Create a prediction model for T20 teams winning rate in future matches.

1. Data Cleaning:

- https://cricsheet.org/downloads/ provides ball-by-ball stats all matches played from 2005-2020. After looking through the data we identified the following attributes which could be relevant to the problem statement.
  - Match_type: T20 , Test,ODI
  - Outcome: win,lose,draw,tie,no result
  - Runs: number of runs per delivery ,extras(leg by,free hit)
  - Wickets: bowled,stumped,caught,run out,lbw
- Stats from all deliveries in all matches are combined to give year wise performance of teams. This performance is recorded separately for each match type. For each of the 3 match types the following stats are extracted
  - Win -number of wins in the match type
  - Loss -number of losses in the match type
  - Other -number of draw,tie,no result in the match type.
  - Sixes -average number of sixes in the match type
  - Fours -average number of fours in the match type
  - Other_runs -average number of runs(except 4's and 6's) in the match_type
  - Bowler_wicket - average number of lbw and bowled
  - Fielder_wicket - average number of run outs and caught
  - Keeper_wicket- average number of stumpings

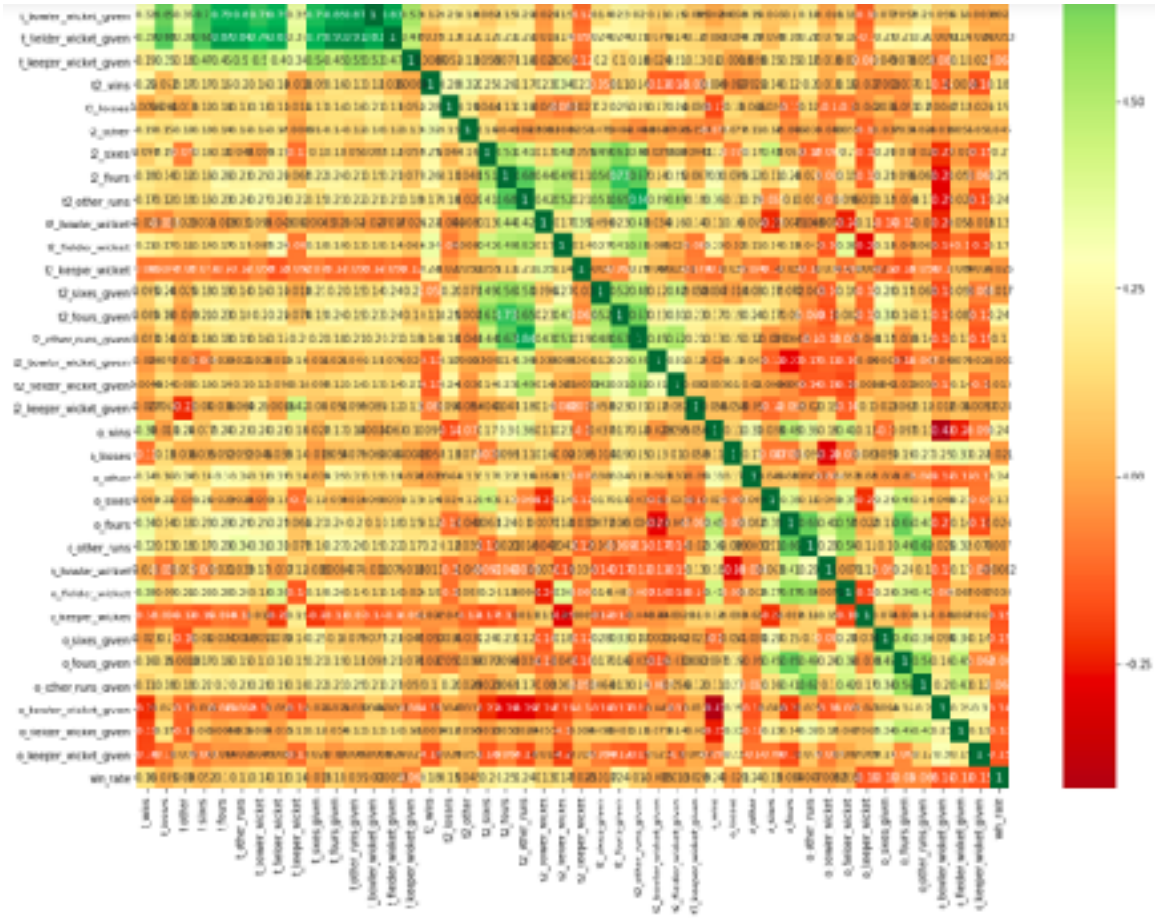
The last 6 features are repeated as sixes_given , fours_given etc, where the runs and wickets conceded in the year are recorded. Thus giving us 45 features for every team per year.

- Analysing this data we observe that the top 10 cricketing nations play more than 30 matches every years whereas nations like Canada and Ireland play about 8-10 matches thus they are discarded from consideration.

