
Workforce Analytics

Individual Assignment

Introduction

This assignment analyses the data containing the NCAA basketball games for Ivy League colleges between 2006 and 2015. The objective is to study the effects of the factors that influence team performance on each game.

The Python code for this homework is submitted in Jupyter notebook format, with filename **WFIndividualHW.ipynb**

Part 1: Performance Measure of NCAA Basketball Team

For the NCAA basketball game, a college's season standing is determined by the number of games won. The actual margin of victory (i.e. difference of points scored over opponent in each game) is relatively unimportant. At the end of the season, the team with the highest number of wins will be declared as season league champion.

For this reason, the important performance measure would be whether a college wins/loses a game, since it directly influences the college's final season standing. To construct this win/lose binary variable, the difference in points between the Ivy League team and its opponent ('Outcome' variable in the CSV file) is used. If the difference is positive, the Ivy League team won the game. Otherwise, the team lost the game. This binary variable is stored as 'win' variable in the Python code (1 represent win, 0 represent loss).

Note that in basketball, there is no "Draw" at game-level. Each game must have a winner and loser. If there is a tie in points scored by both teams by the end of regular time, overtime will be played until a winner emerges. Therefore, the performance measure is a strictly binary variable (win/loss). Binary logistic regression will be employed later in Part 4 to predict the target variable.

Part 2: Additional Variables

In Part 4, a predictive model will be built to examine how the variables (in the CSV file as well as additional attributes obtained from <http://www.sports-reference.com/>) influence team performance. Hence, it is important that the additional variables obtained must relate to information that is available before a game is played, in order to assess their predictive power. For instance, "Field Goal Percentage" of the same game should not be used in the predictive model because this stat is only known after the game has concluded.

Supplement Variables Using Data from sports-reference Website

Going by the above principle, the potential variables (to supplement the predictive model) consist of those that are already known before the match. These two variables are constructed using data from sports-reference website:

- Difference between the Two Teams' Winning Percentages in Previous Season
 - Variable Construction: In sports-reference website, look up the previous season records of the two teams who are competing in a game. Using their previous season's winning percentage, calculate the difference between the two teams. In the case where the opponent team has no record in the previous season, its winning percentage is assumed to be zero.
 - Purpose: This serves as a proxy of the basketball team's relative true strength. A high-performing team should have good track record in the recent past and the previous year record is a perfect indicator of that. This variable also acts as a control variable to control for the difference between opposing teams'

strengths. After controlling for this factor, we can more accurately measure the effects of other factors (e.g. experience, diversity, etc.)

- The reason that the winning percentage is used (instead of other stats) is because it is a simpler and intuitive proxy to the difference in basketball teams' relative true strength. Other game stats (e.g. Field Goals Percentage, Rebounds Per Game) largely depend on the team's playing style and may fluctuate from team to team, hence these match stats might not serve as a good proxy to the difference in team strengths.
- **Position Diversity of a Team**
 - **Variable Construction:** From sports-reference website, obtain a team's current season roster, which contains the playing position (i.e. Guard, Forward, Center) of each player in the team. With this list of positions, the "1 – Herfindahl Index" can be used to measure diversity in playing positions:

$$Diversity = 1 - \sum_{i=1}^N r_i^2$$

where r_i is the proportion of a type of position in the whole list.

- **Purpose:** This variable measures whether a team has different players (regular and reserve players) with diversified playing positions. This may be important when major players become unavailable (e.g. injury), and reserve players may need to stand in. Having a team of players with diversified playing positions allow the coach to employ different playing styles in different circumstances. This may enhance the versatility of the team and improve team performance.

Supplement Variables by Transforming Existing Variables in CSV File

Other than the above two variables supplemented using information from sports-reference website, two other variables are also constructed using the existing information from the CSV file.

1. Winning/Losing Streak

- **Variable Construction:** Using the team's previous match records in the same season, a streak variable is constructed, indicating the number of consecutive wins/losses before the game. A positive number indicates number of consecutive wins (e.g. +2 indicates two consecutive wins before the match) while a negative number indicates number of consecutive losses (e.g. -3 indicates three consecutive losses before the match).
- **Purpose:** This variable measures the effects of momentum. A team which has secure many victories recently may have built-up good momentum (and hence good team morale) and might help them to secure more victories in future. On the other hand, a team which experiences consecutive losses might experience great pressures to get back in form and this might impact their future performance.

2. Difference between the Two Teams' Winning Percentages in Current Season So Far

- **Variable Construction:** Using the team's previous match records in the same season, we can calculate the winning percentage in current season "thus far" before the game. Then we subtract away the "opponent win:loss" to get the difference between the two teams' winning percentages thus far in this season. In the case of first game of the season where there is no previous match records in the same season, the "current season winning percentage difference" is assumed to be zero.
- **Purpose:** This variable measure the team strength displayed in current season so far. A team which exhibits strong performance in current season may indicate recent improvement in team dynamics. Therefore, this variable might be predictive of future performances in the same season.
- **Note** that this variable supplements the variable "Difference between the Two Teams' Winning Percentages in Previous Season" constructed earlier using sports-reference data. Compared to that variable, current season winning percentages is constructed using the limited match data (i.e. based on the previous match results in the same season, prior to the game) may be affected by short-term fluctuations in performance. Hence this variable is more suited to act as a proxy for "Short-Term Performance Differential" while the winning percentages in previous season reflect the intrinsic true strength of the basketball team.

Part 3: Selection of Predictor Variables

Since our objective is to identify how certain factors predict team performance in basketball context, it will be useful to include the nonlinear terms (e.g. quadratic terms and interaction term) to identify if there is any nonlinear relationships between the target variable and predictor variables.

Nonlinear Forms of Predictor Variables Considered

Based on the variables given in the CSV file as well as the variables constructed/transformed in Part 2, their nonlinear cousins, which can potentially be used in the final model, are tabulated below.

Nonlinear Variable	Forms of	Description
Game²		This is the square of Game Number. Towards the end of the season, the games might get more fiercely competitive. This quadratic form investigates if there is any 'reverse' effect towards the end of the season.
(Height Diversity)²		Height diversity is a proxy for inherent diversity. As discussed in class, an optimal team should have moderate amount of diversity, hence this quadratic term is to investigate if too much diversity lowers the team performance.
(Ethnic Diversity)²		Similar to height diversity, a kind of inherent diversity.
(Class Diversity)²		Similar to height diversity, a kind of inherent diversity.
(School Diversity)²		This represents the kind of school players went and is a proxy for acquired diversity. We also add the quadratic term to investigate the nonlinear effects of acquired diversity.
(Degree Diversity)²		Similar to school diversity, a kind of acquired diversity.
(Position Diversity)²		Similar to school diversity, a kind of acquired diversity.
(Experience Opponent)²	with	Gaining more experience playing against the opponent might improve the chances of winning against them. However, playing too many times against the opponent also provides a chance for the opponent to study our team's playing style. Therefore, this quadratic term is intended to study whether there is any nonlinear effect.
Streak²		As discussed in Part 1, the momentum effects of "Streak" may affect the chances of winning. However, if a team loses too many times consecutively, its players may be strongly motivated to win. Conversely, if a team wins too many times consecutively, its players may be under high pressure to keep on the winning momentum. Hence we add the quadratic term to study the nonlinear effect.
Game * Current Season Winning % Diff		<p>This is the interaction term between "Game Number" and "Winning Percentage Difference in current season thus far". The reason behind this interaction is twofold:</p> <ul style="list-style-type: none">• Towards the end of the season, the strong teams would go all out defending their leading positions in the league table, hence the team strength factor might be more evident towards end of season.• The "Current Season Winning % Diff" is calculated using the winning ratio in the matches played <u>thus far</u> in current season. At the start of the season, there might be not enough games to establish an accurate estimate of the current season's team strength. As the season advances, this estimate might grow more accurate and its effects become more evident.

