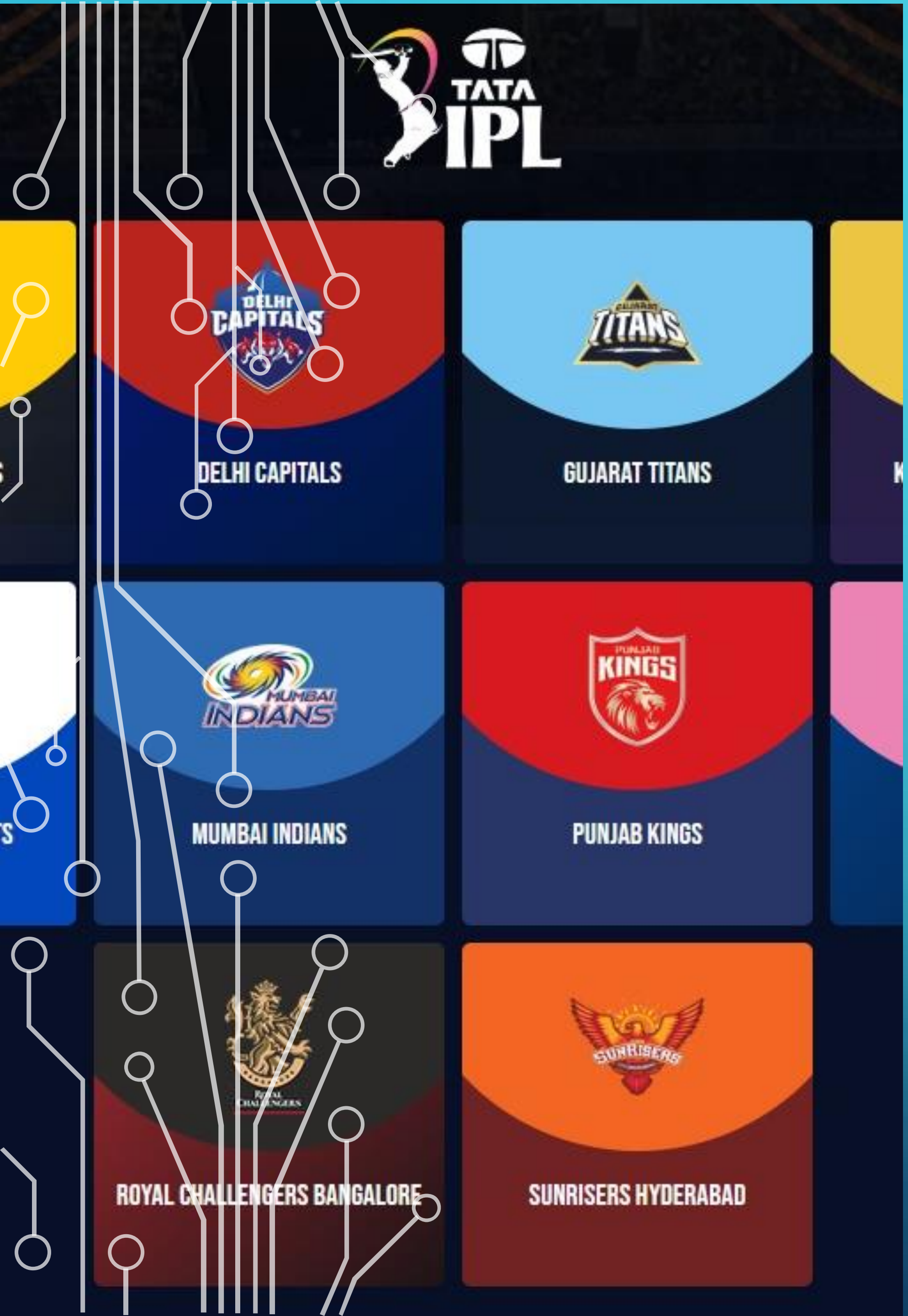
PREDICTING THE MARKET VALUE OF INDIVIDUAL CRICKETERS IN THE INDIAN PREMIER LEAGUE

