Patrick Thygesen

Dr. Chou

CSC485

May 5th, 2024

**Indian Premier League Analysis**

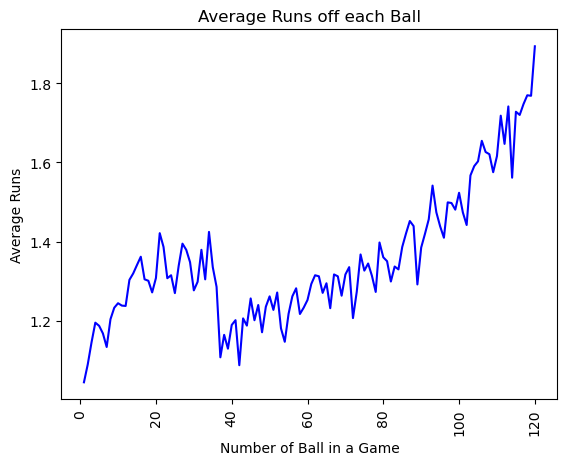
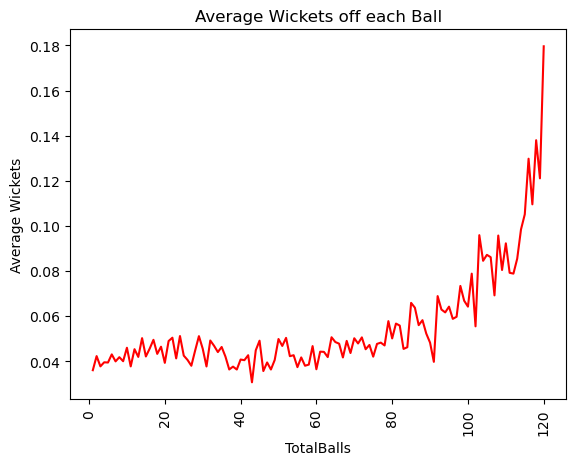
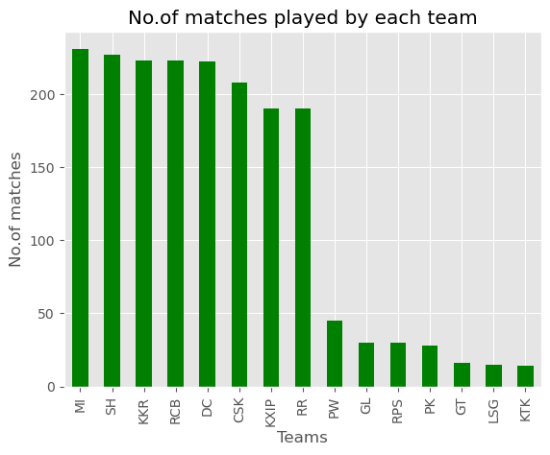
The goals of this project are to identify any significant trends in Indian Premier League (IPL) games that may influence the outcome of the game and to also create a winning prediction model that will calculate the probability of a team winning the game and losing the game after each over in the second innings of a game. Over the past 5 years, the idea of data analytics in cricket has become more prominent in professional teams with each team now having a team of data scientists to analyze the performance of their players and opposition teams to help optimize a team’s strategy going into every game. This project will combine two data sets: one that has information for each game in the IPL from 2008 to 2022 and another that has information on every ball bowled in the IPL from 2008 to 2022. Both datasets will be used for data analysis and the IPL Ball by Ball data set, with some information added from the IPL Game by Game dataset, will be used for the winning prediction model.

