American Express Campus Challenge 2024

Modeling Problem Statement

Predict the Winner of the T20 Cricket Match! 

Background

- Sports analytics has revolutionized the way we analyze sporting events and has generated good buzz
- Given the cricket season, we at American Express are back with an exhilarating T20 Match Prediction Challenge, as a part of **American Express Campus Challenge.**
- We hope all the students take part in this exciting competition and put their cricketing and ML knowledge to task

Problem Statement

Build the best ML model using boosting algorithms to accurately predict “Winning Team” for a T20 match.

Match ID	Team 1	Team 2	Winner Team	Predicted Winner	
9331048	12634	14860	12634	12634	✓
9076815	44904	42573	42573	44904	✗
…	…		….		
9516632	30393	48334	30393	48334	✗
9388483	17583	17744	17744	17744	✓

- T20 games from past two years, across domestic & international tournaments would be provided to be used for training
- Additionally, detailed batsmen & bowler scorecard for each match would also be provided
- Some ready-to-use independent features will also be provided

Evaluation Criteria –

Accuracy = Number of correct predictions / Total number of games to predict

*Evaluation code to be provided to participants

Data Details

Datasets

Data	Dataset type	Desc	# Rows	# Columns	Time period	Location/Name
Train Data	Primary	Train data with game level information	948	23	'22-'23	*To be downloaded from the unstop website (available to all registered candidates)*
Match level data	Additional	All games scorecard without r1 & r2	1689	30	'21-'23	
Batsman level data	Additional	batsman level scorecard w/o r1 & r2	24483	21	'21-'23	
Bowler level data	Additional	bowler level scorecard w/o r1 & r2	18539	18	'21-'23	
Round 1 Submission Data (r1)	Primary	Round1 data with game level information w/o winner	271	21	'22-'23	
Round 2 Submission Data (r2)	Primary	Round2 data with game level information w/o winner	TBD*	21	TBD*	*To be downloaded from the website (available to R1 shortlisted candidates)*

- All match level dataset are unique on match id, batsman level scorecard is unique at match id X batsman_id, bowler level scorecard is unique at match id X bowler_id.
- Additional dataset also include games from 2021 to make features.
- Datasets are delimited by ','
- Please note that *winner* & *winner_id* will not be provided with Round 1 & 2 Submission data
- Additional data to be used for feature creation only.

*To be shared later

Sample of Primary modeling dataset

Train data with important columns

Primary Key

match id	team1_id	team2_id	winner_id	...	ground_id	team_count_50runs_last15	...	...	...	...
9331048	12634	14860	12634	...	7398	0.272	...	...	...	...
9076815	44904	42573	42573	...	14188	0.6	...	...	...	...
...	...		....	...			...	...	...	...
9516632	30393	48334	30393	...	23498	1.125	...	...	...	...
9388483	17583	17744	17744	...	5081	0.6428	...	...	...	...

*winner & winner_id will not be present in round1 & round2 evaluation data.

Variable Name	Description
match id	unique id of a game
team1_id	unique id of team1
team2_id	unique id of team2
winner_id	unique id of winner team
team1	team1 name (masked)
team2	team2 name (masked)
winner	winner team name (masked)
team1_roster_ids	':' separated team1 player ids
team2_roster_ids	':' separated team1 player ids
toss winner	toss winner team name (masked)
toss decision	toss decision - field or bat
venue	Stadium name
city	city the match is held at
match_dt	match date
lighting	lighting condition of match - day/night, day or night match
series_name	name of the series being played. For e.g., different major leagues, International T20s etc.
season	seasons of the series. For e.g., 2021/2022, 2022, etc.
ground_id	unique id of the ground the match is held at
team_count_50runs_last15	Ratio of number of 50s by players in team1 to number of 50s by players in team2 in last 15 games
team_winp_last5	Ratio of team1's win % to team2's win % in last 5 games
team1only_avg_runs_last15	team1's avg inning runs in last 15 games
team1_winp_team2_last15	Team1's win percentage against Team2 in last 15 games
ground_avg_runs_last15	average runs scored in the ground in last 15 games

Sample of Batsman dataset

Batsman level scorecard data with important columns

Primary Key

match id	batsman_id	inning	...	...	runs	balls faced	...	...
8638034	3776849	1	...	...	18	13	...	...
8638034	6718326	2	...	...	91	50	...	...
...	....	...	...	...	...	...	...	...
8587837	7620283	1	...	...	10	10	...	...
8587837	87191	2	...	...	19	17	...	...