SPRINGBOARD CAPSTONE PROJECT – JONNY PEARCE – APRIL 2023

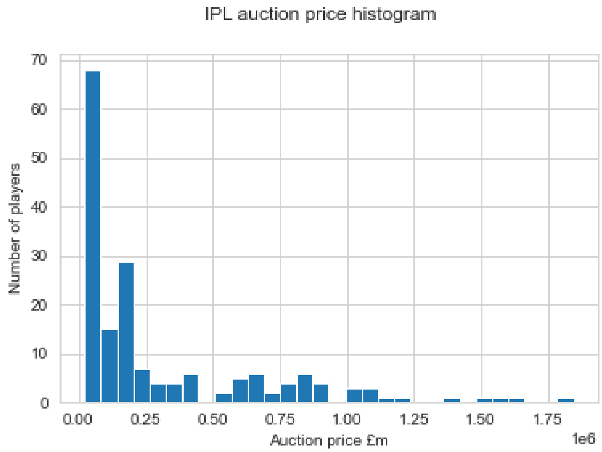


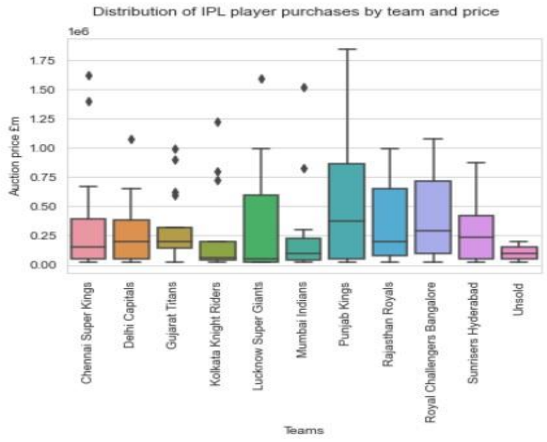
THE PROBLEM

- Indian Premier League – one of the world's biggest international sporting businesses
- US$4.7bn market value in the 2021 season – TV/media, advertising, sponsorship, merchandising and betting markets all depend on success of teams and players
- Operates on a franchise basis, with high profile international teams run by businesses, which bid/compete for the best players in an international auction
- Buying the right players, having the right team purchasing strategy – crucial to business and sporting success
- Overpaid upstarts or worth their weight in gold? What makes the most valuable IPL player and are they worth it?

THE DATA

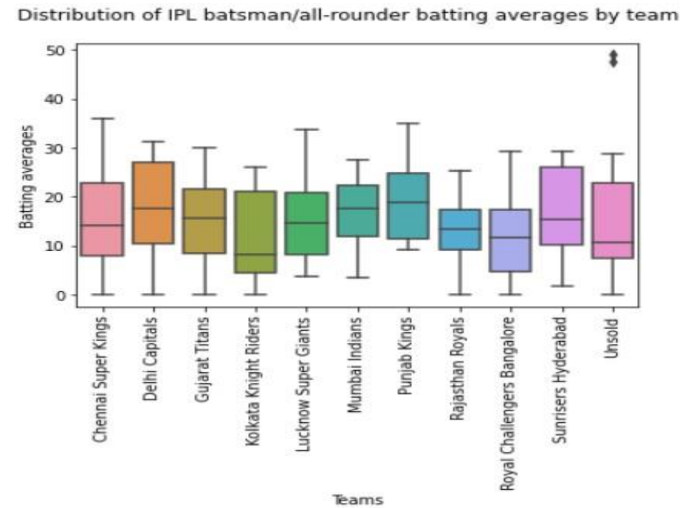
- Last two years of IPL auction data and player performance data
- Scraped from IPL website
  - 1000+ players in the auctions
  - 250+ players' performance data
  - started with 50 features, reduced to 28
- Additional non-IPL performance data gathered, along with international cricketing salary information

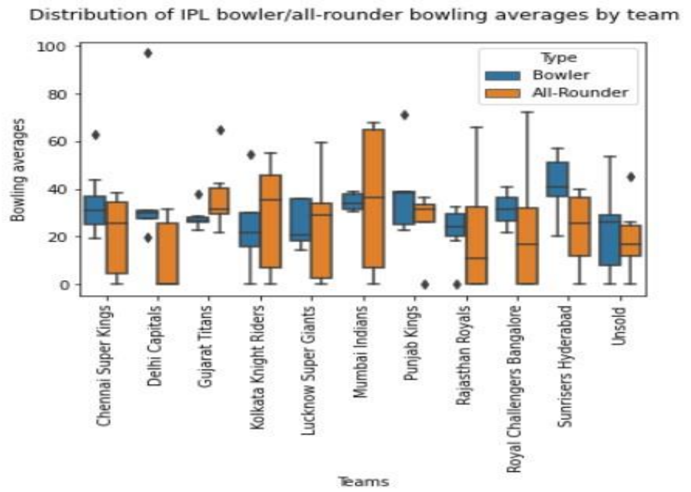




EXPLORATORY ANALYSIS (1)

