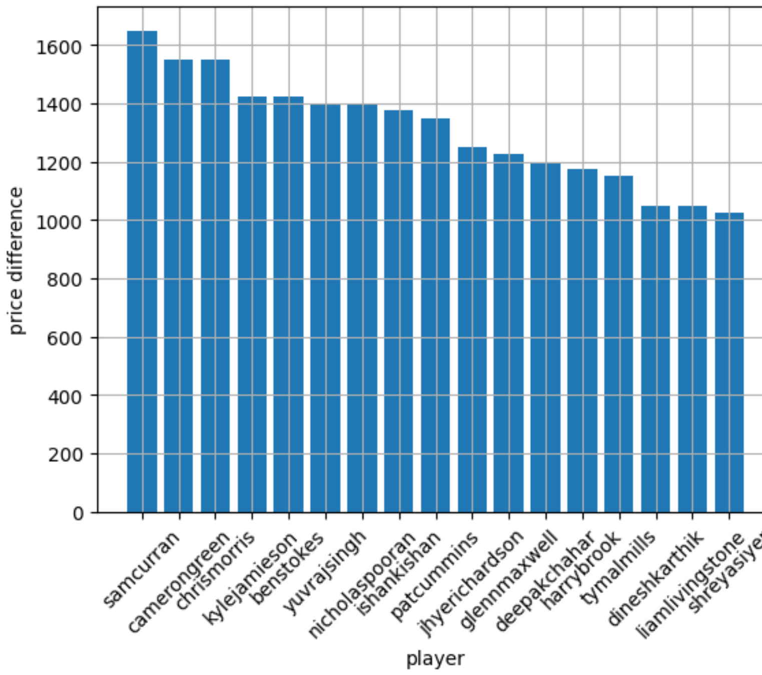




This is the pie graph that represents the Country distribution of players.

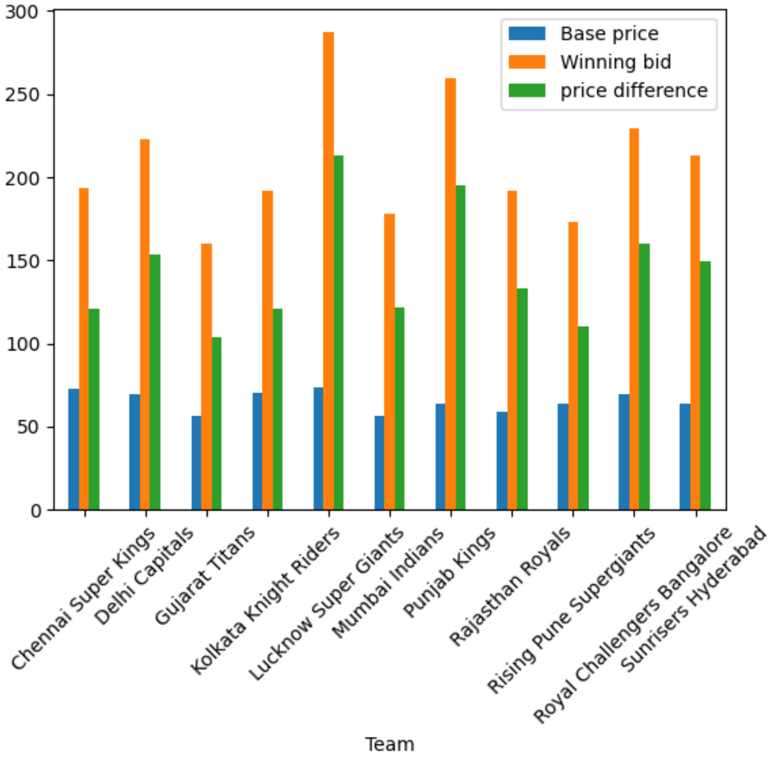
This is the KDE plot of our features which shows their distribution.

