Title:

Tennis match prediction

Abstract:

Tennis is a much loved sport played around the globe. There are different varietes in the game namely Singles(between two players), Doubles(between two teams(2 players per team and both are same gender)) and Mixed Doubles(between two teams(2 players per team and both are different gender)). It is just a thought if we could able to predict the outcome of the singles game at a different stages (start of the match, after the first set etc) with the different factors associated with each match. Because of the the data constraints the prediction can be done at only during the start of the match. Since the dataset contains an information about the matches and not about the players, a new dataset will be derived from the existing dataset with the desired features. These features are derived anticipating that it will help them in the prediction. So there is a littile amount of risk in predicting the match with these new features. Different techniques like LDA, RandomForest, Logistic Regression etc. will be applied to predict the outcome of the match.

Introduction:

Dataset for this project can be downloaded from the below site.

https://www.kaggle.com/gmadevs/atp-matches-dataset

This dataset has data for the matches played from 2000 to 2017 which covers all the tournament types like atp tours, grand slam and masters. It has 27 features/dimensions including the winner of the match. Training set will be from 2000 to 2016. Model is tested against the 2017 matches.

Some of the features are

w_ace – no of aces hit by the winner.
winner_hand – whether the winner is right-handed or left-handed.

In order to predict the outcome of the match, two types of statistics are required. One is match statistics and the other one is player statistics or player current form. The above dataset has only match statistics. i.e the information regarding the match happened between the two players. But player current form is crucial in deciding the outcome of the match. There is a saying “Form is temporary but the class is permanent”. Hence a new dataset will be obtained from the existing dataset which has the statistics about the player current form. The calculation for the new dataset will be discussed later.

Methodology:

This problem is identified as classification technique since it involves only two outcomes (win or loss) with respect to a single player.

Original Data set:

	tourney_id	tourney_name	surface	draw_size	tourney_level	tourney_date	match_num	winner_id	winner_seed	winner_entry
1	2000-717	Orlando	Clay	32	A	20000501	1	102179	NA	

winner_name	winner_hand	winner_ht	winner_ioc	winner_age	winner_rank	winner_rank_points	loser_id	loser_seed	loser_entry
Antony Dupuis	R	185	FRA	27.18138	113	351	102776	1	

loser_name	loser_hand	loser_ht	loser_ioc	loser_age	loser_rank	loser_rank_points	score	best_of	round	minutes
Andrew Ilie	R	180	AUS	24.03559	50	762	3-6 7-6(6) 7-6(4)	3	R32	162

w_ace	w_df	w_svpt	w_1stIn	w_1stWon	w_2ndWon	w_SvGms	w_bpSaved	w_bpFaced	l_ace	l_df	l_svpt	l_1stIn	l_1stWon	l_2ndWon
8	1	126	76	56	29	16	14	15	13	4	110	59	49	31

l_SvGms	l_bpSaved	l_bpFaced
17	4	4

Derived Data set:

	p1	p2	surface	rankDiff	ageDiff	winner	acdDiff	dfDiff	svptDiff	s1stwonDiff	s2ndWonDiff	bpDiff
1	103285	104571	Hard	71	7	TRUE	-0.35	0.44	5.84	6.07	-3.27	0.63
2	104460	111581	Hard	-125	13	TRUE	3.53	-0.84	9.20	6.56	2.75	0.45
3	105777	106233	Hard	9	2	TRUE	-0.10	0.49	8.21	4.46	0.18	0.52
4	111575	122570	Hard	-1224	-1	TRUE	7.36	1.14	30.18	20.73	11.91	-2.64
5	105671	106331	Hard	-157	4	TRUE	3.96	2.08	69.28	25.72	17.28	3.28

The above original dataset contains two type of information.

1. pre-match information like tournament id, surface it is played, type of tournament(ATP, Masters, Grandslam etc), draw size, players, age and rank of the players.
2. Post-match information like number of aces hit by the player. Number of first serve won by the player etc. It also has the winner id of the match.

In this analysis pre-match analysis will be taken as such and post-match information will be fed into the statistics calculator to get the current form of the player.

Hence the new dataset contatins.