Table 1 Potential Nonlinear Terms to be used in the Model

Multicollinearity

Now, we have a large pool of predictor variables that can be used in the predictive model. However, careful selection is required to select only a subset of variables to be included in the final model. This is because if too many correlated predictor variables are added to the model, it might cause multi-collinearity issue and the estimated effects of certain variables may become insignificant. Furthermore, the parameter estimates may become unstable, where small change in data might result in large change in parameter estimates. This would create difficulty in estimating the effects of the predictor variables on team performance.

To combat the multicollinearity issue, we must not add highly correlated predictors to the same model. To identify which predictor variables may create multicollinearity issue, Variance Inflation Factors (VIF) are

calculated when all the variables listed in Table 2 are included in a regression model. A VIF value greater than 10 may indicate the presence of multicollinearity.

Note that the “Description” column of Table 2 describes what the variables measure.

Variable Name	Description	VIF
Game	Game Number	21.60
Game ²	Quadratic form of Game Number	17.61
same_conf	Dummy indicating if the two opposing teams are from same conference	61.79
home game	Dummy indicating whether it is a home game	33.38
height diversity	Height diversity (inherent diversity)	9.12
(height diversity) ²	Quadratic form of Height Diversity	8.60
ethnic diversity	Ethnic diversity (inherent diversity)	52.02
(ethnic diversity) ²	Quadratic form of Ethnic Diversity	51.84
class diversity	Class diversity (inherent diversity)	19.46
(class diversity) ²	Quadratic form of Class Diversity	17.58
school diversity	School diversity (acquired diversity)	68.48
(school diversity) ²	Quadratic form of School Diversity	69.93
degree diversity	Degree diversity (acquired diversity)	60.88
(degree diversity) ²	Quadratic form of Degree Diversity	57.07
experience with opponent	Experience playing with opponent in current season	480.64
(experience with opponent) ²	Quadratic form	241.23
experience in arena	Experience playing in this arena in current season	33.38
opp_conference_size	Number of teams in the opponent’s conference. This could be a proxy of opponent’s experience. Opponent from larger conference may have more experience playing competitively	3.48
opp_conf_wl	Opponent’s performance in its own conference in the season.	3.17
opp_games_played	Total number of games to be played by the opponent in this season. Similar to “opp_conference_size”, a higher number of games to be played may imply the opponent has more experience playing competitively	2.31
Past Win Percentage Diff	Variable constructed in part 2. It is the winning % difference in the previous season and is a proxy for relative team strengths	2.37
Position Diversity	Variable constructed in part 2, to investigate whether a team with diverse playing positions would perform better.	269.19
(Position Diversity) ²	Quadratic form	269.83
streak	Variable transformed in part 2, investigate whether winning/losing momentum affects a team’s future performance	1.39
streak ²	Quadratic form	1.08
Current Season Winning % Difference	Variable transformed in part 2, investigates whether the current season’s performance (before the game) would predict the future game performance in the same season	3.70
Game * Current Season Winning % Difference	Interaction between “Game Number” and the current season winning percentage difference. Rationale was described in Table 1	3.42

Table 2 Predictor Variables and their Variance Inflation Factors

As seen from Table 2, many of the variables have large VIF value. In Figure 1, the bars are sorted by VIF value. The variables higher in the list (with higher VIF values) are to be reviewed. Some of the variables might need to be removed in order to avoid multicollinearity issue.

Correlation between Predictor Variables

After knowing the variables which cause multicollinearity issue, we would need to identify which variables are their highly correlated cousins, in order to determine which variable could be removed from a group of highly-correlated variables. To quickly identify the groups of highly correlated variables, a heatmap (Figure 2) is used to visualise the amount of correlations between pairs of variables.

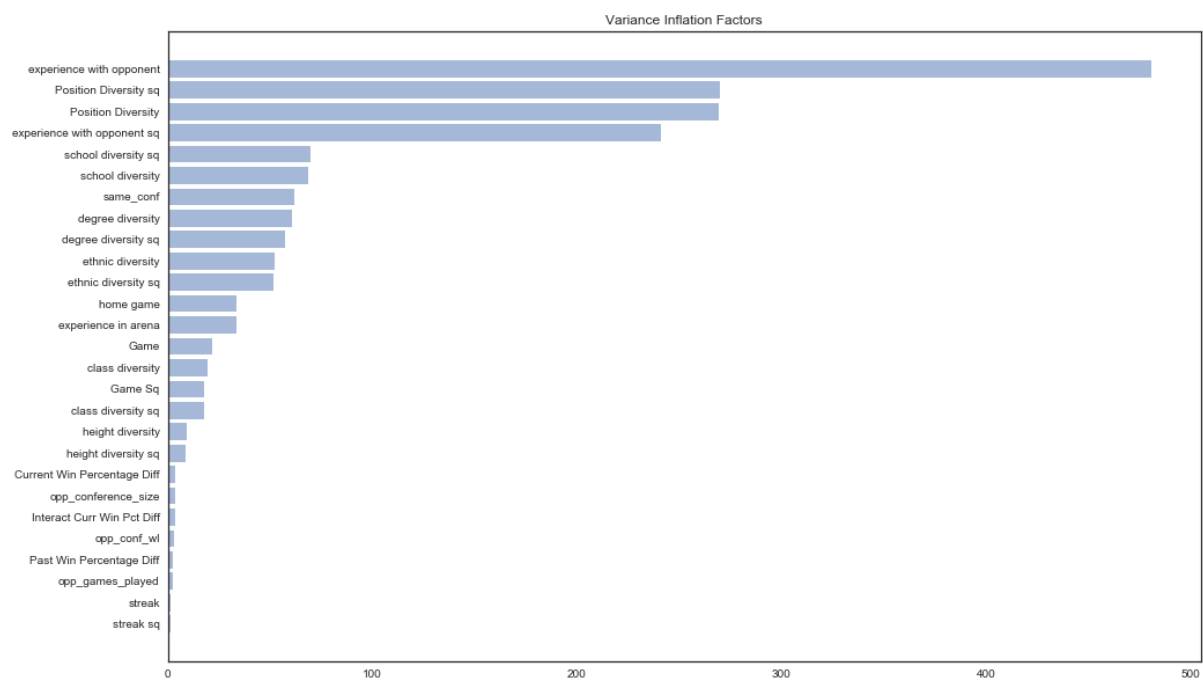


Figure 1 Variance Inflation Factors (sorted by magnitudes)

