Data Analysis of the Indian Premier League

Sahil Khan
Winter in Data Science, Analytics Club
IIT Bombay
Mumbai, India
200020112@iitb.ac.in

I. Introduction

Sports analytics is a field that is becoming widely popular due to the competitive edge that it can give both to sports teams as well as stakeholders involved in the sport. Various data which is available such as the players and team statistics, environment conditions, etc is made use of to predictive models which can help stakeholders make informed decisions on the game. The main objective is to improve the performance of the team and assist in creating strategies which would help the team perfectly counter its opponents. This can be done both prior to a game as well as dynamically as the game progresses. In recent times, it has been observed that the audience themselves are also interested in the data analysis that goes on in the game and hence, sports analysts try to present this data to the audience by making simplifications to it and making use of pictorial elements such as graphs and charts to capture their attention.

A. About Cricket

Cricket is a sport that is played by two teams, each having eleven members. A team consists of batsmen, bowlers, and all rounders. The role of the batsmen is to score as many runs as possible in the limited time/overs available, while the bowlers try to restrict the score that the batsmen try to make. All rounders are players that play both roles and have sufficient expertise in both batting and bowling. The performance of a team depends on various factors such as the constitution of the team in terms of types of players, the venue in which the match is being held, the environmental conditions, and the type of opponents that they're playing against. Data analytics can be made use of to help the teams management figure out which players to play in a specific match, the odds of them reaching a specific stage in a tournament, the environmental conditions that they're going to play in, etc. It can also be used during a match to help the team adjust their strategy according the state at which the match is in, to provide them a competitive edge against their opponent. These days, data science techniques are being made use of by every team that competes in the sport professionally. When used correctly, it can help teams bridge the gap in skill by formulating an effective strategy to counter their opponents.

B. About the Indian Premier League

The Indian Premier League (IPL) is the worlds biggest domestic cricket tournament. It is a 20-over format of the game that makes for short, fast-paced games which is one of the reasons for its massive fanbase. It is an annual tournament and has seen 13 such tournaments conducted so far. There are 8 teams involved in the tournament and the teams themselves consist of players from all around the world. The tournament generates a large revenue and has many stakeholders heavily invested in it. So teams will do everything they can to get an edge over their opponents in a game. Data Analysis is now heavily used by all teams to try and gain this edge.

II. Dataset

The datasets used for analysis and prediction were collected from www.kaggle.com [7], where the data of all editions of the IPL so far was available. Two datasets have been used. One for overall matches data and one for ball-to-ball data for the full 2008-2019 period. Both the datasets are linked by the 'id' column which represents the matches uniquely. Some of the useful features present in the dataset are date of match, venue, run(s) and wicket(if any) on every ball, toss decision, batsman and bowler, result of match with margin etc. There are some minor discrepancies in data such as missing values in 'bowling team' column and duplicate team name but it doesn't hurt the predictions task as team data is also present in 'team1','team2' columns. The dataset consists of 2 lakh data points with 18 features in total.

III. Analysis Pipeline

As observed from the literature survey conducted, a large majority of the predictive models that were made are used to predict the outcome of the match and this prediction is made before the start of the match. This prediction will be useful for the team to make long-term decisions for the team to perform better in the tournament as a whole but is not very useful during the match itself as no changes can be made to the team in the middle of a match. The work discussed in this paper seeks to fill in this gap by providing data to the team at various phases of the match so that the team can make informed decisions such as what batting order and bowling order to use for the rest of the game.

Firstly, an exploratory analysis of the data is conducted to get a better understanding of what parameters affect the performance of the team as a whole as well as the individual contributions of the players.

A. *Interesting Insights drawn from the datasets*

Various manipulations are performed on the available datasets to extract some insightful information from them.

1) Teams with max no. of wins: As seen in the Fig. 1, Mumbai Indian has achieved most wins in 4 season (2010, 2013, 2017, 2019) while Chennai Super Kings achieved most wins in 3 season (2011, 2015, 2018). Hence we can observe the dominance of different teams over the seasons.