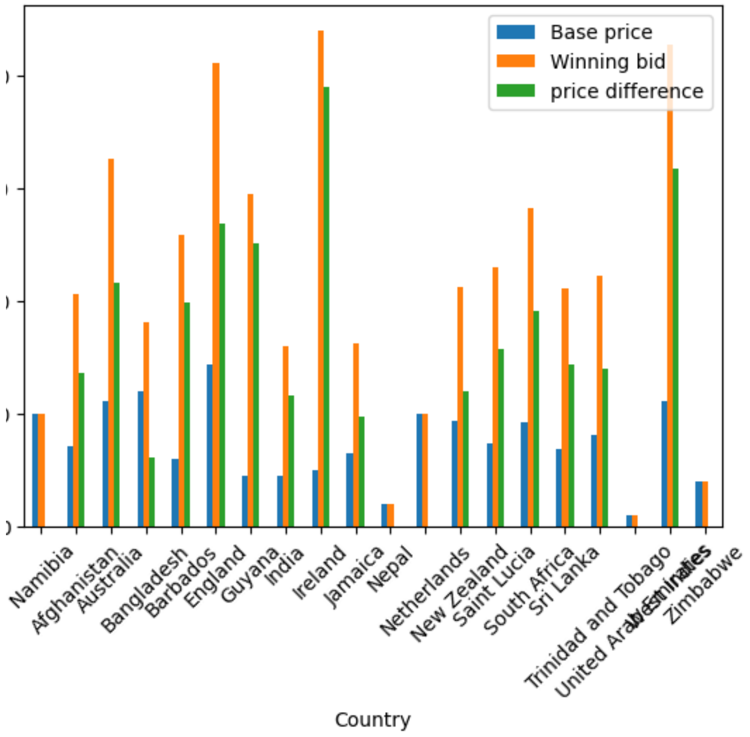




This is the plot that shows players with top 20 Winning bids

This is the plot that shows players with top 20 winning bids.



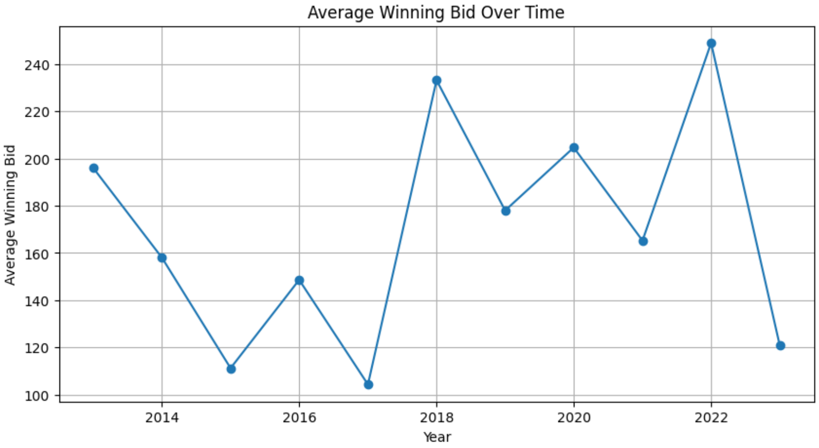


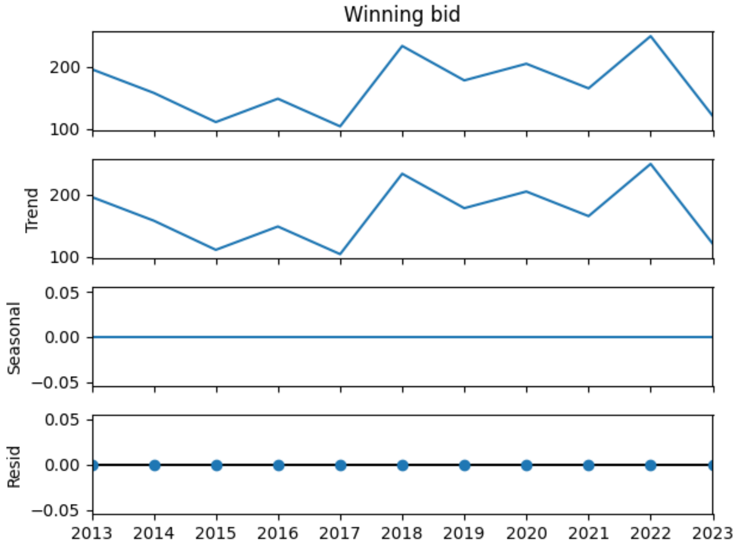
This is the plot that represents teams with With their average winning bid, base price and price difference.

This is the plot that represents countries with their average of winning bid, base Price and price difference.

Machine Learning

<u>Time series Analysis</u>





This is the plot that represents the average winning bid over years.

The 2^{nd} plot represents the trends, seasonality and residuals of the plot.

One-Hot Encoding was applied to the Player feature to make it suitable for machine learning models.