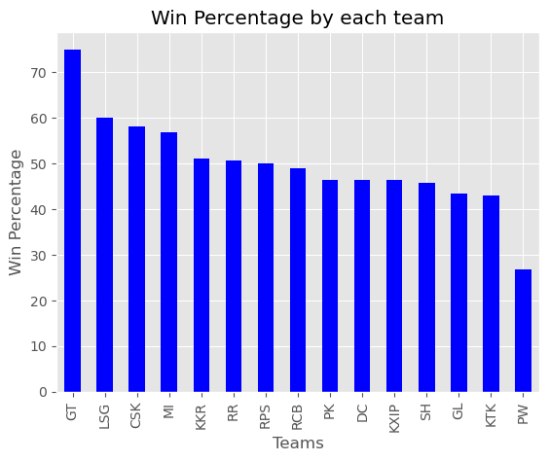
The IPL has occurred every year from 2008 to 2024 and has 8 to 10 teams playing every season. Each season has been played in India apart from the 2009 season (played in South Africa) and the 2020 and 2021 seasons which were played in the UAE due to COVID-19. The IPL is a T20 tournament which means that the team batting first has 20 overs to set a score for the opposition to try to chase in 20 overs. However, if a team loses all 10 wickets before the end of 20 overs the innings the team is bowled out and the innings is over. Just like in all T20 games, for the first 6 overs of an innings, the fielding team is only allowed to players outside the inner ring, and after the 6th over the fielding team is allowed to have a maximum of five players outside the inner ring.