- Right-skewed distribution for player purchases – very few players command high prices
- Vast majority bought at low or mid-range (pile ‘em high, sell ‘em cheap)
- Different strategies and success levels when looked at from a team perspective

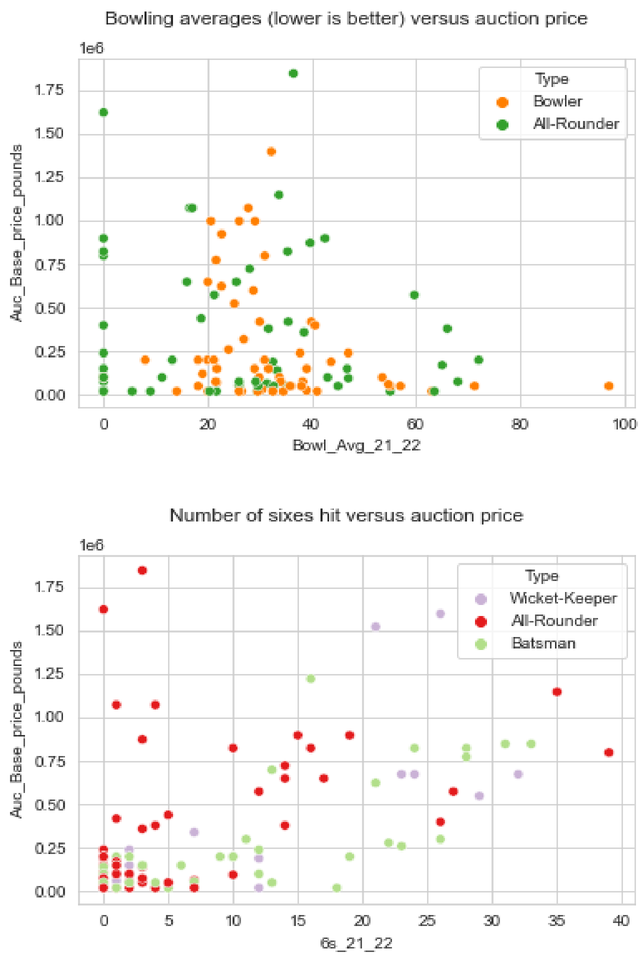




EXPLORATORY ANALYSIS (2)

- Remember – batting averages are better when they're higher, bowling averages when they're lower
- Those spending the big money don't always get the best results
- Significant variation across teams
- Non-specialists often outperform specialists

EXPLORATORY ANALYSIS (3)



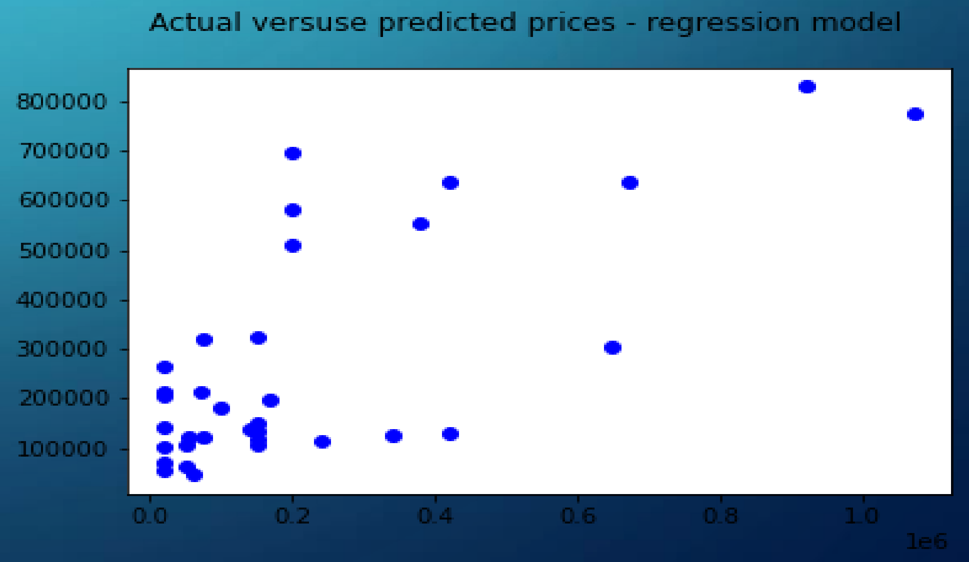
- Scatter plots and heat maps produced
- Hard to identify strong correlations, but interesting difference between types of players
- Positive relationships between pricing and:
- For batting: total runs, number of 50s, 6s scored
- For bowlers: total wickets taken, bowling average, strike rate, dot balls bowled