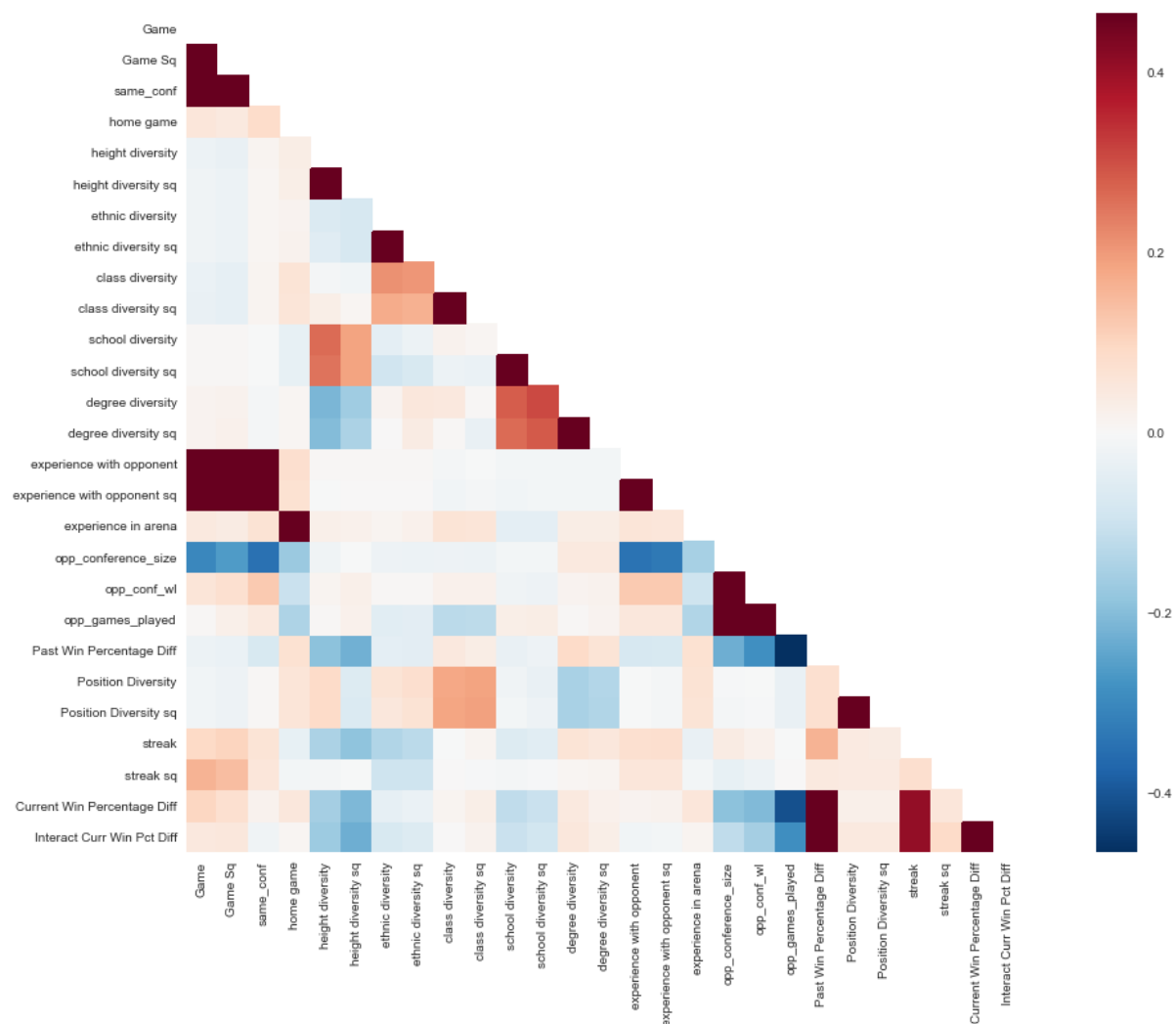


Figure 2 Correlation Heatmap of the Predictor Variables

Referencing the pattern shown in Figure 2, the following groups of highly correlated variables are identified:

No	Group of Highly Correlated Variables	Possible Relationships
1	<ul style="list-style-type: none"> - Game - Game² - same_conf - Experience with Opponent - (Experience with Opponent)² 	<ul style="list-style-type: none"> • At the start of the season, a college may play against teams from other conferences. As the season advances, the college only plays against the teams from the same conference. Therefore, there is high correlation between “Game Number” and “same_conf” dummy variable • Since most of the matches are played against a team from the same conference, as the Game Number increases, the experience playing against the opponent also increases. This explains the high correlation between “Game Number” and “Experience with Opponent”
2	<ul style="list-style-type: none"> - Game * Current Season Winning % Difference - Past Season Winning % Difference - Current Season Winning % Difference 	<ul style="list-style-type: none"> • These three variables are related to the estimated difference in team strength, in terms of winning percentage difference. Hence they are all highly correlated
3	<ul style="list-style-type: none"> - home_game - experience in arena 	<ul style="list-style-type: none"> • If a team is playing in its home ground, then they would have ample experience playing in their home arena, resulting in high correlation
4	<ul style="list-style-type: none"> - Height Diversity - (Height Diversity)² 	<ul style="list-style-type: none"> • The linear term is highly correlated with the quadratic term
5	<ul style="list-style-type: none"> - Ethnic Diversity - (Ethnic Diversity)² 	<ul style="list-style-type: none"> • The linear term is highly correlated with the quadratic term
6	<ul style="list-style-type: none"> - Class Diversity - (Class Diversity)² 	<ul style="list-style-type: none"> • The linear term is highly correlated with the quadratic term
7	<ul style="list-style-type: none"> - School Diversity - (School Diversity)² 	<ul style="list-style-type: none"> • The linear term is highly correlated with the quadratic term
8	<ul style="list-style-type: none"> - Degree Diversity - (Degree Diversity)² 	<ul style="list-style-type: none"> • The linear term is highly correlated with the quadratic term
9	<ul style="list-style-type: none"> - Position Diversity - (Position Diversity)² 	<ul style="list-style-type: none"> • The linear term is highly correlated with the quadratic term
10	<ul style="list-style-type: none"> - Opp_conf_wl - Opp_conference_size - Opp_games_played 	<ul style="list-style-type: none"> • These three variables are related to the opponent’s conference size and conference performance. If the opponent’s conference size is large, then they are scheduled to play more games, resulting in high correlation

Table 3 Groups of Highly Correlated Variables

Feature Importance

After identifying the groups of highly correlated variables, next we would have to determine which variable to remove from each group. Ideally we would like to retain the variables that are highly predictive so that our final predictive model will be highly effective. To assess the importance of the predictor variables, two techniques have been employed.

Feature Importance Estimation Method 1: Random Forest Method

Random forest can be used to estimate the feature importance. Random forest grows multiple trees and each tree performs splitting using a random subset of features. The features which appear many times in multiple trees are deemed to be of higher importance. In other words, they are more predictive with respect to the binary outcome of basketball game (i.e. win or loss).

To estimate the feature importance, the predictor variables are used to build a random forest (consisting of 100 trees) and 10-fold cross validation is used to tune the hyper-parameter (i.e. minimum number of samples

at leaf node). The random forest used achieves a test set accuracy of 68.47%. The final ranking of predictor importance is displayed in Figure 3.