1. Pre-match information like surface, rank difference between the two players, age difference between the two players and their player id's.
2. Post-match information like difference between the average aces(aces/match) hit by the player, difference between the average 1st serve won by the player etc.

For the year 2000, since it doesn't have previous year statistics the current form of the player will be taken as zero, which is not correct solution. Since there is no viable solution, this decision has been made.

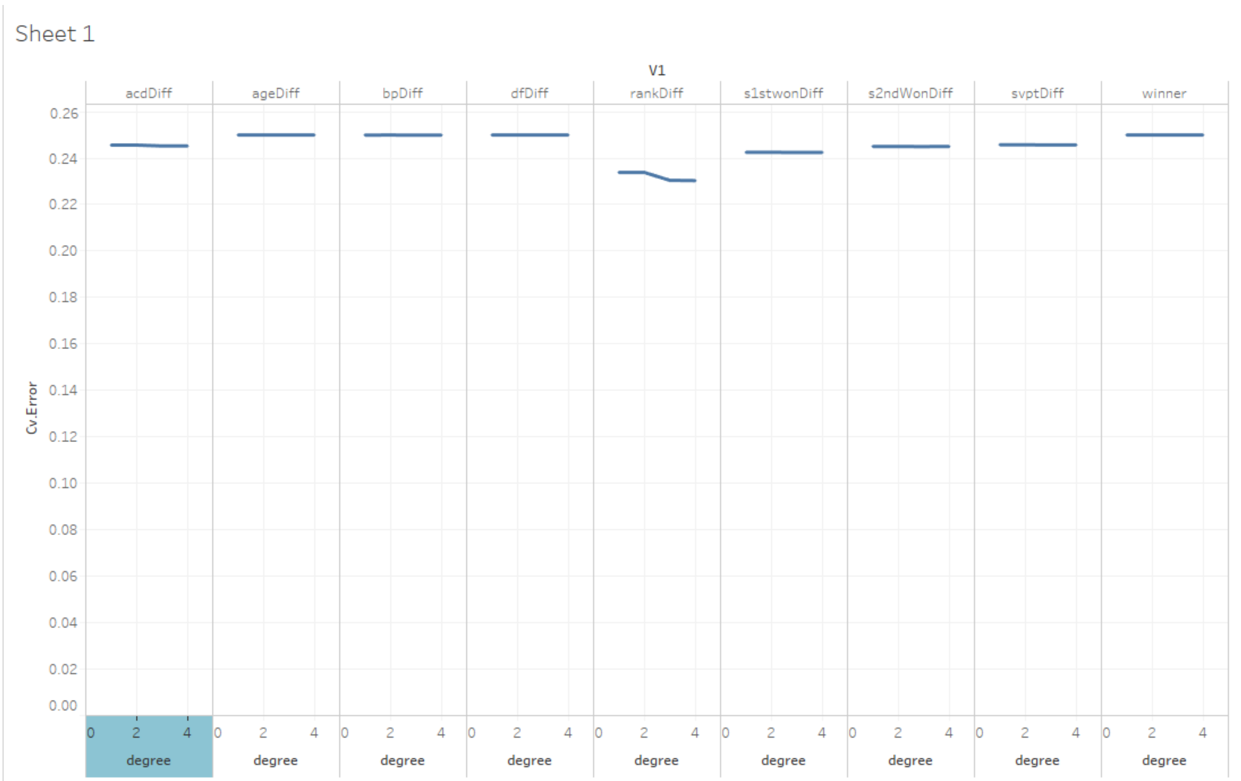
Data Analysis:

Matches played between the year 2000 to 2015 will be taken as a training data. 2016 matches will be the test data for this analysis.

Training data contains close to 50,000 matches.