Variable Name	Description
match id	Unique id of match
batsman_id	Unique player id of the batsman
inning	Inning order – 1st or 2nd.
batsman	Batsman name (masked)
batsman_details	':' separated fields for the batsman - <Nationality>:<Batting style>:<Bowling style>. For e.g., IND:Right-hand bat: Right-arm medium
is_batsman_captain	0/1 field for is batsman captain
is_batsman_keeper	0/1 field for is batsman keeper
runs	Runs scored by the batsman in the inning.
balls_faced	Balls faced by batsman in the inning.
over_faced_first	First over.delivery faced by the batsman.
wicket kind	Kind of dismissal of the batsman.
out_by_bowler	Name of the bowler dimissing the batsman (masked).
out_by_fielder	Name of the fielders assisting in the dismissal (masked).
bowler_id	Unique player id of the bowler.
bowler_details	':' separated fields for the bowler- <Nationality>:<Batting style>:<Bowling style>. For e.g., IND:Right-hand bat: Right-arm medium.
is_bowler_captain	0/1 field for is bowler captain.
is_bowler_keeper	0/1 field for is bowler keeper.
strike_rate	Strike rate of the batsman in the inning.
Fours	Number of Fours scored by the batsman in the inning.
Sixes	Number of Sixes scored by the batsman in the inning.
match_dt	Match date

Sample of Bowler dataset

Bowler level scorecard data with important columns

Primary Key

match id	bowler_id	inning	...	...	runs	wicket_count	...	...
8638034	4950294	1	...	...	21	1	...	...
8638034	3834305	2	...	...	7	0	...	...
...	....	...	...	...	...	...	...	...
8587837	3890984	1	...	...	33	4	...	...
8587837	34061	2	...	...	35	2	...	...

Variable Name	Description
match id	Unique id of a match
bowler_id	Bowler unique player id
inning	inning order - 1st or 2nd
bowler	Name of the bowler (masked)
bowler_details	':' separated fields for the batsman - <Nationality>:<Batting style>:<Bowling style>. For e.g., IND:Right-hand bat: Right-arm medium
is_bowler_captain	0/1 field for is bowler captain
is_bowler_keeper	0/1 field for is bowler keeper
runs	Runs conceded by the bowler
wicket_count	Wickets taken by the bowler
balls_bowled	Number of balls bowled by the bowler
economy	Economy of the bowler - ratio of runs conceded and balls bowled
maiden	Number of maiden overs (overs with 0 runs conceded) bowled by the bowler
dots	Number of dot balls (balls with 0 runs conceded) bowled by the bowler
Fours	Number of Fours conceded by the Bowler
Sixes	Number of Sixes conceded by the Bowler
wides	Number of wides bowled by the bowler
noballs	Number of no-balls bowled by the bowler
match_dt	match date

Sample of Match level dataset

Variable Name	Description
match id	unique id of a game
team1_id	unique id of team1
team2_id	unique id of team2
winner_id	unique id of winner team
team1	team1 name (masked)
team2	team2 name (masked)
winner	winner team name (masked)
team1_roster_ids	':' separated team1 player ids
team2_roster_ids	':' separated team1 player ids
toss winner	toss winner team name (masked)
toss decision	toss decision - field or bat
venue	Stadium name
city	city the match is held at
match_dt	match date
lighting	lighting condition of match - day/night, day or night match
series_name	name of the series being played. For e.g., different major leagues, International T20s etc.
season	seasons of the series. For e.g., 2021/2022, 2022, etc.
ground_id	unique id of the ground the match is held at
by	mode of victory - wickets or runs
win amount	margin of victory
player_of_match_id	Id of Man of the match
umpire1	name of first umpire
umpire2	name of second umpire

Match level scorecard data with important columns

Primary Key

match id	team1_id	team2_id	...	by	...	inning1_runs	inning1_wickets	...	...
8588005	17982	18570	...	runs	...	148	6	...	...
8587907	33914	33956	...	wickets	...	150	4	...	...
...	...	...	...	...	...	...	...	...	...
8587816	33935	33956	...	runs	...	183	4	...	...
8638083	17653	17583	...	runs	...	154	6	...	...

Variable Name	Description
inning1_runs	runs scored by batting first team in first inning
inning1_wickets	wickets taken by bowing first team in first inning
inning1_balls	balls faced by batting first team in first inning
inning2_runs	runs scored by batting second team in second inning
inning2_wickets	wickets taken by bowing second team in second inning
inning2_balls	balls faced by batting second team in second inning
series_type	type of series the match belongs too. It can have one of the three values: International, other_domestic, In Pr Le.