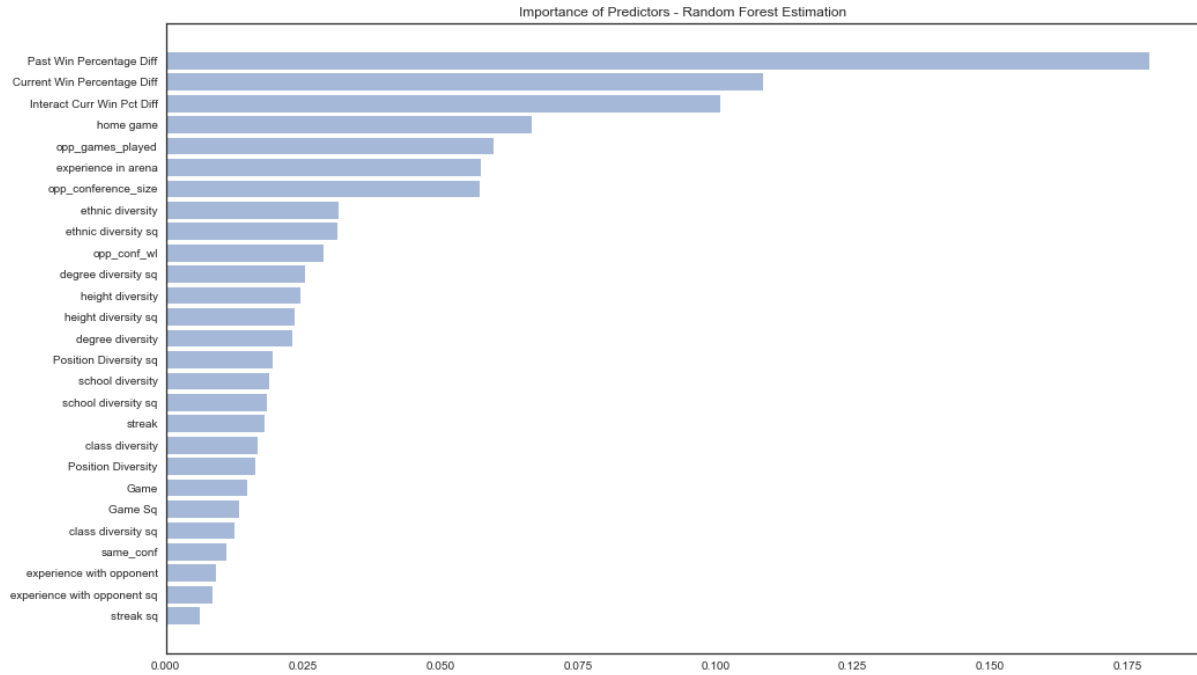


Figure 3 Importance of Predictor Variables – Estimated Using Random Forests

Figure 3 provides the information of the relative importance of the predictor variables, which can be referenced while deciding which variables to remove from the final model.

Other than random forest, another technique (i.e. L1 regularise logistic regression) has also been employed to estimate the importance of the features.

Feature Importance Estimation Method 2: L1 Regularised Logistic Regression

L1 regularised logistic regression adds the L1-norm of coefficient magnitudes to the logistic regression objective function:

$$\min_{w,c} \|w\|_1 + C \sum_{i=1}^n \log(e^{(-y_i(X_i^T w + c))} + 1)$$

As seen from the above equation, the absolute coefficient values are penalised in the objective function. L1-regularisation has a property where the unimportant features will be shrunk towards zero. The penalty parameter C determines the relative amount of penalty to be imposed on absolute coefficient values, this parameter should be fine-adjusted according to the dataset characteristics.

To estimate the importance of all the predictor variables, they are used to build a L1-regularised logistic regression model to predict game outcome (i.e. win or loss). Ten-fold cross validation is used to tune the optimal value of penalty parameter C . This L1-regularised logistic regression model achieves a test set accuracy of 70.61%.

The absolute coefficient values of this model are shown in Figure 4. In Figure 4, only eight variables have nonzero coefficients. These eight variables are deemed to be the important predictors. Other predictors are relatively unimportant.

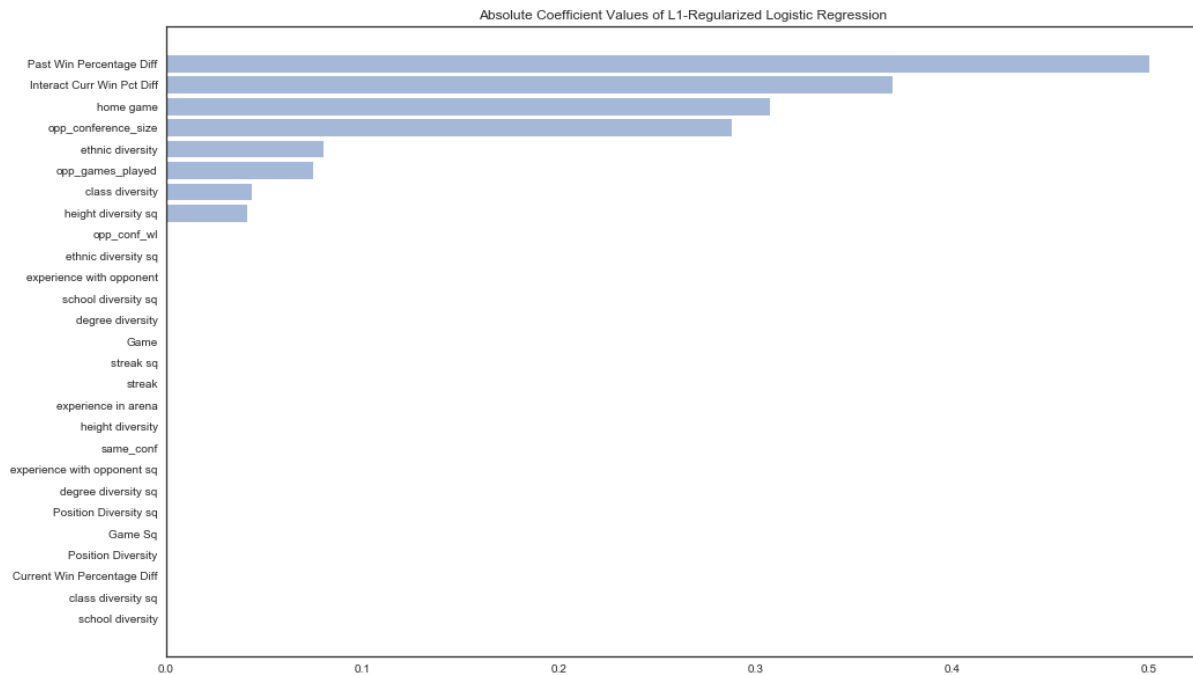


Figure 4 Importance of Predictor Variables – Estimated Using L1-Regularised Logistic Regression

After knowing the importance of the predictor variables, we can now make an informed decision in removing the variables with low predictive power yet highly correlated with other important predictors.

Feature Selection

Using the information from the VIF table, correlation heatmap and feature importance scores (from Random Forest and L1-regularised logistic regression); we can now decide which variables are to be removed from final model. Table 4 lists down the variables to be removed from each of the groups of highly correlated variables.

No	Group of Highly Correlated Variables	Removed Variables	Reasons of Removing / Not Removing
1	<ul style="list-style-type: none"> - Game - Game² - same_conf - Experience with Opponent - (Experience with Opponent)² 	<ul style="list-style-type: none"> - Game² - same_conf - (Experience with Opponent)² 	<p>These three variables have low importance scores in both Random Forest and L1-Regularised Logistic Regression estimations.</p> <ul style="list-style-type: none"> • Since Game and Game² have high VIF values, the quadratic term needs to be removed to avoid multicollinearity • Since the teams tend to play against team from the same conference later in the season, "Game Number" should provide sufficient information and "same_conf" is redundant • (Experience with Opponent) and (Experience with Opponent)² have high VIF values, the quadratic term needs to be removed to avoid multicollinearity
2	<ul style="list-style-type: none"> - Game * Current Season Winning % Difference - Past Season Winning % Difference - Current Season Winning % Difference 	None	<p>Since the VIF values of these three variables are small (all smaller than 4), they do not cause much multicollinearity issues. Thus, none of them is required to be removed.</p> <p>These three variables are also deemed (by</p>