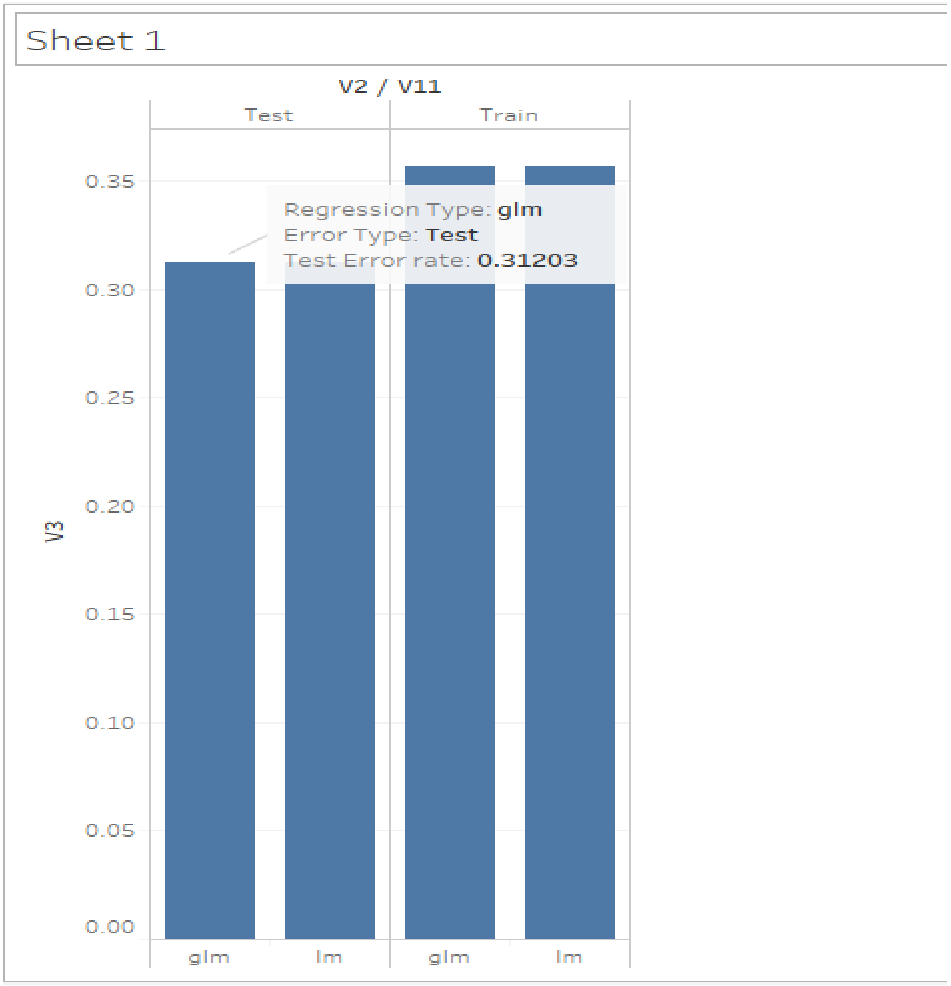
Test data contains close to 265 matches.

Results



10 k -fold cross validation has been perfomed to verify whether the polynomial degree of the features helps in reducing the error. But from the above graph it is observed that there is no improvement even if the attributes/features are non-linear.