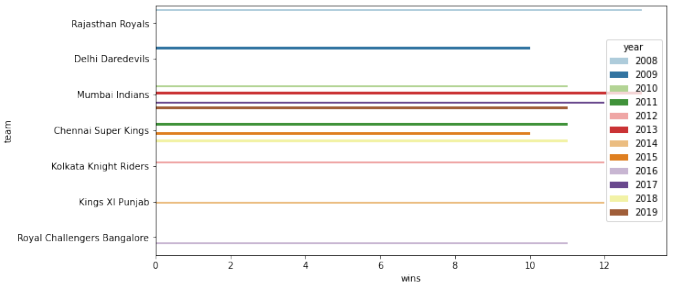

Fig. 1. Teams with max no. of wins in a season

2) Stadiums hosting IPL matches: In Fig. 6, we can see that Eden Gardens hosted the most no. of IPL matches (77), the stadium with second, third positions here are Wankhede Stadium (73 matches) and M Chinnaswamy Stadium (71 matches) respectively.

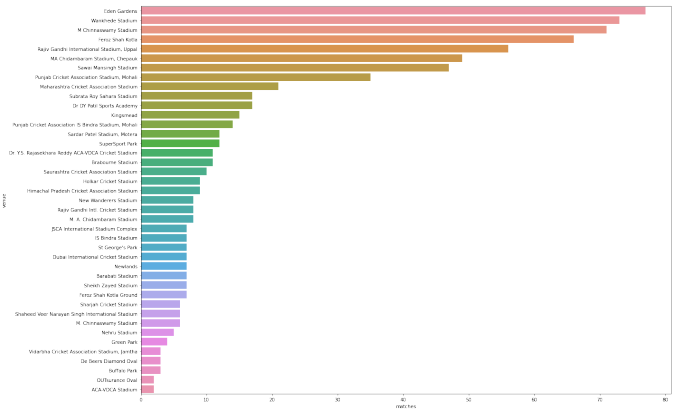

Fig. 2. Number of matches hosted in different Stadiums

3) Most successful teams: In Fig. 3, we can see most wins by all teams ever competed in IPL. As being the most crowned team of IPL, Mumbai Indians has the most number of wins as expected. What is interesting to note is that despite not competing in 2 full tournaments (due to ban), CSK are just 14 wins away from MI in terms of total wins, a proof why MI-CSK rivalry is most popular in IPL.

Fig. 4 show the percentage wins of teams, Delhi capitals has highest success rate followed by Chennai Super Kings and than Mumbai Indians. But the thing to keep in mind is that Delhi Capitals played less no. of matches and has won 10 till now.

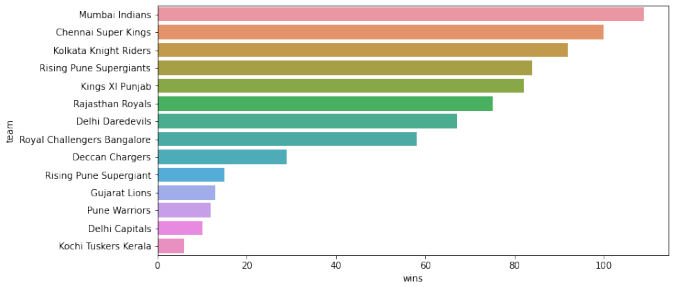

Fig. 3. Total wins by all teams till 2019

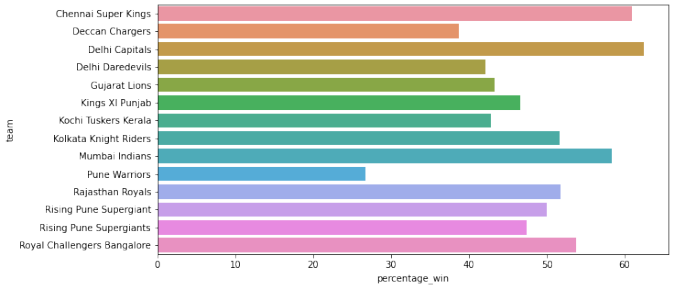

Fig. 4. Percentage wins by all teams till 2019

4) Most Influential Players: Using Man of the Match (MoM) data available in the dataset, we can extract players with most MoM awards in IPL. Fig. 5 shows us that despite RCB not winning no IPL yet, three of it's players feature in top 10 list with CH Gayle and AB de Villiers at number 1 and 2 respectively. This highlights the incompetence of RCB's bowling which hampers their title-winning chances.