			Random Forest and L1-Regularised Logistic Regression) as very important predictors
3	- home_game - experience in arena	- experience in arena	“home_game” has high importance score in both measures while “experience in arena” is relatively unimportant. To avoid multicollinearity effect, “experience in arena” is removed
4	- Height Diversity - (Height Diversity) ²	None	The VIF values of these two variables are not high (less than 10), hence both of them are kept
5	- Ethnic Diversity - (Ethnic Diversity) ²	- (Ethnic Diversity) ²	Since both of these variables have low importance scores, we remove the quadratic term to avoid multicollinearity issue
6	- Class Diversity - (Class Diversity) ²	- (Class Diversity) ²	Since both of these variables have low importance scores, we remove the quadratic term to avoid multicollinearity issue
7	- School Diversity - (School Diversity) ²	- (School Diversity) ²	Since both of these variables have low importance scores, we remove the quadratic term to avoid multicollinearity issue
8	- Degree Diversity - (Degree Diversity) ²	- (Degree Diversity) ²	Since both of these variables have low importance scores, we remove the quadratic term to avoid multicollinearity issue
9	- Position Diversity - (Position Diversity) ²	- (Position Diversity) ²	Since both of these variables have low importance scores, we remove the quadratic term to avoid multicollinearity issue
10	- Opp_conf_wl - Opp_conference_size - Opp_games_played	- Opp_conf_wl	“Opp_conf_wl” is removed because the similar overall winning ratio information is already embedded in “Current Season Winning % Difference”. Although “opp_conference_size” and “opp_games_played” are highly correlated, they are both kept in the final model because: <ul style="list-style-type: none"> • Their VIF values are small and not going to cause much multicollinearity issue • They are important proxy variables to the opponent’s experience in competitive basketball, hence we keep them to control for this factor

Table 4 Variables to be removed from Final Model

VIF calculation is again performed on the set of variables that are kept in the final model, and tabulated in Table 5.

Variable Name	Description	VIF
Game	Game Number	2.78
home game	Dummy indicating whether it is a home game	1.05
height diversity	Height diversity (inherent diversity)	8.21
(height diversity)²	Quadratic form of Height Diversity	7.78
ethnic diversity	Ethnic diversity (inherent diversity)	1.10
class diversity	Class diversity (inherent diversity)	1.11
school diversity	School diversity (acquired diversity)	1.32
degree diversity	Degree diversity (acquired diversity)	1.26
experience with opponent	Proxy for experience playing with opponent	2.92
opp_conference_size	Number of teams in the opponent’s conference. This could be a proxy of opponent’s experience. Opponent from larger conference may have more experience playing competitively	1.90
opp_games_played	Total number of games to be played by the opponent in this season. Similar to “opp_conference_size”, a higher number of games to be played may imply the opponent has more experience playing	2.10

	competitively	
Past Win Percentage Diff	Variable constructed in part 2. It is the winning % difference in the previous season and is a proxy for relative team strengths	2.29
Position Diversity	Variable constructed in part 2, to investigate whether a team with diverse playing positions would perform better.	1.25
streak	Variable transformed in part 2, investigate whether winning/losing momentum affects a team's future performance	1.37
streak ²	Quadratic form	1.07
Current Season Winning % Difference	Variable transformed in part 2, investigates whether the current season's performance (before the game) would predict the future game performance in the same season	3.53
Game * Current Season Winning % Difference	Interaction between "Game Number" and the current season winning percentage difference. Rationale was described in Table 1	3.35

Table 5 Final Model Variables and their VIF values

As observed from Table 5, all the VIF values are at an acceptable level now. The variables listed in Table 5 will be the subset of variables to be used in the final model.

Part 4: Final Model & Results Presentation

The selected variables are now used in the final logistic regression model. The data is fitted to a logistic regression model (with intercept) and result is summarised in Figure 5.

Logit Regression Results						
=====						
Dep. Variable:	win	No. Observations:	2334			
Model:	Logit	Df Residuals:	2316			
Method:	MLE	Df Model:	17			
Date:	Fri, 09 Jun 2017	Pseudo R-squ.:	0.1809			
Time:	08:37:20	Log-Likelihood:	-1324.3			
converged:	True	LL-Null:	-1616.6			
		LLR p-value:	2.399e-113			
=====						
	coef	std err	z	P> z	[95.0% Conf. Int.]	

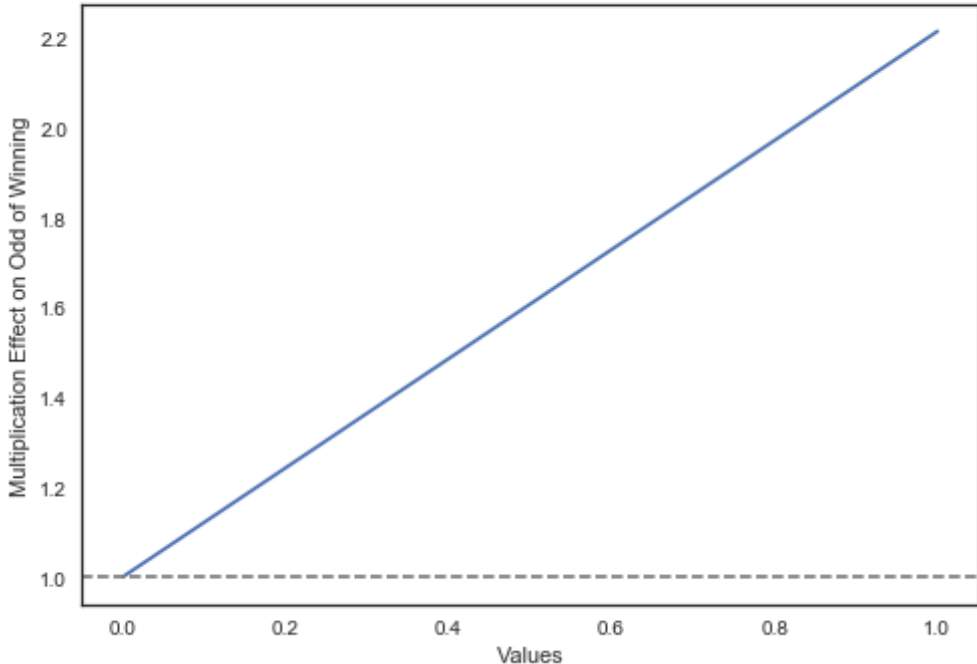
const	3.1262	0.853	3.664	0.000	1.454	4.799
Game	-0.0031	0.009	-0.342	0.733	-0.021	0.015
home game	0.7946	0.098	8.104	0.000	0.602	0.987
height diversity	2.3611	1.766	1.337	0.181	-1.100	5.822
height diversity sq	-15.6314	7.409	-2.110	0.035	-30.153	-1.110
ethnic diversity	-1.3356	0.359	-3.718	0.000	-2.040	-0.632
class diversity	-0.2981	0.195	-1.531	0.126	-0.680	0.084
school diversity	-0.5972	0.758	-0.788	0.431	-2.083	0.888
degree diversity	0.2024	0.358	0.565	0.572	-0.500	0.905
experience with opponent	-0.1379	0.157	-0.877	0.381	-0.446	0.170
opp_conference_size	-0.1390	0.025	-5.570	0.000	-0.188	-0.090
opp_games_played	-0.0209	0.012	-1.717	0.086	-0.045	0.003
Past Win Percentage Diff	2.1650	0.282	7.688	0.000	1.613	2.717
Position Diversity	-0.8825	1.133	-0.779	0.436	-3.102	1.337
streak	0.0130	0.019	0.682	0.495	-0.024	0.050
streak sq	-0.0027	0.003	-0.920	0.358	-0.008	0.003
Current Win Percentage Diff	0.0288	0.296	0.097	0.922	-0.551	0.609
Interact Curr Win Pct Diff	0.0862	0.021	4.158	0.000	0.046	0.127
=====						