Multiple Linear Regression and Logistic Linear regression

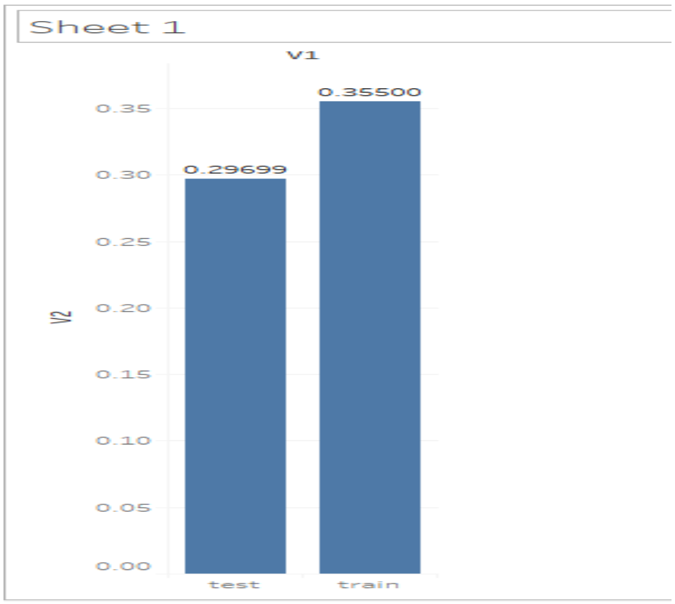


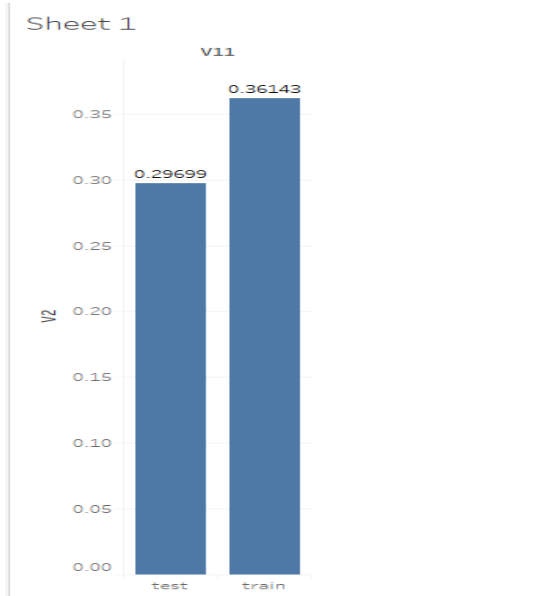
Multiple Linear regression and logistic linear regression obtains the same test error rate of 31%. Initially the model was built with all the features and then based on the p-value, some of the features have been removed. Hence the final model contains below attributes.