Ready to use features provided - Sample example vars

Features	Feature level	Description
team_count_50runs_last15	Team level (aggregated from players)	Ratio of number of 50s by players in team1 to number of 50s by players in team2 in last 15 games
team_winp_last5	Team level	Ratio of team1's win % to team2's win % in last 5 games
team1only_avg_runs_last15	Team level	team1's avg inning runs in last 15 games
team1_winp_team2_last15	Team1 x Team2 level	Team1's win percentage against Team2 in last 15 games
ground_avg_runs_last15	Ground level	average runs scored in the ground in last 15 games

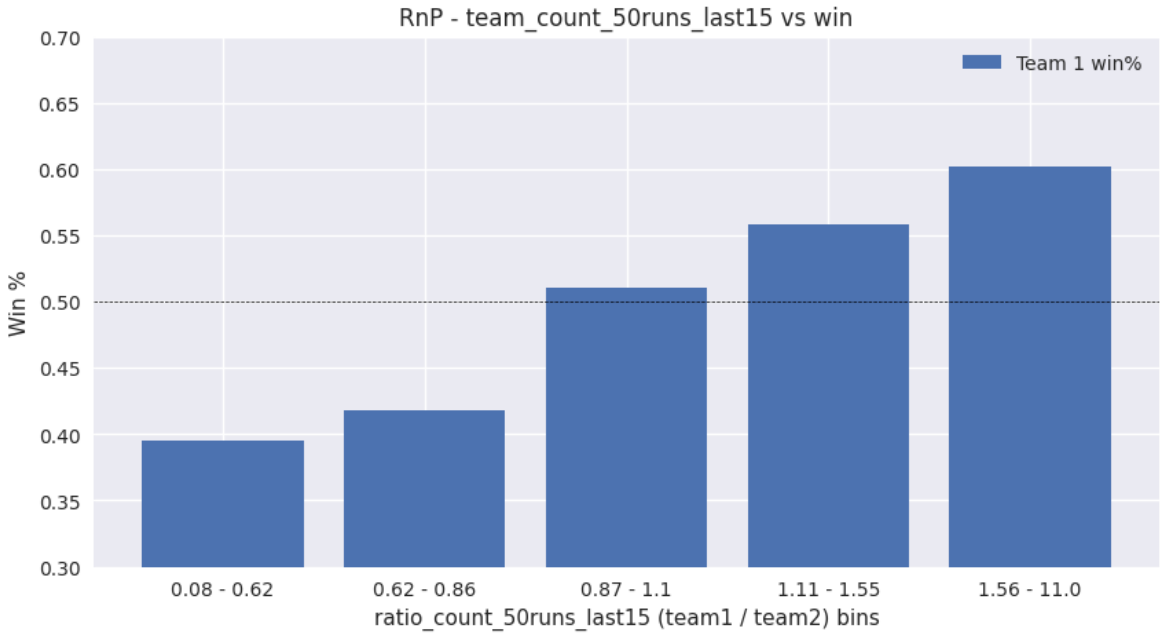
Logic to *team_count_50runs_last15* for a game.

- For both teams, take last 15 games played by players present in the roster for that game.
- Sum up the number of 50s scored in last 15 games across players.
- Take ratio of team1's sum to team2's sum.

