1.A.i

	Div	Y	HomeTeam	AwayTeam	home_agd	away_agd	agd_diff	home_games	away_games
5326	Ligue_1	14	Evian Thonon Gaillard	Paris SG	-3.5	1.000000	4.500000	2.0	2.0
7214	Serie_A	14	Sassuolo	Sampdoria	-3.5	1.000000	4.500000	2.0	2.0
6464	Ligue_1	17	Strasbourg	Lille	-4.0	0.078261	4.078261	1.0	115.0
1910	La_Liga	14	Cordoba	Celta	-2.0	2.000000	4.000000	1.0	1.0
1912	La_Liga	14	Elche	Granada	-3.0	1.000000	4.000000	1.0	1.0
7197	Serie_A	14	Empoli	Roma	-2.0	2.000000	4.000000	1.0	1.0
7212	Serie_A	14	Palermo	Inter	-0.5	3.500000	4.000000	2.0	2.0

1.A.ii

	Div	Y	HomeTeam	AwayTeam	home_agd	away_agd	agd_diff	home_games	away_games
2940	La_Liga	16	Granada	Barcelona	-0.875000	2.192308	3.067308	104.0	104.0
3393	La_Liga	17	Levante	Barcelona	-0.705357	2.140000	2.845357	112.0	150.0
3008	La_Liga	16	Granada	Real Madrid	-0.936937	1.900000	2.836937	111.0	110.0
3293	La_Liga	17	Las Palmas	Barcelona	-0.623762	2.208633	2.832395	101.0	139.0
3370	La_Liga	17	La Coruna	Barcelona	-0.621622	2.142857	2.764479	148.0	147.0
2921	La_Liga	16	La Coruna	Barcelona	-0.519608	2.225490	2.745098	102.0	102.0
3190	La_Liga	17	Barcelona	La Coruna	2.186047	-0.527132	2.713178	129.0	129.0

1.A.iii

Strasbourg was a newly promoted team that had never played in Ligue 1 before. They lost their first game -4 so had a very low average goal difference.

1.B.i

Coefficient Value: -1.1669

1.B.ii

Brier Score: 0.2473

1.C

A logistic regression model only gives a value of 1 or 0. We would only see a 1 if the home team won every time. Since this is not the case, we get a negative intercept.

1.D

Coefficient Values:
- Intercept: -0.1791
- Historical Average Goal Differential (Home): 0.7853
- Historical Average Goal Differential (Away): -0.7619

Brier Score: 0.2173

2.A

Brier Score: 0.2140

2.B

Logit Model

2.C

The features used in the final model were:

- AGD_Home: Historical average goal differential by the home team
- AGD_Away: Historical average goal differential by the away team
- Form5g_Home: The goals scored by the home team in the last 5 games of the current season
- Form5g_Away: The goals scored by the away team in the last 5 games of the current season
- AxG_Home: Historical average expected goals for the home team
- xPts_season_Home: The points the home team should have based on results calculated via expected goals
- Rank_Home: The table position of the home team before the game started
- xRank_Away: What the table position of the away team should be based on expected points
- Reason_Home: A variable that outlines the teams that are in the bottom 4-6 spots in the table towards the end of the season

2.D

When starting the model building process, I tried to include statistics from both the entire dataset as well as just the season metrics. There's much season to season variability in leagues so historical data between seasons isn't always the most telling - for example Leicester City finished 14th in the premier league in the 14-15 season and won the premier league the next season. Additionally, there's a saying in fantasy premier league "form over fixtures" - which basically implies that a player's or a team's form can be an important metric to look at when deciding on a player or team.

I first started the model based on the homework question, only looking at home win through the average goal differential at home and average goal differential away. While our dataset included shots and shots on target, I felt that those weren't as telling as goals - a team can have many more shots but if they don't score, they won't win. I started including the form metrics as I felt that they were the most important to predict the next win and saw that the goal scoring form that looks at the last five games had the highest z-scores and included those in the model.

Next, I tried to breakdown the most important metrics for just the home team. These included the average expected goals which is a better indicator of how many goals a team should score over a season, the expected points in a season, which shows how many points a team should have based on the expected goals, and their current rank. I would assume that the teams with a higher rank in the table would generally be stronger and have a higher chance than the lower ranked teams. Additionally, I created a variable called "Reason" which signified teams that have a reason to play towards the end of the season. As the bottom three teams in top level leagues get relegated to the lower level competitions, I believed that those teams would be extra motivated to do well towards the end of the season and try and avoid relegation. Finally, I looked at the away team metrics and only included expected rank as that would should what the away team is actually playing like, irrespective of their position in the table.

Some interesting ideas that I tried to implement included forward selection and backward selection algorithms on Random Forest, Logistic Regressions, and SVMs. These didn't end up working in the end because they weren't able to identify a better subset of features than I had. These two algorithms only evaluate one feature at a time when adding to the set. A feature like "Reason" didn't seem to be useful via those algorithms but may have become more useful in tandem with some of the other features I included.