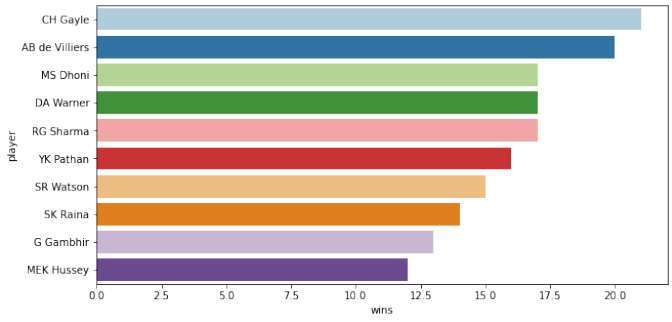

Fig. 5. Players with most Man of the Match awards till 2019

5) Toss winnings: From Fig. 6 & Fig. 7, it can be estimated that Mumbai Indian has won the most no. of tosses, but percentage of toss winning is 53%. While Delhi Capitals has 62% of toss winning rate, But it has played only 16 matches, out of which 10 tosses were won.
6) Greatest victories: Fig. 8 & Fig. 9 represents wins by run and wickets respectively. More over the 1st team in Y-axis is the winner team. So one of the greatest victories was when Mumbai Indians & Delhi Daredevils competed and MI won by 146 runs.

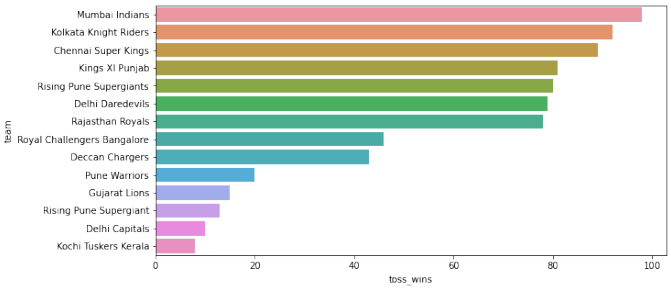

Fig. 6. Number of tosses won by teams

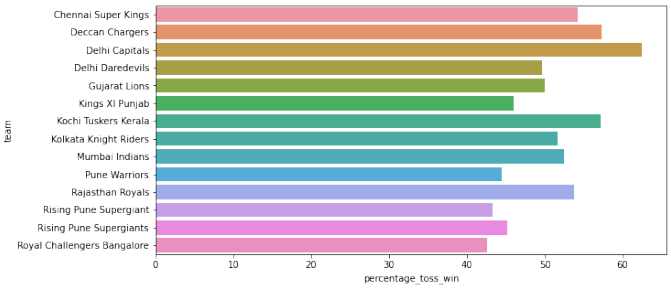

Fig. 7. Percentage of tosses won by teams

7) Most 50s and 100s scored: The number of centuries and half centuries made by player shows his consistency and a better chance to win the match. This information is represented in Fig. 10 & Fig. 11.
8) Comparison of Teams over seasons: Bi-histogram plots will give us an estimate of how the two teams perform against each other over the years. For example in Fig. 12, comparison between MI & KKR is shown.