winner~ rankDiff+ageDiff+acdDiff+svptDiff+s1stwonDiff+s1stwonDiff+s2ndWonDiff

+bpDiff

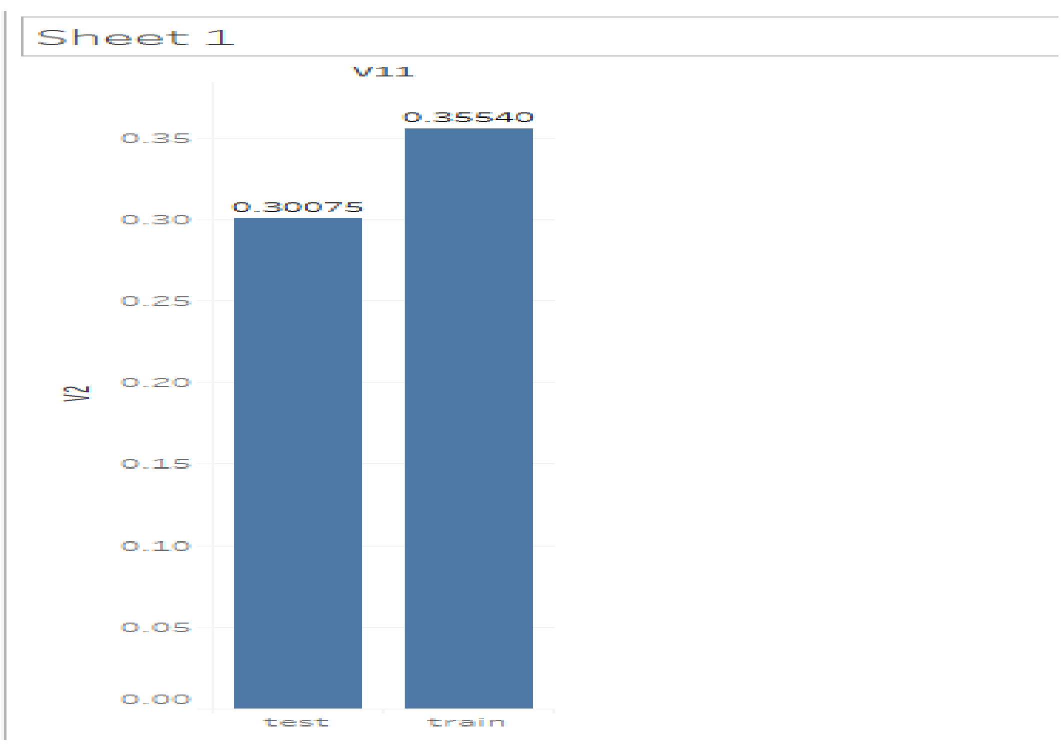
LDA QDA



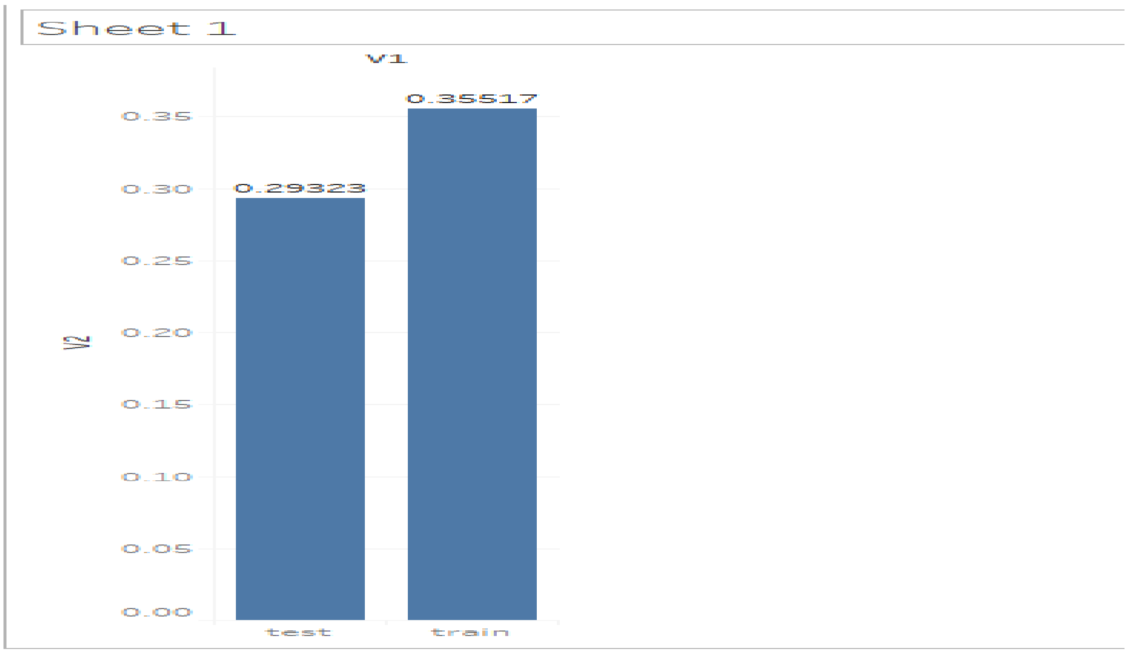


LDA and QDA obtains the same test error rate of 29 %. This is relatively better compared to the logistic regression model.