2. Data Preprocessing:

- In order to calculate the win rate ther results of first 5 T20 matches at the start of every year are used.
- The win rate of a team in the first five matches of the year are appended to the stats of the team of the previous year.
- To understand the relationship among the features we compute correlation between them and visualize it in form of the following heatmap.

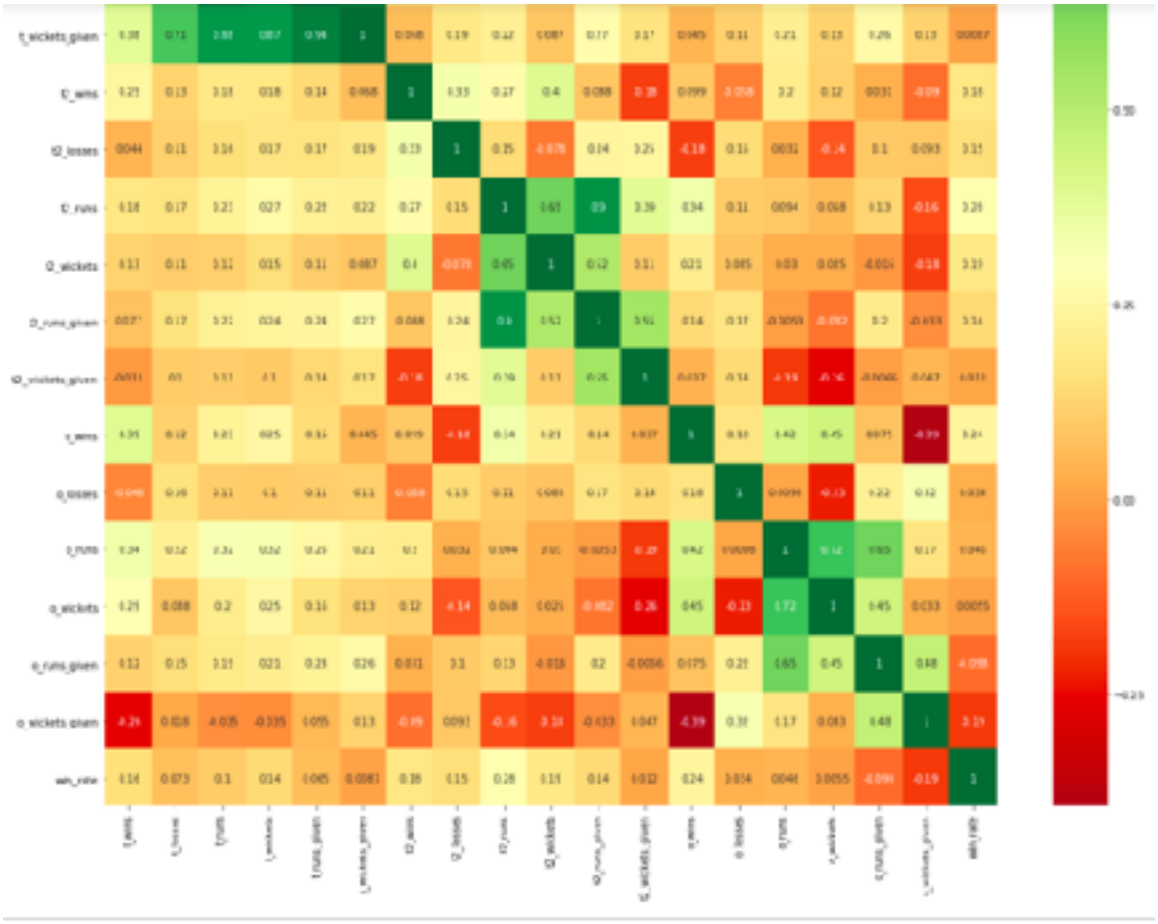
Correlation heat map for all features



The red values show low correlation whereas green shows high corelation. Looking at the last row we see that the win_rate has moderate correlation with all the features.

The different types of runs may not give any additional information when compared to the total runs scored in a match by a team. Same can be said for the type of wicket. Also since we need to predict just the wins losses and draws can be merged.

Correlation heat map after merging features



As expected we see similar correlation of features with the win_rate as we saw with all features.

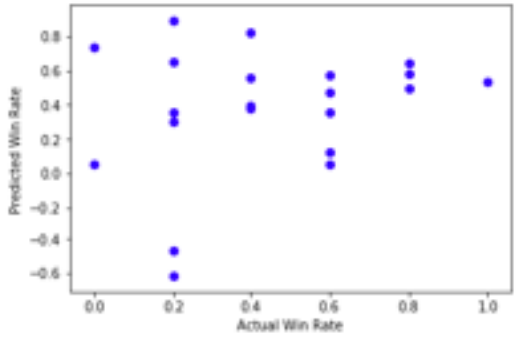
Most suitable features can be extracted using metrics like ANOVA f-value, regression f-value,and mutual_information. Regression f-value is found to be most suitable after testing.

