Dota 2 matches predictive analysis

Vladyslav Bilyk

April 25, 2020

Introduction

This research is based on a famous multiplayer online video game Dota 2. Now, I will briefly explain how the game works so that you can understand what I've been working on.
Dota 2 is a team-based game, in which 2 teams (Radiant and Dire) of 5 players compete against each other. Both teams have their own base located on either side of the map. During the game every player controls their own hero (there are more than 100 heroes in the game, each having their own unique abilities/skills), earning experience to level-up and gold for buying useful items to defeat and kill the enemy heroes. Match is won by a team which is the first to destroy the enemies' Ancient, which is the core of their base. Dota 2 is also the biggest eSports game, having more than 1000 professional players and 40 teams constantly competing with each other on a big stage. There are many various match types and game modes, but to simplify things I divided them into the following types: Ranked All Pick, Professional games, and other game modes. I already mentioned professional games, so I will proceed with the explanation of Ranked All Pick, which is the most popular among Dota 2 players. Basically, you just play in a match ,and afterward you earn a rating, based on whether you won or lost. Professional players are among the ones with the highest rating.
Since the release of Dota 2 in 2013, more than 5 billion matches were played. Unfortunately, it was impossible for me to analyze all of them, so I had to be satisfied with a little smaller sample.
As you will soon see in the next paragraphs, my research was focused on the analysis of matches with the available data, trying to find out how particular factors influence the outcome of the game. Apart from data analysis, I attempted to create a prediction model for match results. All coding was done using Python.

Data description

The cornerstone of my data is the dataset from Kaggle, containing many csv files with various information about each of 50 000 matches played in 2015. Amidst the match information this dataset offers there are:
- match outcome
- heroes chosen in a match
- every player's gold and experience earnings at each minute of a match
- chat messages in a match
- player's hero ability upgrades timings
- teamfights in a match.

Apart from that I also used the OpenDota API, which helped me obtain data on heroes' matchups — hero vs other heroes (games played, games won)

Example:

```
{"1": {
        "2": [26, 14],
        "3": [33, 12],
        "4": [24, 13]
}}
```

The biggest part of my work was exploring and transforming the data mentioned above.
Firstly, I extracted the data on hero picks (what heroes were chosen in a match) and adjusted them by matches' ids. Then, I took hero picks for each match and calculated the average win rate of each of the radiant heroes against the dire. Radiant heroes' ids are the first 5 numbers in a list.

[9]:

	match_id	heroes	rad_win_rate
0	0	[86, 51, 83, 11, 67, 106, 102, 46, 7, 73]	0.470
1	1	[7, 82, 71, 39, 21, 73, 22, 5, 67, 106]	0.497
2	2	[51, 109, 9, 41, 27, 38, 7, 10, 12, 85]	0.514
3	3	[50, 44, 32, 26, 39, 78, 19, 31, 40, 47]	0.489
4	4	[8, 39, 55, 87, 69, 101, 100, 22, 67, 21]	0.553

Next step was to transform the data on players' timings (each minute of the game). This file was quite big, having more than 1 million lines in total. After the reconstruction of that file I had 5 new separate files, each containing 10 000 matches with players' timings at 5th, 12th and 20th minute. To be specific, these timings are players' gold and experience earned at a certain time. Then, I summed up these values for every player and calculated Radiant's gold/experience advantage.

[16]:

	match_id	times	rad_gold_adv	rad_xp_adv
0	0	300	-1037	-90
1	0	720	-1124	176
2	0	1500	6316	10032

And lastly, I created a file containing information on the match winner adjusted by matches' ids.