Lasso

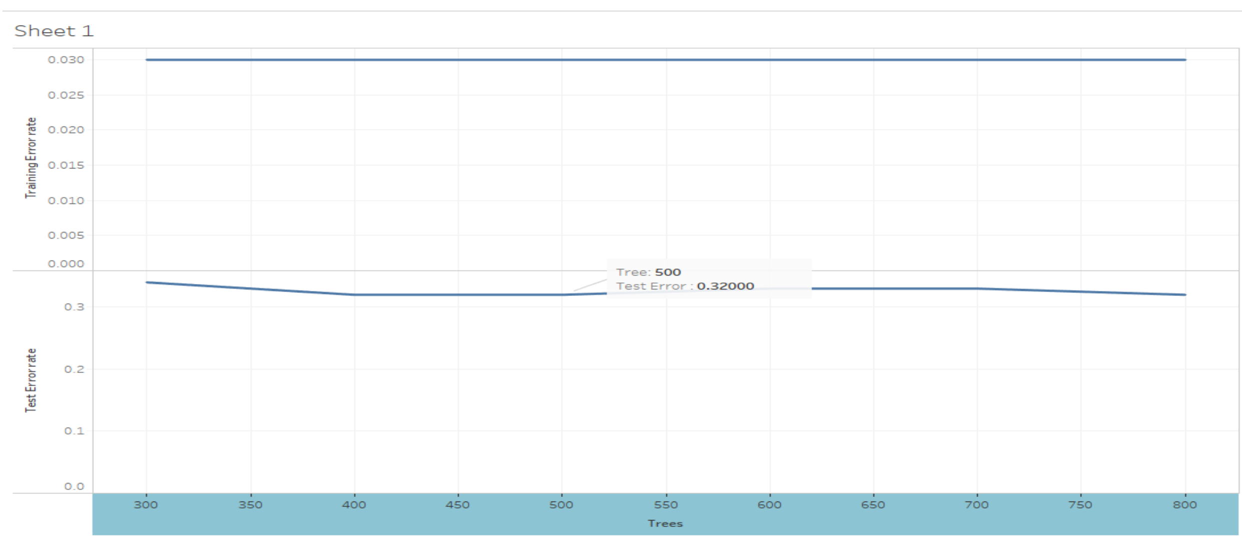


Ridge



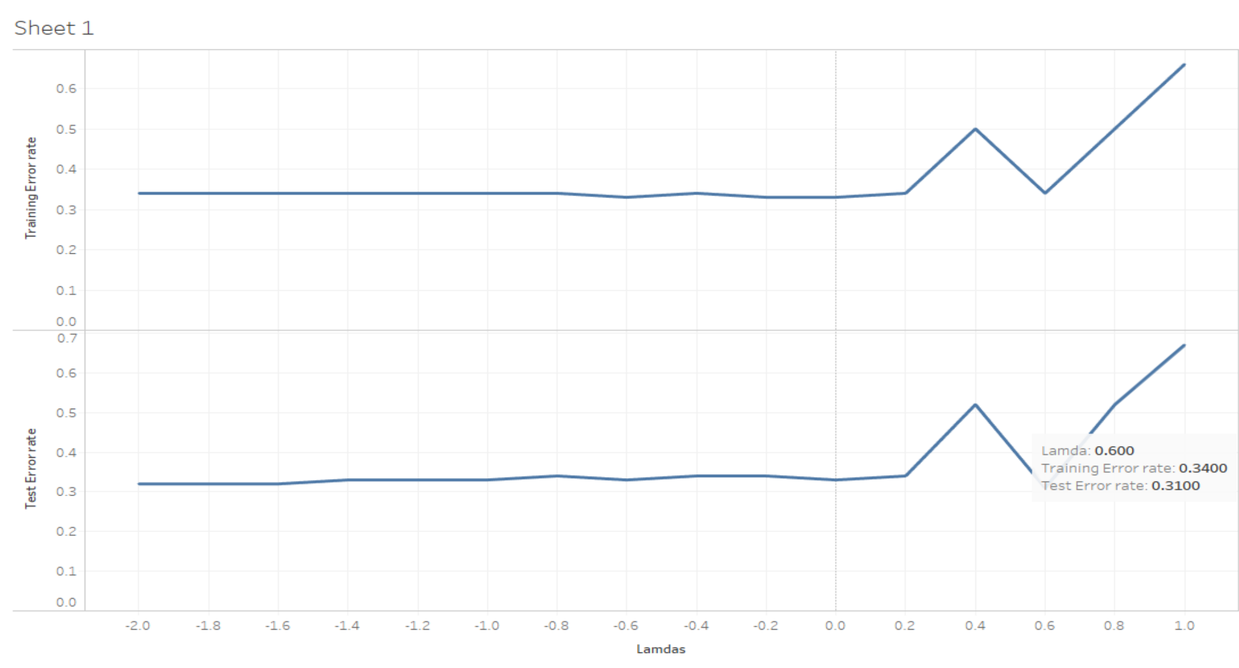
Ridge regression performs slightly better than lasso and other models. It records the lowest error rate of all the models. It is understandable that n>p ridge performs better than lasso and other models.