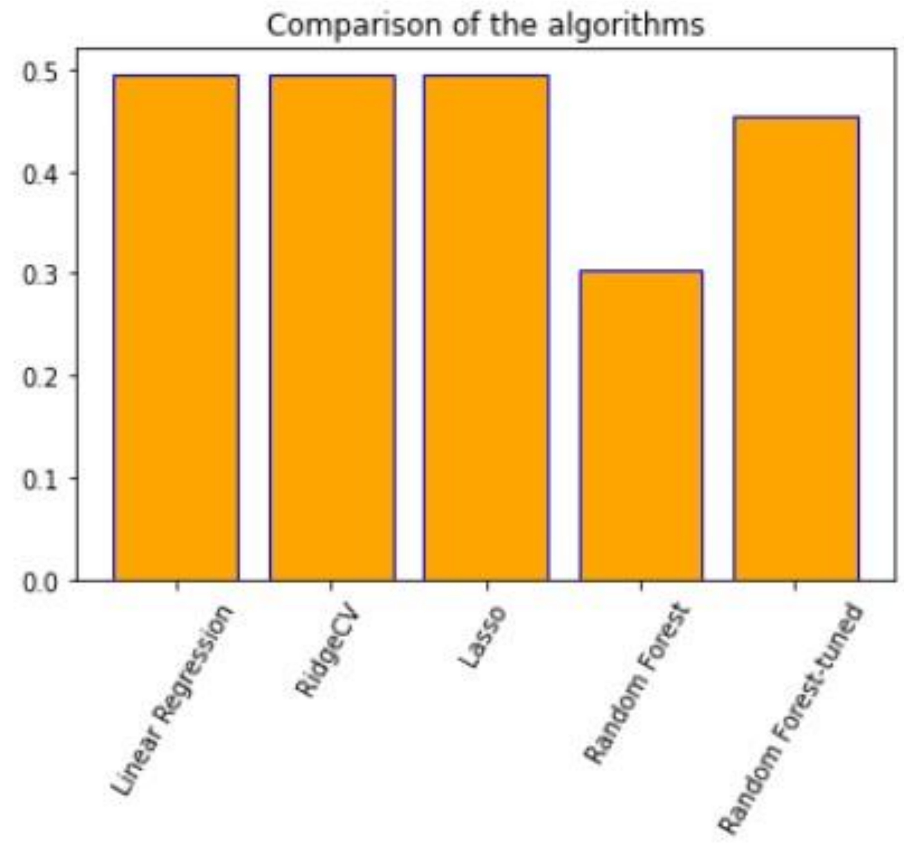
PRE-PROCESSING

- Feature data standardised:
  - Categorical data (eg, player type) – one-hot encoded
  - Continuous data (averages, strike rates, etc) – StandardScaler()
  - Discrete data (number of not-outs, 100s, 50s, 5-wicket hauls, etc) – MinMaxScaler()
- Principal Component Analysis:
  - Of the 26 final features, first four or five PCA components accounted for the majority of the variance
  - Key features:
    - Bowling economy
    - Bowing strike rate
    - Batting average
    - Batting runs total
    - Bowling average

MODEL SELECTION

- Regression problem:
  - Linear Regression
  - Ridge Regression
  - Lasso Regression
  - RandomForestRegressor
- All performed poorly – regression coefficient of 0.49





RESULTS

- Even with hyperparameter running including:
  - alpha and solver for Ridge Regression (involving the use of GridSearchCV),
  - the use of an adjusted alpha for Lasso Regression (found from cross-validation on the training data), and
  - a full parameter grid for the Random Forest Regressor (using GridSearchCV again).

CONCLUSIONS AND NEXT STEPS

Number of possible reasons for poor model performance

- Dataset is too small
- Distortion within some of the data
- Reality that players underperform or overperform in relation to their sale price
- Limited engineering of the existing features
- Highly priced players not playing, due to injury or unavailability
- The player categories may be unhelpful
- Lack of corresponding data on team success

Next steps:

- Get/buy more data on previous years
- Include data on team performance
- Address data distortions and create additional features
- Try again, with different models

QUESTIONS?