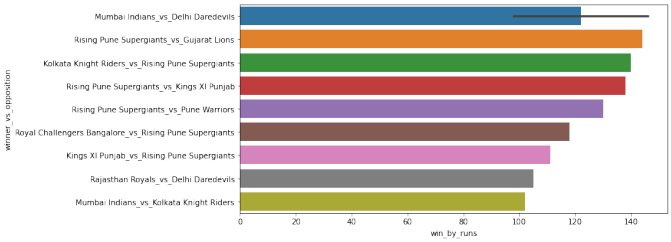

Fig. 8. Win by runs

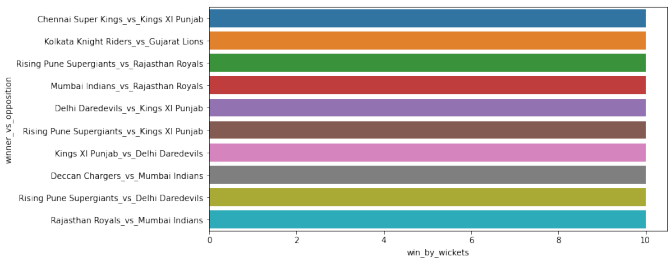

Fig. 9. win by wickets

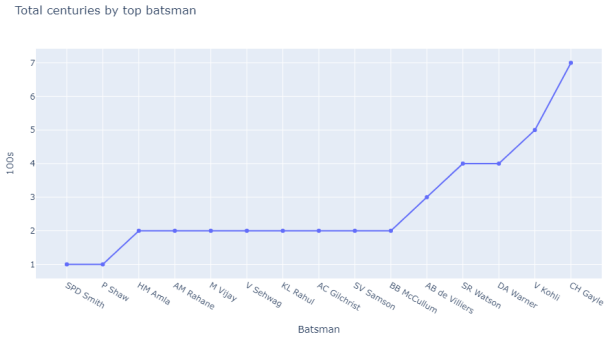

Fig. 10. Full centuries made by Top Batsmen

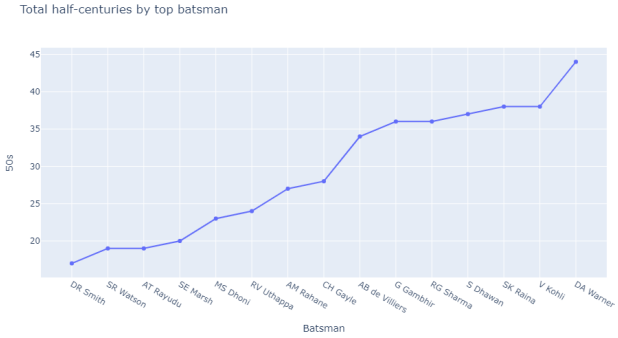

Fig. 11. Half centuries made by Top Batsmen

IV. Winner Prediction

In this section, we train a ML model for this Classification task to predict the winner of the match, using previously available data from past years that we analysed in *Analysis Pipeline* section.

A. Data Cleaning and Processing

Following steps performed before training the final model:

1) Data Pre-processing: Removed Nan values in any column in the dataset. Converted All features to numerical values by utilising label encoding and pd.dummies to

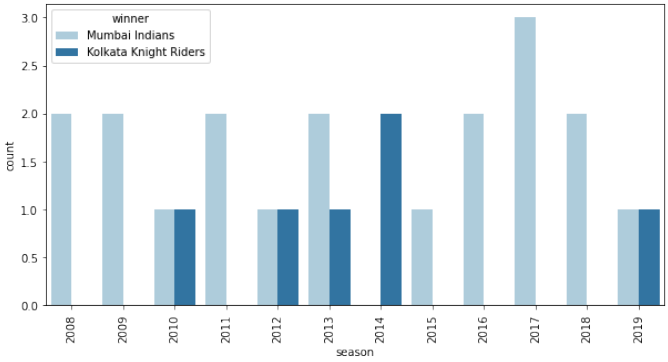

Fig. 12. Wins achieved when played against each other

convert categorical data into numerical data. Finally, brought all features to the same scale by MinMaxScaler and StandardScaler.

2) Features Engineering : Performed correlation analysis to find similar features, and then removed similar features from the training data to let the ML model run faster. Then made new features which would be useful to predict the target. For example, to predict the final score, we can engineer features like current score, current wickets, current batsmen runs, current bowler's wickets, runs scored in the last 5 overs, wickets taken in the last 2 overs etc.

B. Prediction and Results

For this classification task, we leveraged different machine learning and deep learning models and got a sequence of results as follow:

1) ML models: We used Logistic Regressoin, Support vector machine, Decision Tree, and Random Forest models. Fig 13 show different accuracy level in the form of F1 score. So accuracy of Decisoin Tree and Random Forest are nearly equal, so compairing as per confusion matrix, best suitable model for this task is Random Forest Classifier.

model	f1_score
Logistic Regression	0.503311
SVM	0.523179
Decision Tree Classifier	0.543046
Random Forest Classifier	0.543046

Fig. 13. Summary of results obtained from different models

2) Nueral Netowrks(NNs): Neural networks are the modern times way-to-go prediction models. We modelled a network with 2 hidden layers with 32 units each.
results: After 200 epochs, the results obtained are as follows: Train Loss: 0.6784 — Val Loss: 0.6854 — Train Accuracy: 56.5789 — Val Accuracy: 56.6667%. And on the Final test set accuracy come out to be 63.1578%, which is not usable for practical purposes. Limited amount of data is the primary reason why neural networks are not giving decent results even after training for around 1 hour. But, this is not entirely useless as can be seen in the training graphs as shown in Fig. 14 15.