Divided the complete data set into train(80%) and test(20%)

Regression Models

Initial Approach

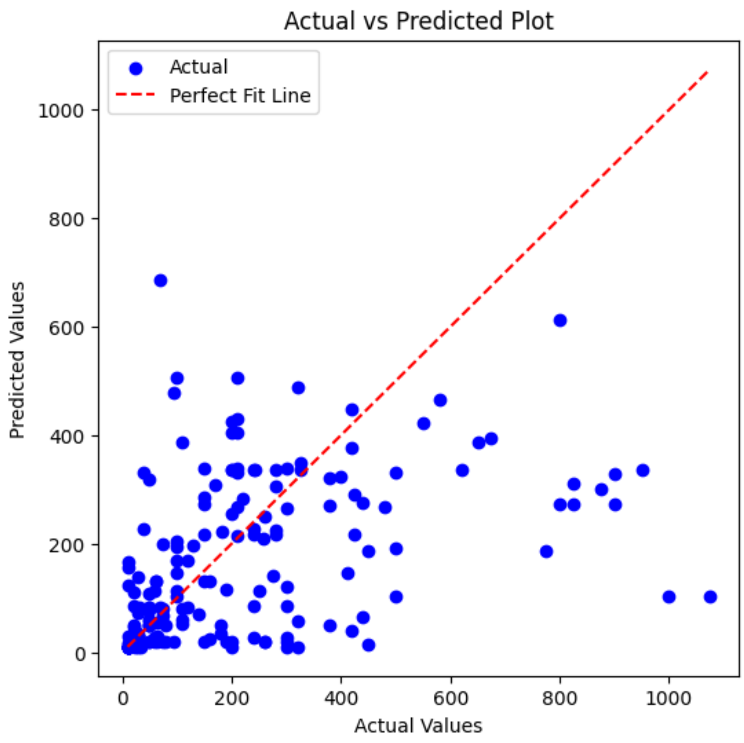
Firstly, a Random Forest Regression model was trained using 'player' and 'base price' as features. The model's performance was evaluated using Root Mean Squared Error (RMSE), resulting in an RMSE of 211.374.

Enhanced Feature Set

To improve model performance, additional features were added: 'previous year's base price', 'previous year's winning bid', and 'average of the previous 2 years' winning bids'. Using these features, the Random Forest Regression model's RMSE decreased to 199.079.

Additional Models

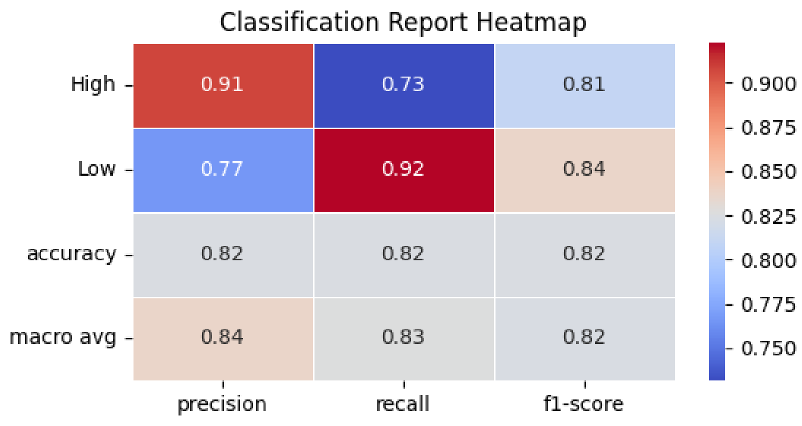
1. **Gradient Boosting Regression**:
   - RMSE: 188.082
2. **XGBoost Regression**:
   - RMSE: 193.896

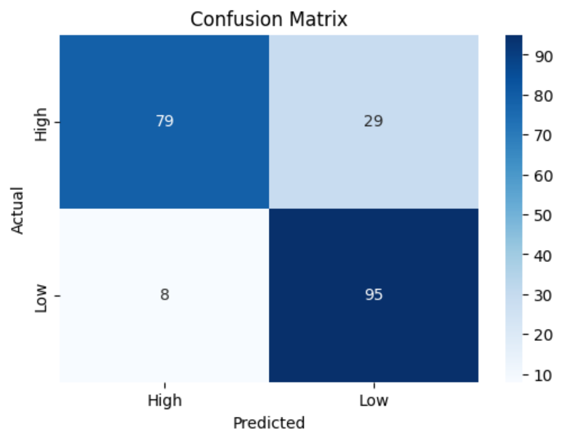


Classification Models

Binary Classification

Players were classified into two classes based on their winning bid. Players with winning bids greater than the median were classified as 'high', and the rest as 'low'. A Random Forest Classifier was trained using 'player' and 'base price' with an accuracy of 0.82.