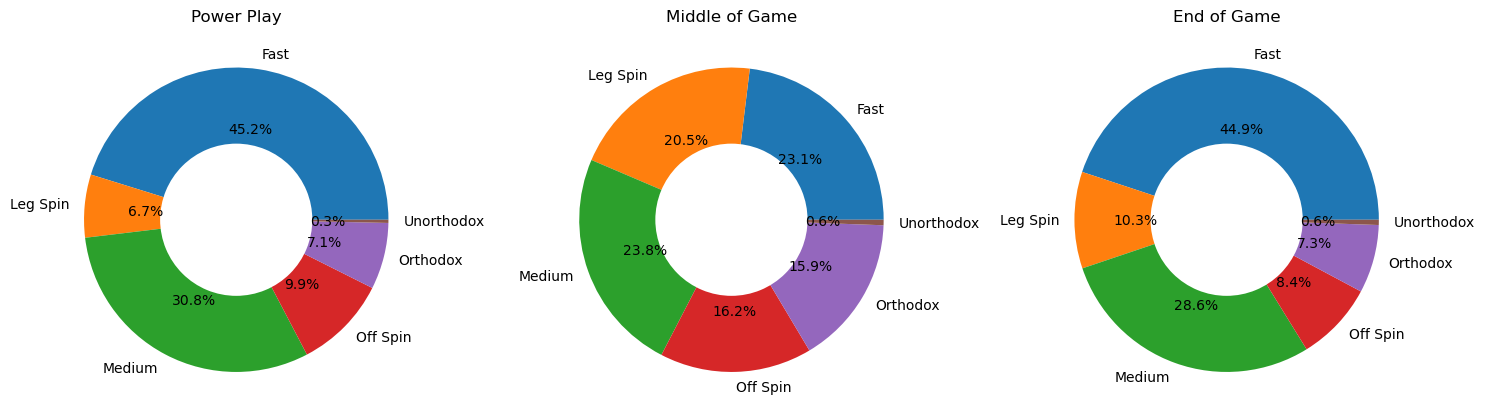
**Data Creation:**

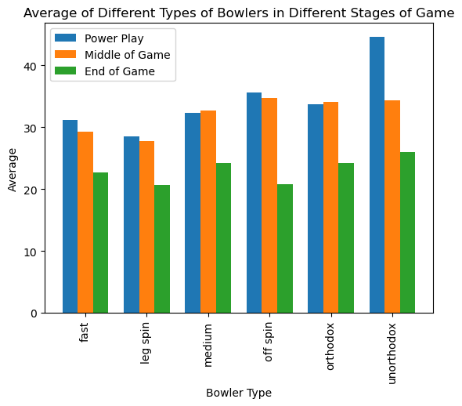
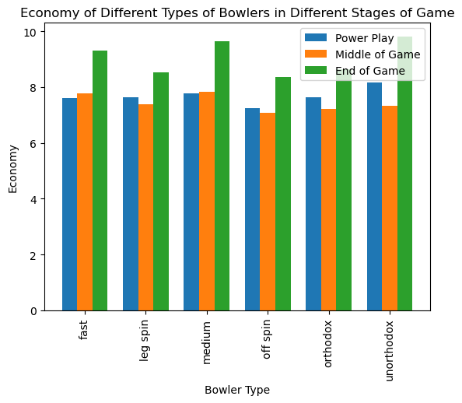
The Ball by Ball data set has a total of 225,954 observations with each ball being uniquely identified by the game ID, innings, over number, and the ball number. Other important information in this data set includes runs and wickets off each ball, the number of extras (added runs for illegal deliveries), the bowler’s and batsman’s names, and the bowling and batting teams. For each bowler and batsman in the data, information about the type of bowler they are (fast, medium, off-spin, wrist spin, orthodox, and unorthodox) and what type of batsman they are was collected from ESPN Cricinfo and added to the data set as well. The Game by Game dataset has a total of 946 observations and has information on the teams playing, the location of the games (city and venue), the toss decision, the winning team, and the winning margin. Each game also has a unique game ID which is supplied by ESPN Cricinfo and aligns with the game ID in the Ball by Ball dataset. The game ID was integral in combining data from both data sets. For example, the city and venue, and the winning team of each game were added to the ball-by-ball dataset using the game ID. Similarly, the number of runs and wickets in each innings was added to the game-by-game dataset. The only challenge in the datasets was having to correct the names of teams and stadiums that had changed their names due to changes in owners or branding. For example, Delhi Daredevils changed to Delhi Capitals for branding reasons and Deccan Chargers changed to Sunrisers Hyderabad due to a change in ownership.

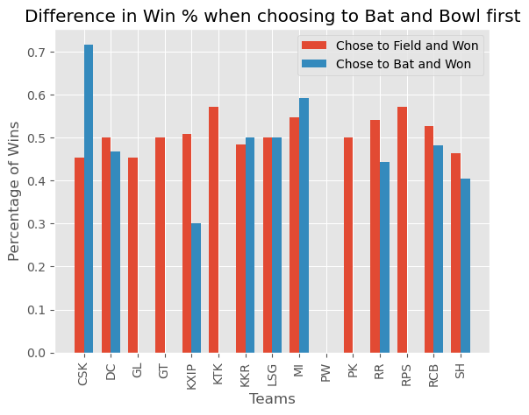
**Exploratory Data Analysis**

Given the structure of the game, teams tend to play more or less aggressively at different stages of the game. This means that there may exist certain identifiable patterns in a game. In Figure 1, we can see that teams tend to score much quicker in the first 6 over (36 balls) of the game, and then there is a steep drop off in runs scored. This is likely due to the powerplay where there are fewer players in the outfield making it easier to score runs. This may be important for teams to see, suggesting that it is important for teams to score quickly early so using more aggressive batsmen is important. Teams also tend to score much quicker at the end of an innings. There are no significant findings from the amount of wickets throughout a game, with the only observation being that wickets are more likely to be taken at the end of the game when teams are trying to score runs at a faster rate which creates more risks to lose wickets.

In the IPL there have been 15 different teams play, with the original 8 teams being the Chennai Super Kings, Delhi Daredevils (now Capitals), Kolkata Knight Riders, Mumbai Indians, Punjab Kings, Rajasthan Royals, Deccan Chargers (now Sunrisers Hyderabad), and Royal Challengers Bengaluru were all founded in 2008 when the IPL began and have played in every season, except for Chennai and Rajasthan who were suspended in 2016 and 2017 for financial breaches. In place of them, the Gujarat Lions and Rising Pune Super Giant played in IPL for those two seasons. Kochi Tuskers Kerala (2011) and Pune Warriors (2011-13) also played in the IPL for small stints, and the two most recent additions have been Gujarat Titans and Lucknow Supergiants which were founded for the 2022 season and still play in the IPL. This information explains the distribution of games played by each team (Figure 1). The Lucknow Supergiants and Gujarat Titans may have the best winning percentage but they have only played one season so this is biased. So the most successful teams would be Chennai Super Kings and Mumbai Indians, which is understandable since they have also won the most IPL titles (5 each, 3 more than the next team KKR).