Model selection

- Win rate can be considered as ratio of wins.A simple Multi-Linear regression model can be trained to predict the values.
- We can also consider the number of wins and losses separately . Our output variable y then has a binomial distribution thus we can create a binomial regression model.
- We have tested both , a Multi-Linear regression model trained with least squared cost function and a Binomial regression model with a logistic link function trained using Iteratively reweighted least squares.

Model evaluation

Baseline: linear regression is run on all features to obtain a baseline

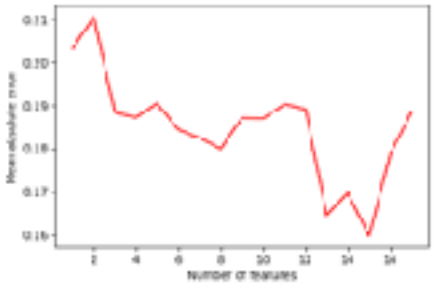


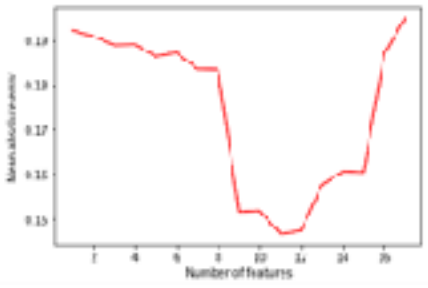

```
explained variance: -1.2246456320798988
r2 score: -1.2750585796258131
mean squared error: 0.1700361242278159
mean absolute error: 0.3261389105159034
```

Tests:

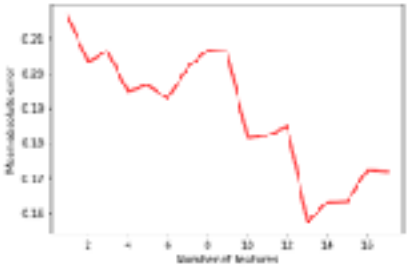
Linear regression

Score_function: ANOVA f-value Score_function: regression f-value



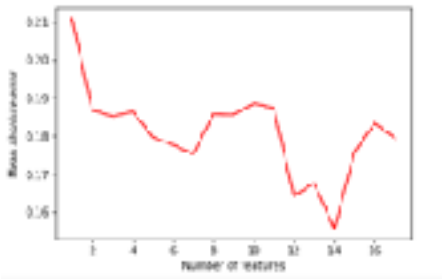


Score_function: mutual-information

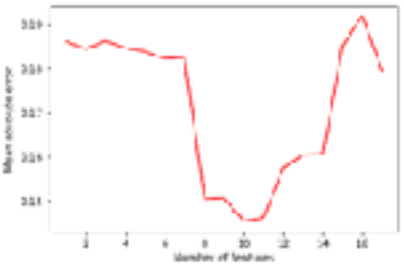


Binomial regression

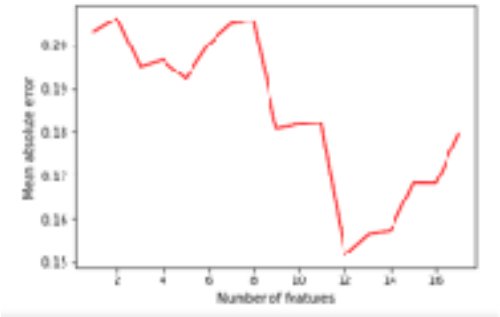
Score_function: ANOVA f-value



Score_function: regression f-value



Score_function: mutual-information



Results

Based on above tests the following model and parameters are selected:

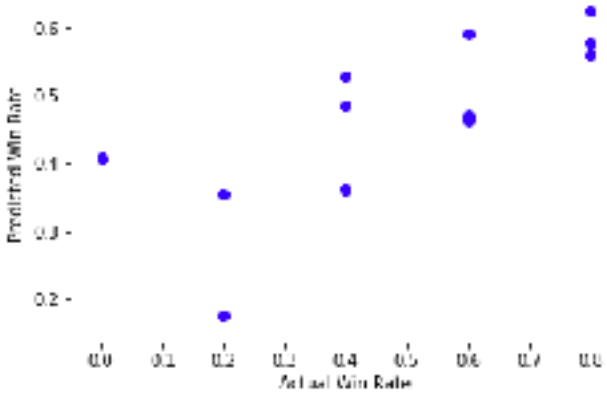
Selected score function: regression f-value

Selected features:

(t2 stands for t20 ,o stands for odi)

Specs	Score
t2_runs	8.640423
o_wins	6.325335
o_wickets_given	3.842940
t2_wickets	3.727767
t2_wins	3.534660
t_wins	2.660683
t2_losses	2.430104
t_wickets	2.113380
t2_runs_given	2.097697
t_runs	1.153775
o_runs_given	1.014611
t_losses	0.562724

Evaluation of selected model




```
explained variance: 0.48977367261100413
r2 score: 0.48502459288052824
mean squared error: 0.03247206039336659
mean absolute error: 0.14615133304820346
```