Random Forest



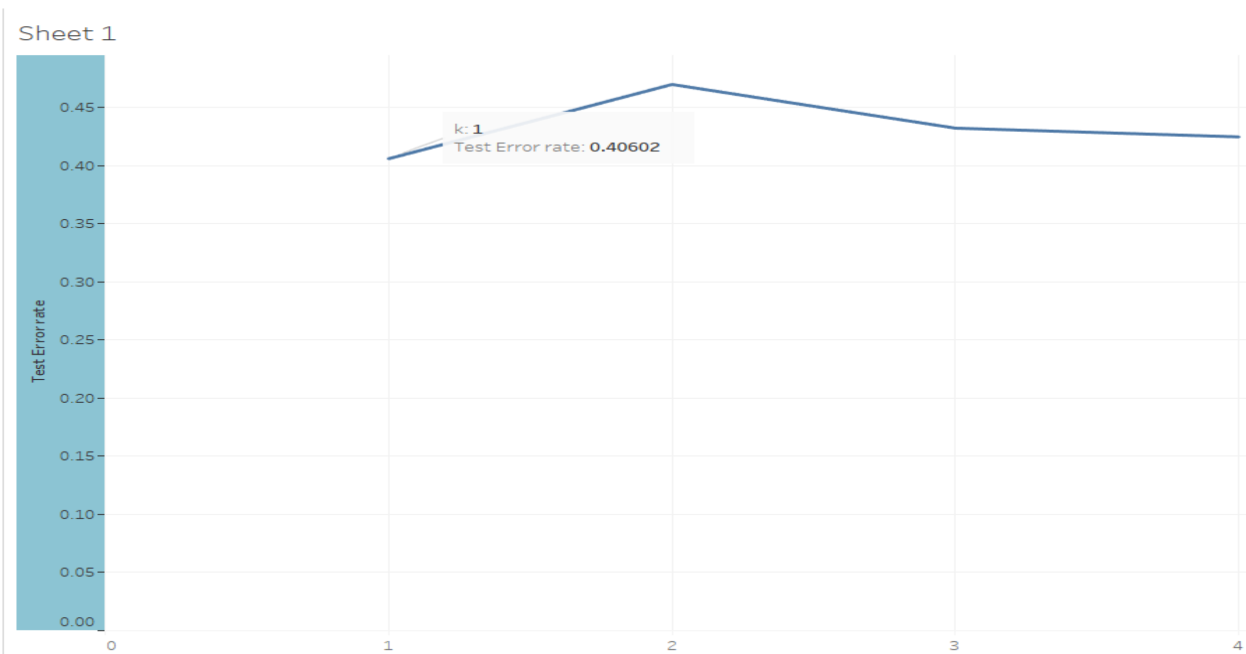
Random Forest records the lowest test error rates of 32% when the tree size is 500.

Boosting



Boosting lowest test error rate is 31% when the lambda is 0.6

KNN



KNN performs very poorly as it's best test error rate is 40% when k=1.

Conclusions:

From all the above techniques which has been used for the prediction, the lowest test error rate we could able to achieve is 29% by LDA. This means this model could able to predict 71% of the matches. This is reasonably a fair accuracy considering there is still a randomness in the sports. One more interesting thing we could able to observe is both the test error rate and training error rate are similar in most of the techniques. This explains that bias is more and variance is less when fitting the models. If the case is vice versa, then the training error rate will be far lesser than the test error rate. This analysis can be further improved by adjusting a bias-variance trade off. And many of the factors are derived as a new feature from the existing dataset, interpretation of the original feature has been lost.

Limitations and Improvements

The first year 2000 doesn't have data to calculate a player current form/player statistics. This may could lead to a bad model design since we are potentially nullifying the player's current form.

Player's current form has been calcuated by taking the statistics from the last year. This can be further improved by taking the statistics till the last match he played.

Other factors like player's health condition(smoking, drug use) also can be used as a feature in the future if it's been provided.

References:

1] T. Barnett and S. R. Clarke. Combining player statistics to predict outcomes of tennis matches.

IMA Journal of Management Mathematics, 16:113–120, 2005.

[2] T. Barnett and G. Pollard. How the tennis court surface affects player performance and injuries.

Medicine Science Tennis, 12(1):34–37, 2007.

[3] J. E. Bickel. Some Comparisons among Quadratic, Spherical, and Logarithmic Scoring Rules.

Decision Analysis, 4(2):49–65, 2007.

[4] L. Breiman. Bagging predictors. Machine Learning, 24(2):123–140, 1996.

[5] S. R. Clarke. An adjustive rating system for tennis and squash players. In Mathematics and

Computers in Sport, 1994.