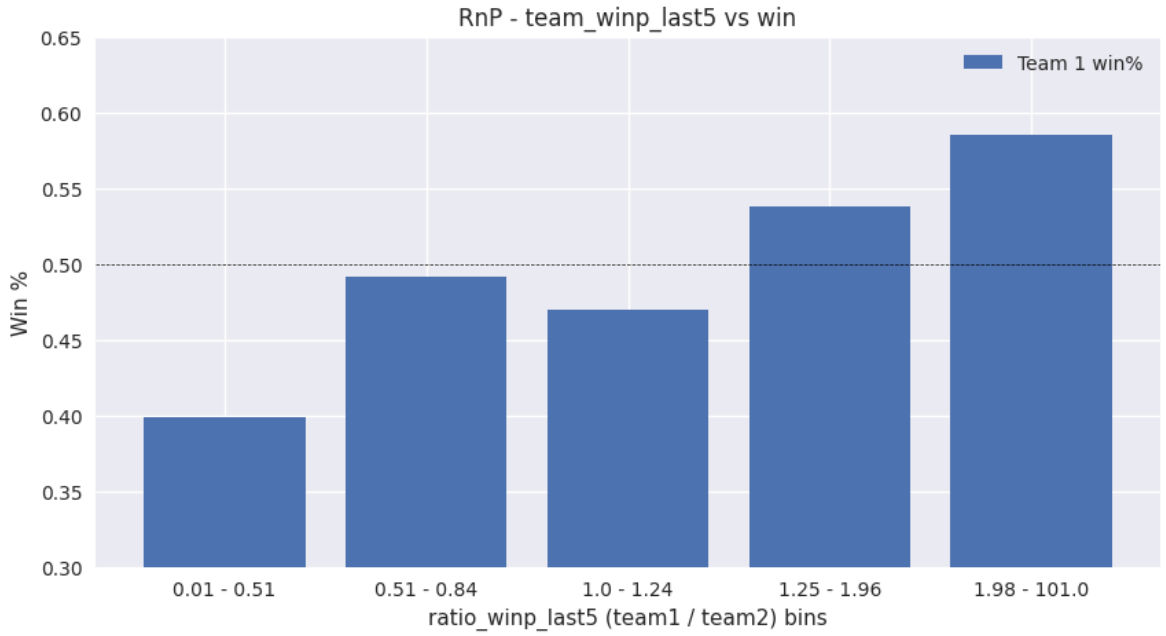
Logic to *team_winp_last5*

- For both teams, take last 5 games played by each team.
- Take the ratio of team1's win% to team2' win%.

Ready to use features trend plot



Plot [I]



Plot [II]

- Rank and Plots show trend of the variable with dependent variable.
- X axis represents range of values taken by the bin. Y axis represents win % of Team1.
- As the count of 50runs scored by players in Team1 is more than count of 50runs scored by players in Team2, their ratio is greater than 1, and we expect Team1 to win more often. This is evident from plot [I] as the bin value increases as feature value increases.
- Similarly, as the win % of Team1 in last 5 games is more than win % of Team2 in last 5 games, we expect Team1 to win more often. This is evident from plot [II] as the bin value increases as feature value increases.

Sample of submission - Sample 1

File 1

	Fixed Variables									Dynamic Variables (Prefix can be fixed indep_feat..) -Top 10 features		
	Amex will provide		Participants will add							Participants will add		
	match id	dataset_type	win_pred_team_id	win_pred_score (for win_pred_team_id)	train_algorithm	Ensemble? (if yes, then comma separated train_algo)	train_hps_trees	train_hps_depth	train_hps_lr	indep_feat_id1	....	indep_feat_id10
Evaluation Data Scores (271 matches)	9250275	r1	5143	0.905	xgboost	no	100	8	0.1	...	...	...
	9262189	r1	6437	0.765	xgboost	no	100	8	0.1	...	...	...
	9128776	r1	2	0.345	xgboost	no	100	8	0.1	...	...	...
	9586919	r1	1992	0.5	xgboost	no	100	8	0.1	...	...	...
	9128538	r1	1409	0.031	xgboost	no	100	8	0.1	...	...	...
	...	...	...	...	xgboost	no	100	8	0.1	...	...	...
	...	...	...	...	xgboost	no	100	8	0.1	...	...	...
	...	...	...	...	xgboost	no	100	8	0.1	...	...	...
	...	...	...	...	xgboost	no	100	8	0.1	...	...	...
Train Scores (948 matches)	9331181	train	4340	0.44	xgboost	no	100	8	0.1	...	...	...
	8797060	train	5159	0.665	xgboost	no	100	8	0.1	...	...	...
	9433269	train	1479	0.499	xgboost	no	100	8	0.1	...	...	...
	...	train	...	...	xgboost	no	100	8	0.1	...	...	...

File 2

Participants will add				
feat_id	feat_name	feat_description	model_feat_imp_train	feat_rank_train
1	avg_runs_last10	Avg. Runs scored by Team 2 in last 20 matches	15%	1
2	...	...	...	...
3	...	...	...	...
4	...	...	...	...
5	...	...	...	...

*Submission template for file1, file2 can be downloaded from unstop website.

Sample of submission - Sample 2 (Ensemble model)