Figure 5 Results of Final Logistic Regression Model

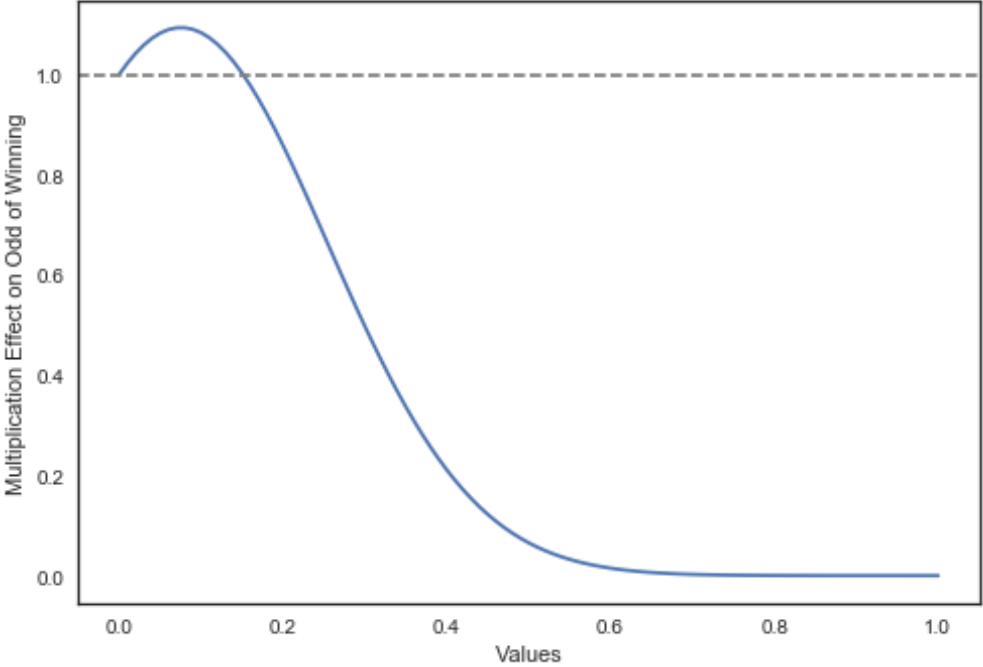
The in-sample accuracy score of this logistic regression model is 70.05%, using 0.5 as the threshold probability in estimating the game outcome.

Statistically Significant Variables

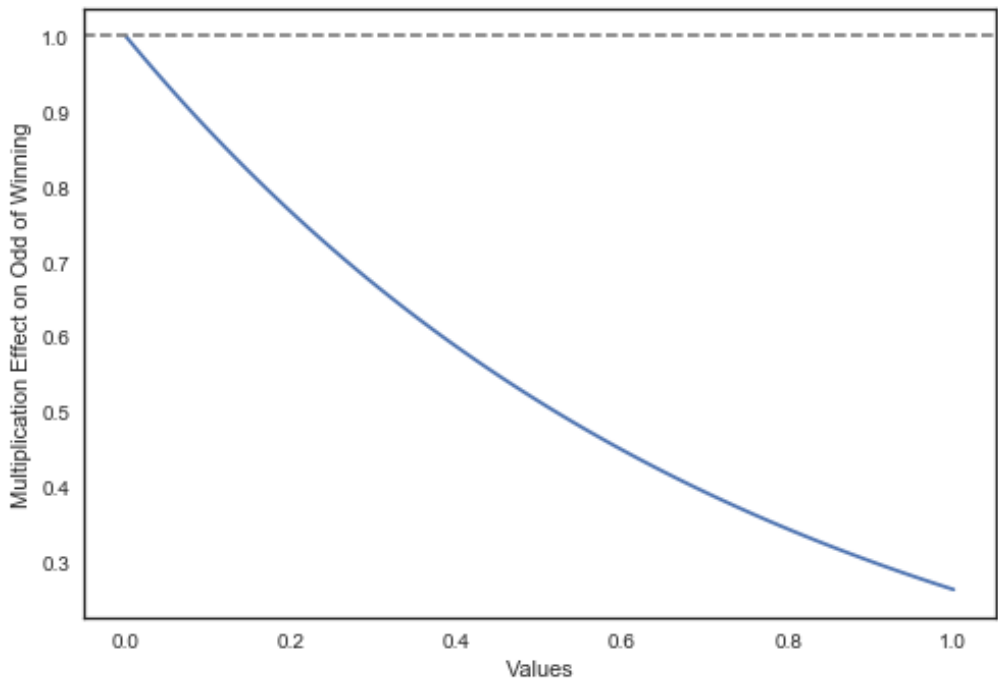
From Figure 5, it can be observed that a number of predictor variables are statistically significant (at 95% confidence level). Their effects of these significant variables are discussed in separate tables below.

Multiplication Effects of home_game on Odds of Winning	
Graph	<p style="text-align: center;">Effects of home_game on Odd of Winning</p>  <p>Note that “home_game” dummy variable only has two possible values (0 or 1). Therefore, playing at home (i.e. home_game = 1) will increase the odds of winning by factor of around 2.2.</p>
Incremental Effects Calculation	<p><i>Note: home_game is a dummy variable where value 1 indicates the team is playing at home ground.</i></p> <p><u>Coefficient</u> home_game = 0.7946</p> <p><u>Multiplication Effects on Odds of Winning the Match if playing at home</u> On average, the odds of winning will be $e^{(0.7946)(1)} = 2.21$ times higher than playing an away game.</p> <p><u>Interpretation</u> This result is not surprising because playing at home grants the team certain advantages such as familiarity with arena, fans support, etc.</p>

Multiplication Effects of Height Diversity on Odds of Winning

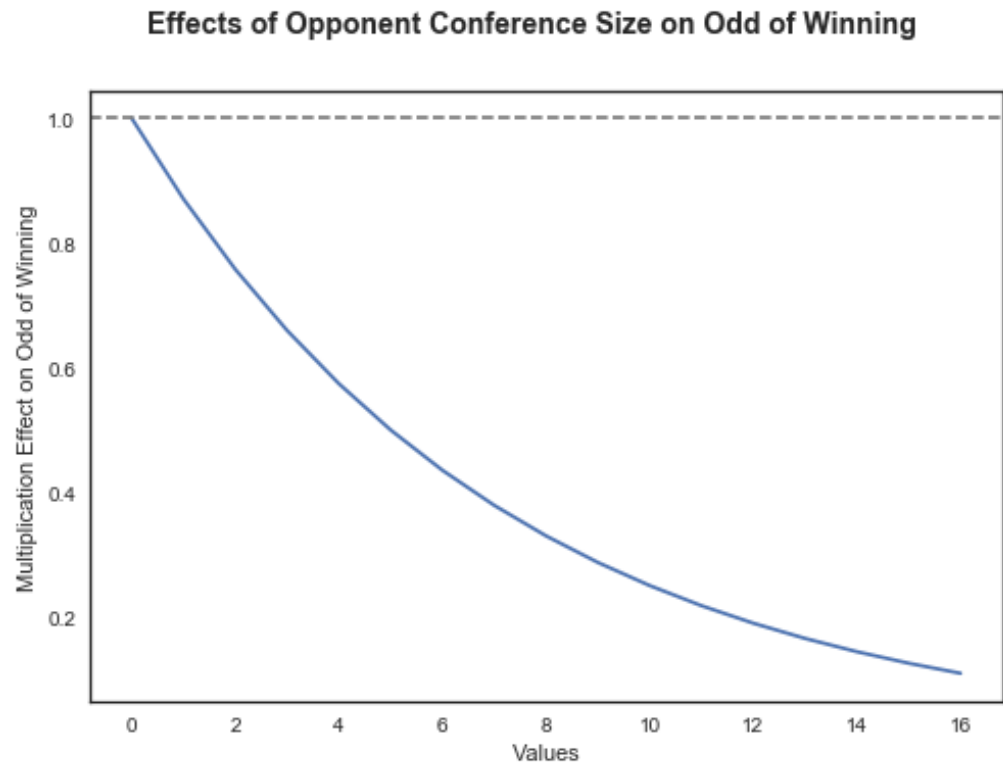
Graph	<p>Note: There are two "Height Diversity" terms, the linear term and the quadratic term. Although the linear term is not statistically significant, the quadratic term is significant at 95% confidence level. Thus, we consider the total effects (linear term + quadratic term)</p> <p style="text-align: center;">Effects of Height Diversity on Odd of Winning</p>  <p>The grey dashed line indicates the level where multiplication effect = 1. The curve that lies above this line implies that low amount of height diversity (between around 0 and 0.17) improves team performance. When the diversity increases beyond certain level (around 0.17), it becomes detrimental to team performance.</p>
Incremental Effects Calculation	<p><u>Coefficient</u> Height Diversity = 2.3611 (Height Diversity)² = -15.6314</p> <p>Since this variable involves quadratic term, we will need to assess the incremental effects using two fixed values of height diversity, hence we analyse the incremental effects of increasing height diversity 1 standard deviation above the mean.</p> <p><u>Distribution of Height Diversity Variable</u> Mean = 0.1018 Standard Deviation = 0.0774</p> <p><u>Multiplication Effects on Odds of Winning the Match if Height Diversity increases by 1 standard deviation above the mean</u></p> <ul style="list-style-type: none"> • Multiplication Effects at Mean $= e^{(2.3611)(0.1018) + (-15.6314)(0.1018)^2} = 1.0815$ • Multiplication Effects at (Mean + 1 Standard Deviation) $= e^{(2.3611)(0.1018+0.0774) + (-15.6314)(0.1018+0.0774)^2} = 0.9242$ • Ratio of the two Multiplication Effects $= \frac{0.9242}{1.0815} = 0.8545$ <p><u>Interpretation</u> As we can see, increasing the height diversity 1 standard deviation beyond the mean actually decreases the odds of winning by around 15%. This is expected because from the graph, we can</p>