V. Final Score Prediction

In this section, we train ML models, Neural Network to forecast the final score of the team which is presently batting, based on the current score of the team.

A. Data Cleaning and Processing

Data Cleaning and Pre-processing including feature engineering was performed for this dataset also.

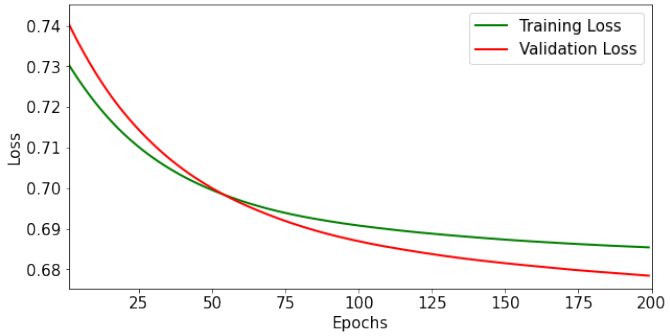

Fig. 14. Training and Validation losses for classification Task

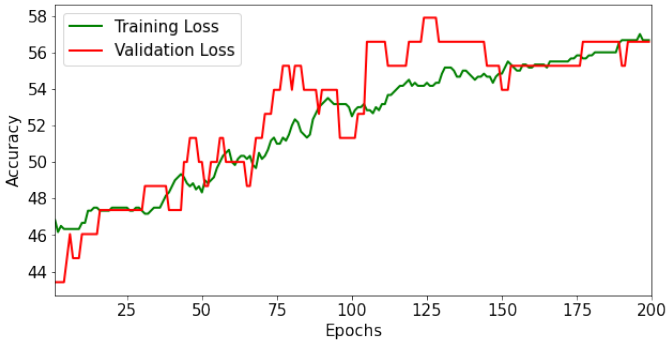

Fig. 15. Training and Validation Accuracy for Classification Task

The exact features which are used for prediction are listed as follows:

- Current runs
- Current wickets
- Runs on every ball
- Overs completed
- Batsman runs

B. Prediction and Results

Mean squared error is used as a loss function for back-propagation. R2 score and custom accuracy (predicted score being in margin of 10 of actual final score) is used for evaluating the model.

1) ML Models: We used Linear Regression, Random Forest and Linear SVR. Fig. 16 show custom accuracy & Mean squared error values for different models. Random Forest give the largest accuracy of 76.1859% among these three models.

models	mean squared error	custom accuracy
Linear Regression	522.667463	41.727084
Random Forest	156.065917	76.185914
Linear SVR	525.211956	42.076644

Fig. 16. Summary of results obtained from different models

2) Neural Network: For the architecture of this Neural Net, we constructed 6 hidden layers, more info is provided in Fig. 17. So the model has 100,609 trainable parameters.

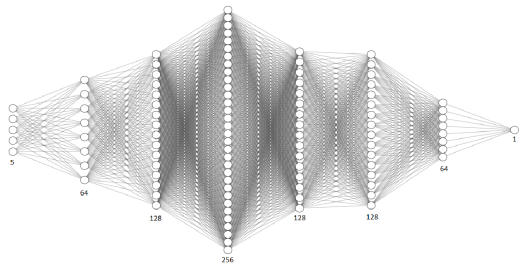

Fig. 17. Neural network architecture with hidden layer sizes

After 200 epochs, the results obtained are as follows: Train Loss: 0.0111 — Val Loss: 0.0111 — Train R^2: -0.5650 — Val R^2: -0.5825. And on the Final test set accuracy come out to be 53.5456%, which is not usable for practical purposes. Limited amount of data is the primary reason why neural networks are not giving decent results even after training for around 1 hour. But, this is not entirely useless as can be seen in the training graphs as shown in Fig. 18 19.