File 1

	Fixed Variables									Dynamic Variables (Prefix can be fixed indep_feat..) -Top 10 features		
	Amex will provide		Participants will add							Participants will add		
	match id	dataset_type	win_pred_team_id	win_pred_score (for win_pred_team_id)	train_algorithm	Ensemble? (if yes, then comma separated train_algo)	train_hps_trees	train_hps_depth	train_hps_lr	indep_feat_id1	....	indep_feat_id10
Evaluation Data Scores (271 matches)	9250275	r1	5143	0.905	xgboost;lightgbm	yes	100;200	8;6	0.1;0.2	...	...	...
	9262189	r1	6437	0.765	xgboost;lightgbm	yes	100;200	8;6	0.1;0.2	...	...	...
	9128776	r1	2	0.345	xgboost;lightgbm	yes	100;200	8;6	0.1;0.2	...	...	...
	9586919	r1	1992	0.5	xgboost;lightgbm	yes	100;200	8;6	0.1;0.2	...	...	...
	9128538	r1	1409	0.031	xgboost;lightgbm	yes	100;200	8;6	0.1;0.2	...	...	...
	...	...	...	...	xgboost;lightgbm	yes	100;200	8;6	0.1;0.2	...	...	...
	...	...	...	...	xgboost;lightgbm	yes	100;200	8;6	0.1;0.2	...	...	...
	...	...	...	...	xgboost;lightgbm	yes	100;200	8;6	0.1;0.2	...	...	...
	...	...	...	...	xgboost;lightgbm	yes	100;200	8;6	0.1;0.2	...	...	...
Train Scores (948 matches)	9331181	train	4340	0.44	xgboost;lightgbm	yes	100;200	8;6	0.1;0.2	...	...	...
	8797060	train	5159	0.665	xgboost;lightgbm	yes	100;200	8;6	0.1;0.2	...	...	...
	9433269	train	1479	0.499	xgboost;lightgbm	yes	100;200	8;6	0.1;0.2	...	...	...
	...	train	...	...	xgboost;lightgbm	yes	100;200	8;6	0.1;0.2	...	...	...

File 2

Participants will add (Max – Top 100)				
feat_id	feat_name	feat_description	model_feat_imp_train	feat_rank_train
1	team_winp_last5	Ratio of team1's win % to team2's win % in last 5 games	13%	1
2	...	...	...	...
3	...	...	...	...
4	...	...	...	...
5	...	...	...	...

Top 10 or 100 should be decided on average feature importance of all models within the ensemble

*Submission template for file1, file2 can be downloaded from unstop website.

Stages of competition

Round 1 Guidelines

- Participants will train their model using labeled Training Data only. Additional data can only be used to create independent features and cannot be used to add more matches (rows) to the Training data.
- They are free to split the Training data into Train & Out-of-sample/In-time data into whatever ratio they deem fit. They are also free to use any sampling technique on Train data
- Using any future information, that happened during or after the match is strictly prohibited and would lead to disqualification. For e.g. for predicting outcome of a game on 5th Jan 2024 only data before 4th Jan 2024 to be used.
- Match information already provided in Training data like venue, team rosters, toss info, lightning etc. can be used to create independent features.
- Only the following algorithms are allowed – GBM, LightGBM, XGBoost, CatBoost
- An evaluation custom code (in python) to calculate "Accuracy" would be provided to participants to run on their training/out-of-sample data scores to mimic the exact evaluation process that will run on their submitted Round 1 & Round 2 data. Directions to run the code have been mentioned in the beginning of the code. This can't be used on scored Round 1/2 submission data as it doesn't have 'winner_id' column.
- Any names present in the datasets have been masked. Participants are encouraged to use IDs for any merging operations.
- Two CSVs need to be uploaded for Round 1. Participants are required to download and follow exact templates of the submission files from Unstop website.
- Max submissions allowed per team in Round 1 are 20, Leaderboard would be public in this round and team would be rank ordered basis max accuracy
- Amex will be thoroughly evaluating all solutions to ensure integrity & guarding against any misuse or gaming.

Round 2 Guidelines

- Top Participants with respect to ‘Accuracy’ value on Round 1 data & who qualify Amex sanity checks, will be shortlisted for Round 2
- Only Participants shortlisted from Round 1 will be shared Round 2 submission data
- Re-training is not allowed in Round 2. Participants are required to just use the Trained model that led to max accuracy score & use it to re-score Round 2 data.
- Participants will be allowed to submit their scores only once
- Leaderboard would be private
- Only 1 CSV need to be uploaded for Round 1. Participants are required to download and follow exact templates of the Round 2 submission file from Unstop website.

Final Round Guidelines

- Top Participants with respect to ‘Accuracy’ value on Round 1 data & Round 2 data & who qualify Amex sanity checks, will be shortlisted for the Final Round

- Shortlisted teams from Round 2 will be asked to share details of the codes & the datasets used to arrive at the model.

- Shortlisted teams will also create a presentation detailing their approach including (but not limited to) sampling technique, Feature Innovation, Intuitiveness, Selection, Algorithm/Modelling Framework used, Presentation, QnA etc. They will be asked to present the same to a panel.

Top 3 teams will be selected as winners based on Round 1 & Round 2 scores, as well as scores from the presentation

All the Best!!!