	observe that the mean height diversity (around 0.1018) is nearly optimal in enhancing team performance. If it goes too far from this level, the positive effects start to reverse.
--	--

Multiplication Effects of Ethnic Diversity on Odds of Winning	
Graph	<p style="text-align: center;">Effects of Ethnic Diversity on Odd of Winning</p>  <p>The above graph is a monotonically decreasing function, this implies that increasing ethnic diversity is expected to decrease team performance</p>
Incremental Effects Calculation	<p><u>Coefficient</u> Ethnic Diversity = -1.3356</p> <p><u>Distribution of Ethnic Diversity Variable</u> Standard Deviation = 0.1366</p> <p><u>Multiplication Effects on Odds of Winning the Match if Ethnic Diversity increases by 1 standard deviation</u> On average, the odds of winning will be multiplied by a factor of $e^{(-1.3356)(0.1366)} = 0.8332$. This indicates a decrease in odds of winning by around 16.7% if ethnic diversity increases by 1 standard deviation.</p> <p><u>Interpretation</u> The result shows that ethnic diversity actually deteriorates the team performance. It seems that ethnic diversity brings relationship conflict and resulting in worse team performance in college basketball context.</p>

Multiplication Effects of Opp_Conference_Size on Odds of Winning

Graph



The above graph is a monotonically decreasing function, this implies that a team is less likely to defeat an opponent from larger conference size

Incremental
Effects
Calculation

Coefficient

Opp_Conference_Size = -0.1390

Distribution of Opponent Conference Size Variable

Standard Deviation = 2.83 (rounded to 3)

Multiplication Effects on Odds of Winning the Match if Opponent Conference Size increases by 3 (approximately 1 standard deviation)

On average, the odds of winning will be multiplied by a factor of $e^{(-0.1390)(3)} = 0.6590$. This indicates a decrease in odds of winning around 34.10% if the opponent's conference size increases by 3 teams.

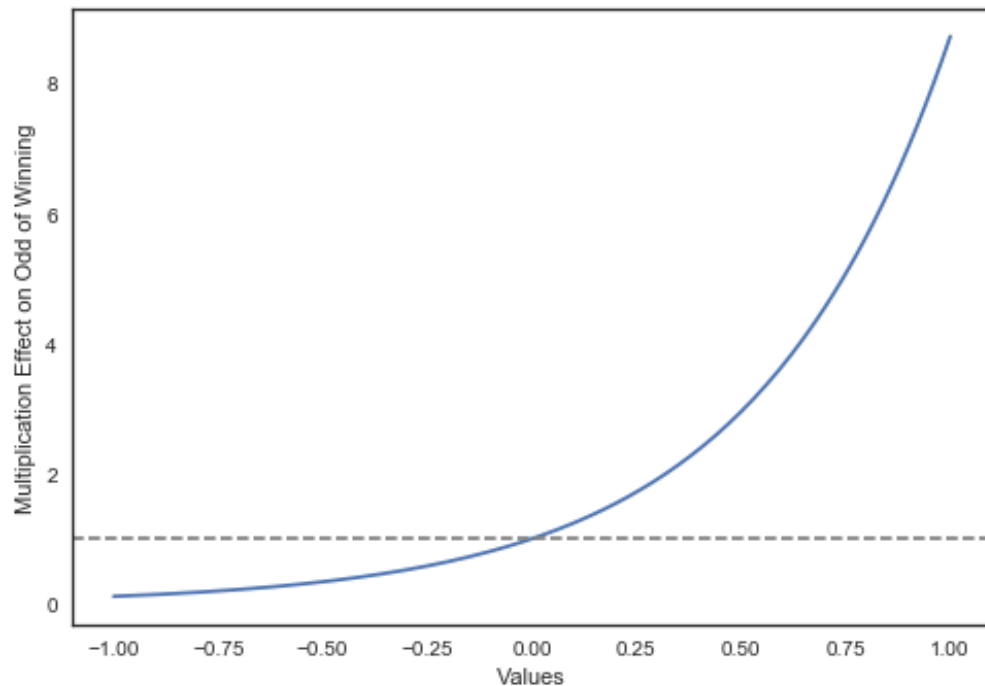
Interpretation

The result shows that the Ivy League team is less likely to win an opponent from larger conference size. Opponent's conference size is a proxy to the opponent's experience in playing competitive basketball. An opponent from larger conference is scheduled to play more conference matches. Consequently, the Ivy League team is less likely to defeat an experienced opponent.

Multiplication Effects of "Past Season Winning Percentage Difference" on Odds of Winning

Graph

Effects of Past Season Winning % Difference on Odd of Winning



The above graph is a monotonically increasing function and the multiplication effect is 1 when "Past Season Winning % Difference" is 0. This implies that a team is more likely to win if it has positive "Past Season Winning Percentage Difference" over its opponent. Otherwise, it is more likely to lose the game.

Incremental
Effects
Calculation

Coefficient

Past Season Winning % Difference = 2.1650

Distribution of "Past Season Winning % Difference" Variable

Standard Deviation = 0.2618

Multiplication Effects on Odds of Winning the Match if "Past Season Winning % Difference" increases by 1 standard deviation

On average, the odds of winning will be $e^{(2.1650)(0.2618)} = 1.7626$ times higher when "Past Season Winning % Difference" increases by 1 standard deviation.