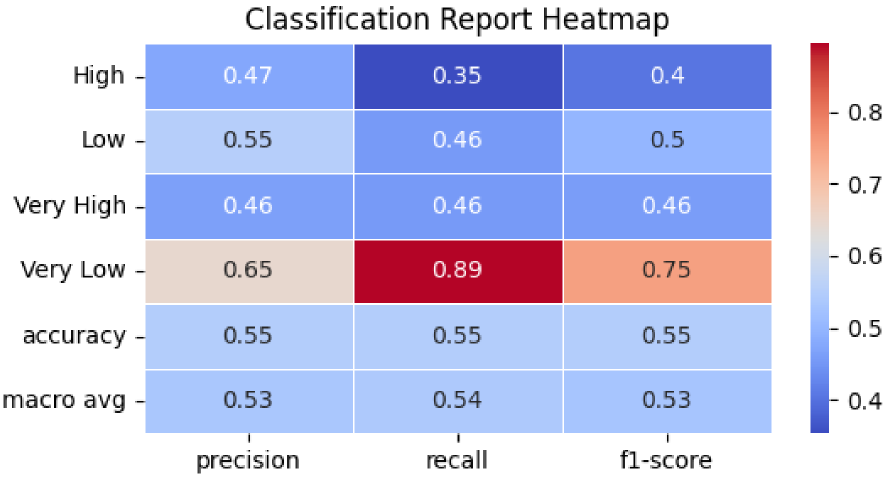


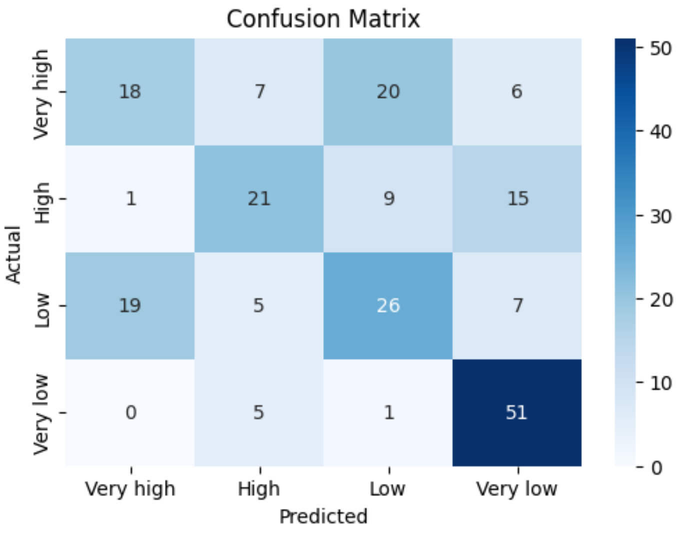


Multiclass Classification

Attempting to classify players into four classes resulted in lower accuracy:

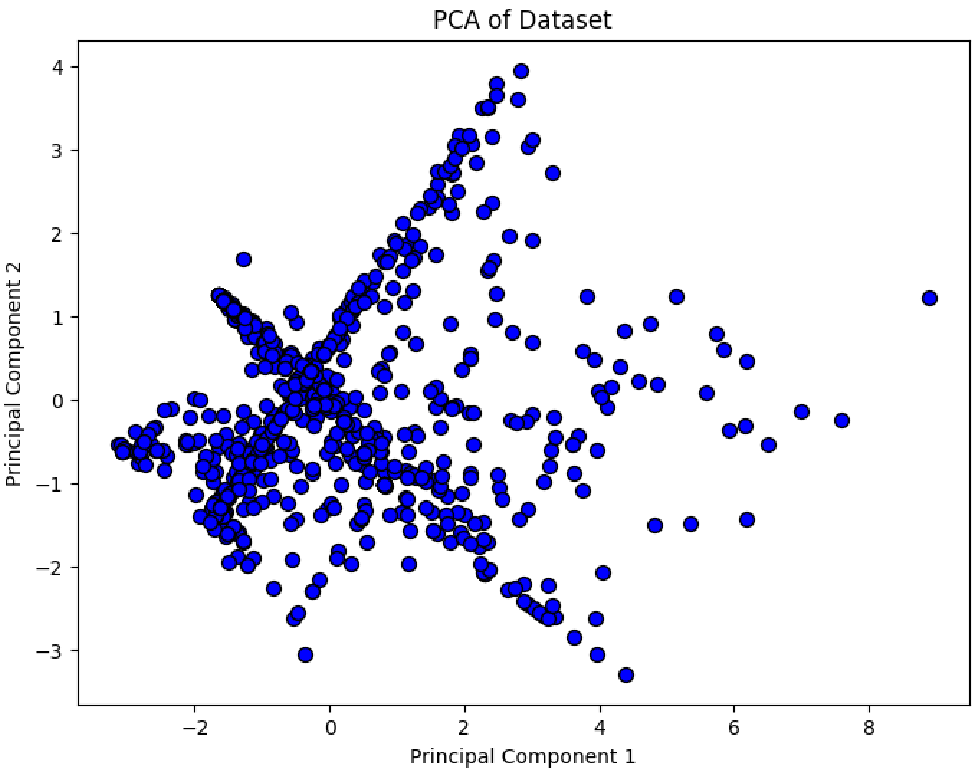
- Logistic Regression: 0.658
- Decision Tree Classifier: 0.663
- Random Forest Classifier: 0.635
- Gradient Boosting Classifier: 0.654





Principal Component Analysis (PCA)

Principal Component Analysis was performed using 'base price', 'previous year's base price', 'average of the previous 2 years' base prices', 'previous year's winning bid', and 'average of the previous 2 years' winning bids' as features.

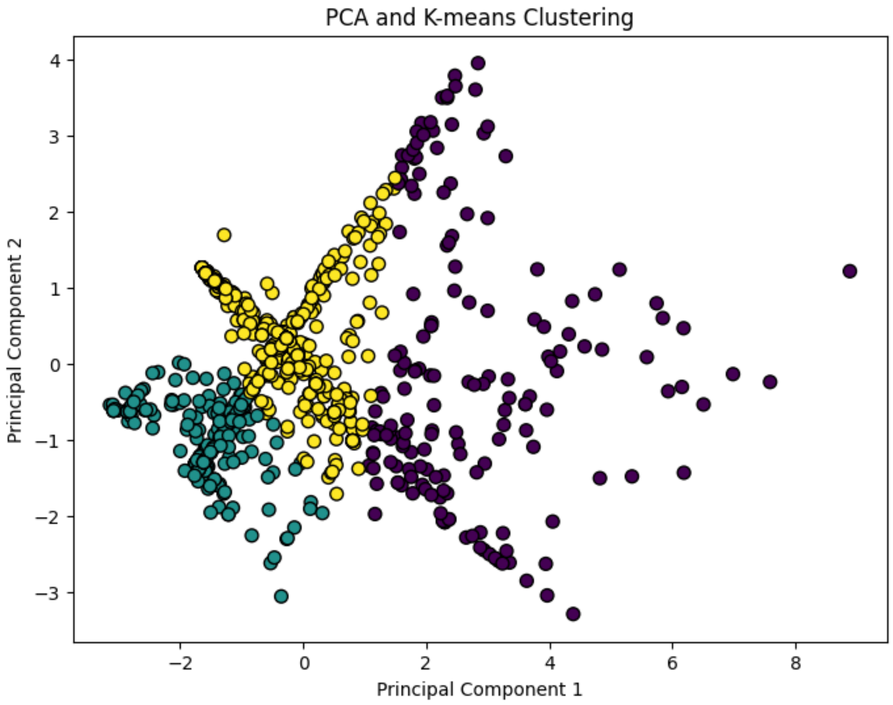


Various regression models were trained using the principal components with 'winning bid' as the target:

1. **Random Forest Regression**:
   - RMSE: 113.234
2. **Gradient Boosting Regression**:
   - RMSE: 158.979
3. **XGBoost Regression**:
   - RMSE: 119.364

Clustering

K-means clustering was used to create clusters of players, providing additional insights into player groupings.



Conclusion

Throughout the project, different models and feature sets were explored to improve the prediction of player winning bids. Random Forest Regression and Principal Component Analysis proved to be effective approaches. The models achieved notable accuracy and reduced RMSE, indicating their potential for practical application in predicting player values.