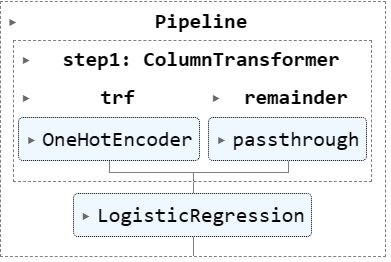
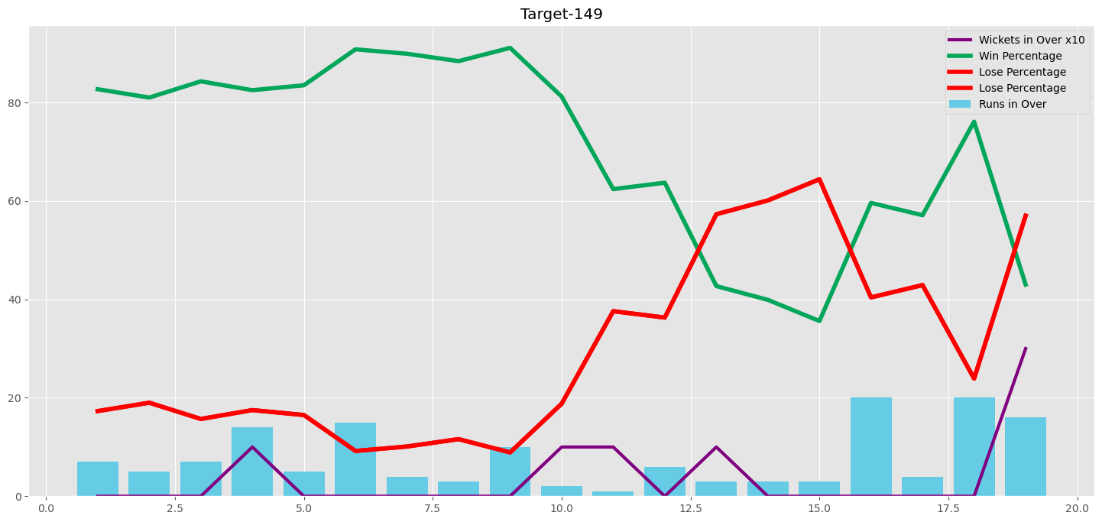
As mentioned earlier, there are different types of bowlers and batsmen in cricket which presents various challenges for batsmen, and how to use these different bowlers is a large part of the game tactically. The distribution of right-handed batsmen to left-handed batsmen is 67% to 33%, with the distribution of runs scored being 66.6% to 33.4%. Since these are both similar, we know that right-handed and left-handed batsmen must score runs at a similar pace with strike rates (how many runs are scored every 100 balls) of 130 and 132 respectively. So there is no significant difference between the performance of right-handed and left-handed batsmen. Over the entire IPL, fast-paced and medium-paced bowlers have been used most commonly with off spin, leg spin, and orthodox spin bowlers being used about 10 – 12% of the deliveries as well. Unorthodox spin is the least common bowling type with only 0.5% of deliveries being bowled. Over the entire IPL, leg spin bowlers tend to be the best-performing bowlers since they have the best bowling average (number of runs conceded per wicket) and one of the best economies (number of runs conceded per over. A more important analysis is to see how these different bowlers are used and perform in different stages of the game. We can split the game into three parts – the power play (overs 1 to 6), the middle of the game (overs 7 to 14), and the end of the game (overs 15 to 20). The distribution of how bowlers are used is below:

As we can see, fast and medium-paced bowlers are used more prominently at the beginning and end of the game, whereas the spinners bowl the most during the middle of the game. Comparing the bowling average and economy in different stages of the game explains why bowlers are used when they are. Firstly, fast and medium-paced bowlers are the only type of bowlers that have a better economy in the power play than in the middle of the game. This suggests that teams have realized that fast and medium-paced bowlers will have their most effective overs in the first 6 overs. Additionally, spin bowlers have a much lower economy in the middle of the game which may explain why they are used the most during this stage of the game. One important finding from the data is that leg spinners and off-spinners seemed to be underused at the end of the game. These bowlers have the best economy and average at this stage of the game yet are used much less than fast and medium-paced bowlers. By having a greater economy and average, leg-spin and off-spin bowlers are the most effective bowlers at this point of the game since they have a higher tendency to take wickets and teams seem to struggle to score off them the most.

The final part of the game that will be analyzed is the team's decision at the toss. At the beginning of each game, both teams meet for the toss and the winner chooses to either bat or bowl first. In general, teams tend to choose to field first more often (67.5% of games) with some seasons seeing teams choose to field first in over 90% of matches (2016). This implies that teams believe that it is easier to chase a target than defend a target, but does this decision often pay off? The answer is typically yes, with more teams having a greater likelihood of winning when choosing to field first instead of batting first.

**Winning Prediction Model:**

The first step in making the model was to create an appropriate dataset through feature engineering. The initial part of the feature engineering was to combine data from the game-by-game datasets with the ball-by-ball datasets. Some of these variables included the city, venue, and winning team. The next stage of the feature engineering was to only take the data from the second innings and remove games that were rain-affected or shortened for other reasons since this is which innings the win predictor will be used for. Once this data frame was created, several new variables were generated to be used in the winning prediction model. These variables were the current run rate, required run rate, wickets remaining, current score, runs left to score (to win), and balls left in the innings. The final data frame was taken and includes the variables batting team, bowling team, city, runs left, balls left, wickets left, runs, bowler type, current run rate, required run rate, and result.

For this model, a pipeline was used with the first step using a Column Transformer to transform the values of the categorical variables (city, batting team, bowling team, and bowler type) into binary dummy variables that are suitable for modeling. This transformation was done using the One Hot Encoder function. The remainder of the variables would then pass through to the modeling stage. Three different models were initially trained and tested for the model. These were a logistic regression model, a random forest classification model, and a decision tree classification model. All three models had very high accuracy scores with the logistic regression being over 80% and the other two models had accuracy scores of 99%. However, the decision tree and random forest classification models seemed to overfit the data when they were used to model win prediction such that it would only say 100% or 0% win prediction. Additionally, these two models were not able to generate a loss prediction. Therefore, a logistic regression was used. The way the model interpreted the data was to look at the given information after each ball and would be trained to understand whether that situation was favorable toward a team winning the game or not. The model was made to analyze the win prediction of any game in the IPL data set based on its unique ID. An example of the win prediction model is below which is the 2019 IPL final between Chennai Super Kings and the Mumbai Indians.

As the graph shows, there are significant changes in win percentage based on whether a large number of wickets were lost, or runs were either scored quickly or limited. For example, at the end of the game going into the 19th over, Chennai scored many runs in the previous over increasing their likelihood of winning. But in the 19th over, despite scoring runs, they still lost three wickets which the model predicted to be more detrimental. In the end, this was accurate as Chennai would not be able to win the game in the last over. This indicates that the model has picked up trends in the data about how changes in the game may impact the outcome.

Ideally, I would make this model such that you could draw from another spreadsheet and add your ball-by-ball data so you could track whether different choices in a game could change the result, but unfortunately, I ran out of time to do so. The main limitation of this model is the lack of qualitative data for bowlers and fielders. A model that could base predictions on the strength of the batsmen at the crease or the quality of the bowler bowling the next over would greatly improve the accuracy and precision of the model.