Interpretation

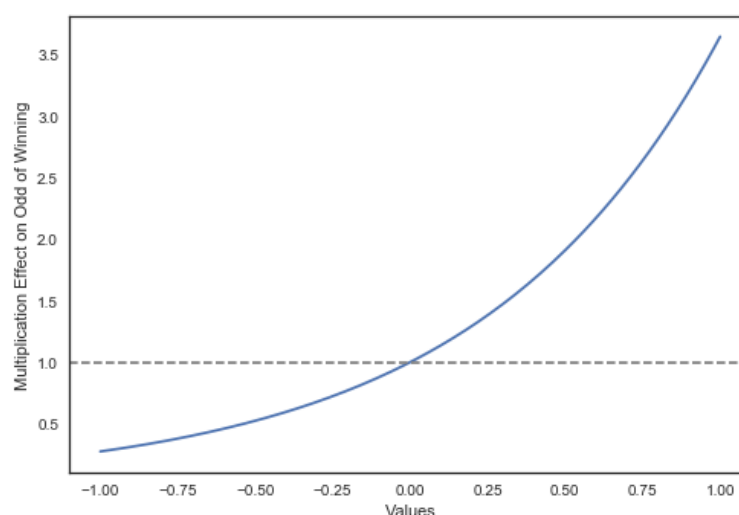
Unsurprisingly, this is one of the strongest predictors in the whole regression (the effect on odds of winning is around 76.26% higher per 1 standard deviation increase). This is because this variable measures the relative true team strength of both teams. A team can experience short-term setback in current season but the previous season team stats would be a good indicator of the team's true strength.

The result shows that if a team has positive "Past Season Winning % Difference" over its opponent, the odds of winning would be higher. This is expected since we expect an intrinsically stronger team would have higher chance of winning the match.

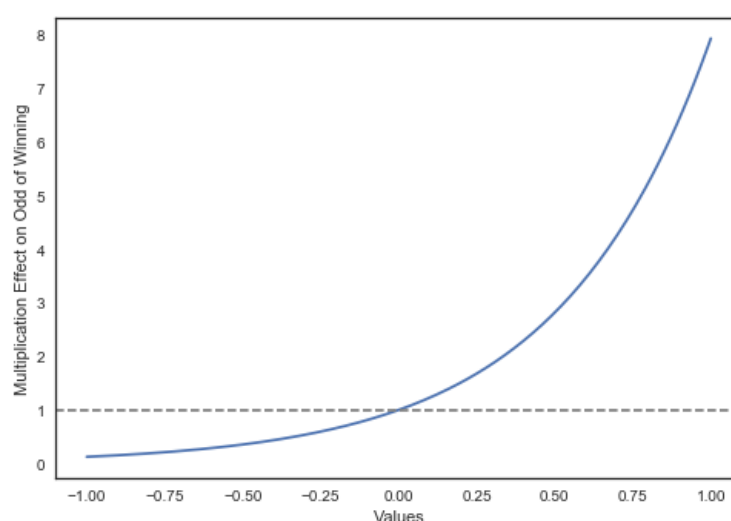
Multiplication Effects of “Game * Current Season Win % Difference” (Interaction between Game Number and Current Season Winning Percentage Difference) on Odds of Winning

Graph Since there is an interaction between “Game Number” and the “Current Season Winning Percentage Difference”, the multiplication effects will be different at different Game Number. To illustrate this, two graphs are plotted to visualise the effects of “Current Season Winning Percentage Difference” at Game = 15 (mid-season) & Game = 24 (late-season)

Effects of Current Season Winning % Difference (with Interaction, Game = 15) on Odd of Winning



Effects of Current Season Winning % Difference (with Interaction, Game = 24) on Odd of Winning



As expected, the graphs are monotonically increasing functions. Note that the multiplication effects get steeper when Game Number is higher (i.e. later in the season).

Incremental
Effects
Calculation

Coefficient
(Game * Current Season Winning % Difference) = 0.0862

Distribution of “Current Season Winning % Difference” Variable
Standard Deviation = 0.3057

Multiplication Effects on Odds of Winning the Match if “Current Season Winning % Difference” increases by 1 standard deviation

(When Game = 15)

	<p>On average, the odds of winning will be $e^{(0.0862)(15)(0.3057)} = 1.4848$ times higher when “Current Season Winning % Difference” increases by 1 standard deviation when Game = 15 (mid-season).</p> <p>(When Game = 24)</p> <p>On average, the odds of winning will be $e^{(0.0862)(24)(0.3057)} = 1.8822$ times higher when “Current Season Winning % Difference” increases by 1 standard deviation when Game = 24 (late-season).</p> <p><u>Interpretation</u></p> <p>Similar to “Past Season Win % Difference”, the variable “Current Season Win % Difference” quantifies teams’ past performance in current season. It is expected to be positively correlated with the outcome variable and the graphs show that this is indeed the case.</p> <p>The interesting part is its interaction with the Game Number. With a positive coefficient, this indicates that the effect of “Current Season Win % Difference” gets larger when we advance to later part of the NCAA basketball season. This verified our earlier hypotheses:</p> <ul style="list-style-type: none"> • Towards the end of the season, the strong teams would go all out defending their leading positions in the league table; hence the team strength factor might be more evident towards end of season. • The “Current Season Winning % Diff” is calculated using the winning ratio in the matches played <u>thus far</u> in current season. At the start of the season, there might be not enough games to establish an accurate estimate of the current season’s team strength. As the season advances, this estimate might grow more accurate and its effects become more evident.
--	---

Statistically Insignificant Variables

There are also a number of statistically insignificant (95% confidence level) predictor variables in the final logistic regression model:

Variable	Findings
Game	The data does not show significant difference in team performance at different stages of the season. However, “Game” does have an interesting interaction effect with “Current Season Winning % Difference” discussed earlier
Class Diversity	The regression result shows that class (i.e. freshman, sophomore, junior, or senior) diversity does not significantly influence Ivy League basketball team’s performance
School Diversity	The regression result shows that school (i.e. private or public) diversity does not significantly influence Ivy League basketball team’s performance
Degree Diversity	The regression result shows that the diversity in players’ degrees of study does not significantly influence Ivy League basketball team’s performance
Experience with Opponent	In general, the Ivy League basketball teams do not perform better against the opponent they have experience with. Having played more with the opponent also allows the opponent to study the Ivy League teams’ playing styles. Therefore, more experience with opponent does not necessarily translate to better team performance
Opp_games_played	Having controlled for ‘opp_conference_size’, the variable ‘opp_games_played’ becomes statistically insignificant. They are both proxy for opponent’s experience in playing in competitive basketball matches. It seems that ‘opp_conference_size’ is more informative in this regards, and the correlated cousin ‘opp_games_played’ becomes statistically insignificant after controlling

	for 'opp_conference_size'
Position Diversity	This is one of the variables constructed in Part 2, the data shows that playing position diversity (in the regular and reserve team) does not significantly influence the team performance
Streak / Streak²	This is one of the transformed variables used to measure the effects of winning/losing momentum. The regression results show that the momentum effects are not significant in influencing game outcome
Current Season Winning % Difference	After controlling for the interaction effect of "Game * Current Season Winning % Difference", this variable becomes not significant by itself. It can be said that the effect of this variable is concentrated in the interaction term

Table 6 List of Statistically Insignificant Variables in the Model

Note that although the above variables are not statistically significant, they are to be kept in the logistic regression model to act as control variables so that we can accurately assess the effects of the other significant predictors.

Conclusion

In this assignment, the Ivy League basketball team performance has been analysed and the factors that significantly influence team performance include whether it is a home game, height & ethnic diversity, the opponent's conference size (proxy for opponent's experience) and the difference in the two opposing teams' strengths.

It is also concluded that certain factors, which are intuitively believed to influence the game outcome, turns out being not statistically significant. These insignificant variables include winning/losing streak, experience playing with opponent, class/school/degree/position diversity, etc.