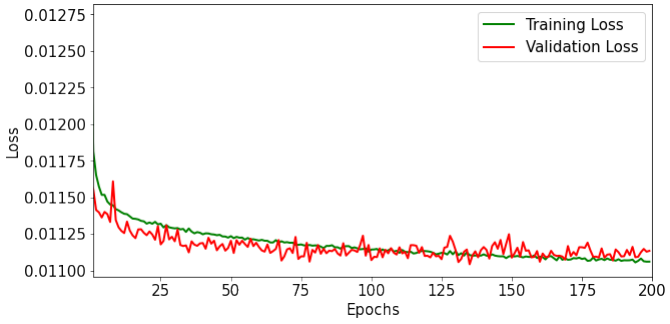

Fig. 18. Training and Validation losses for classification Task

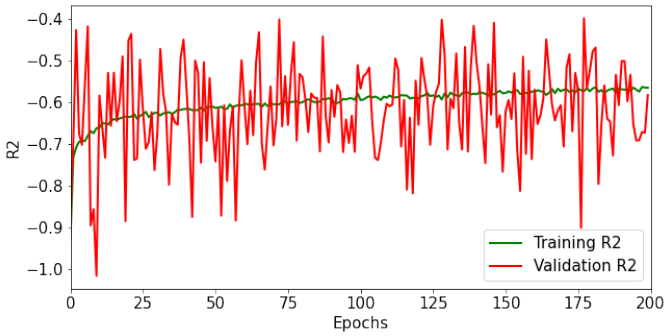

Fig. 19. R^2 value for Training & Validation sets

VI. Conclusion and Future Scope

This paper provides useful insights from IPL dataset about what are the best performing teams and players. Toss decisions and their importance in winning matches prove the overall winning toss has more or less no influence on winning chances. Best performing players of IPL can be listed with the most MoM awards analysis. Sponsors can focus on which cities host the IPL matches most to analyze the audience in those areas specifically and make their plans accordingly. The prediction of final score at any given moment of match is currently done with the help of Current Run Rate(CRR), while it is one of the useful features, it doesn't take into account what are the remaining overs and scores of the batsmen at crease. The models proposed in the work take these features into account to predict the final score given these features at any point in the game. Due to limited data, the best model is 76% accurate on an error margin of ±10 runs. Future Work can be pre-training the neural network models on an ODI or T20 international datasets and then fine tuning them for ipl predictions as direct training with datasets is not possible due to different formats and playing conditions.

References

[1] H. Barot, A. Kothari, P. Bide, B. Ahir and R. Kankaria, "Analysis and Prediction for the Indian Premier League," 2020 International Conference for Emerging Technology (INCET), Belgaum, India, 2020, pp. 1-7, doi: 10.1109/INCET49848.2020.9153972.

[2] Passi, Kalpdrum Pandey, Niravkumar. (2018). Increased Prediction Accuracy in the Game of Cricket Using Machine Learning. International Journal of Data Mining Knowledge Management Process. 8. 19-36. 10.5121/ijdkp.2018.8203.

[3] Lamsal, Rabindra Choudhary, Ayesha. (2018). Predicting Outcome of Indian Premier League (IPL) Matches Using Machine Learning.

[4] Deep Prakash, Chellapilla Patvardhan, C. Vasantha, C.. (2016). Data Analytics based Deep Mayo Predictor for IPL-9. International Journal of Computer Applications. 152. 6-11. 10.5120/ijca2016911875.

[5] Priyanka, Sachi. (2020). Prediction of Indian Premier League-IPL 2020 using Data Mining Algorithms. International Journal for Research in Applied Science and Engineering Technology. 8. 790-795. 10.22214/ijraset.2020.2121.

[6] Thenmozhi, D. Palaniappan, Mirualini Sakthi, S.M.Jai Vasudevan, Srivatsan Kannan, V Sadiq, S. (2019). MoneyBall - Data Mining on Cricket Dataset. 1-5. 10.1109/ICCIDS.2019.8862065.

[7] @miscWinNT, author = Prateek Bhardwaj, title = IPL Complete Dataset (2008-2020), year = 2020, url = https://www.kaggle.com/patrickb1912/ipl-complete-dataset-20082020, urldate = 2020-11-23

[8] Cricsheet, url = https://cricsheet.org/downloads/, urldate = 2020-11-30

[9] Exploratory Data Analysis of IPL Matches-Part I, url = https://towardsdatascience.com/exploratory-data-analysis-of-ipl-matches-part-1-c3555b15edbb, urldate = 2019-10-16

[10] Predictive Analysis of an IPL Match, url = https://towardsdatascience.com/predicting-ipl-match-winner-fc9e89f583ce, urldate = 2020-03-06