[11]:

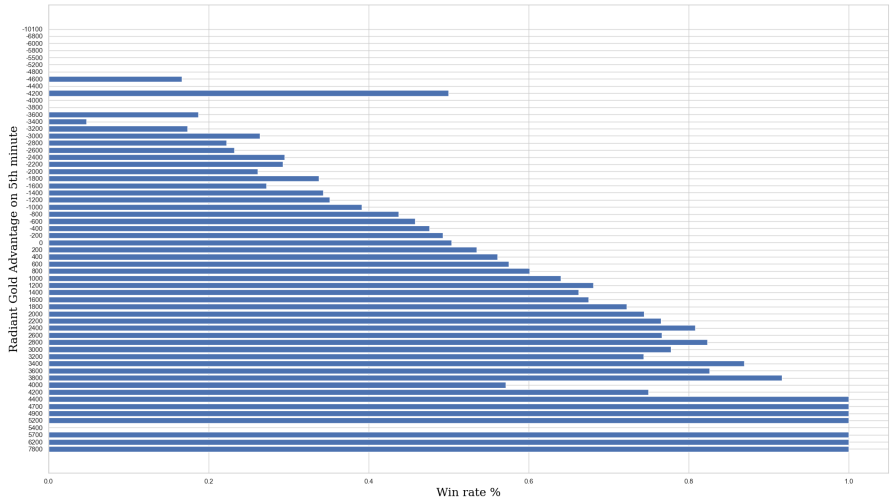
	match_id	radiant_win
0	0	True
1	1	False

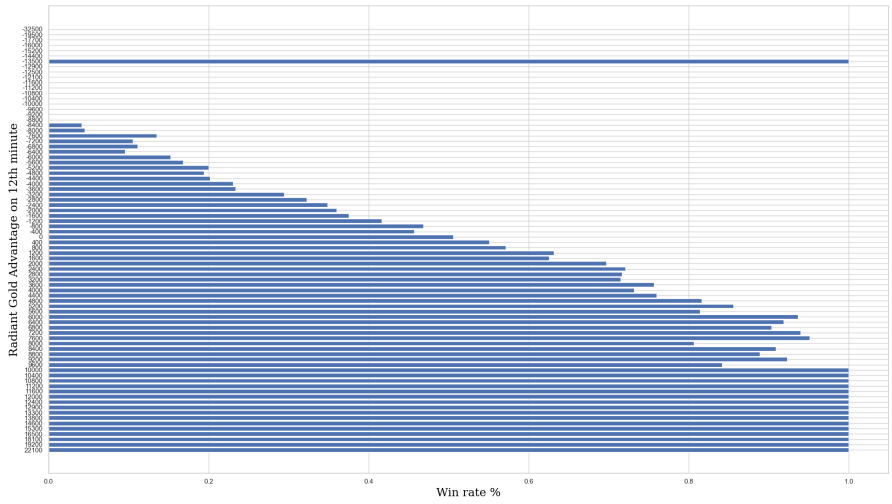
```
2          2         False
3          3         False
4          4          True
```

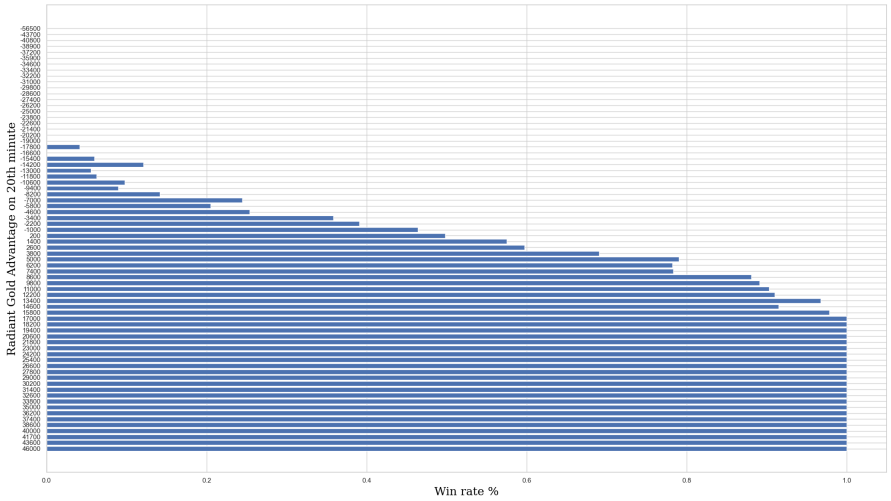
Data visualization

After having collected the data needed, I performed an analysis, in particular, on the above mentioned Radiant's gold/experience advantage to find out how much impact does this criteria has on the game outcome at 5th, 12th and 20th minute of a match.
Below are the obtained histograms. The order is: 5th, 12th, 20th minute advantage

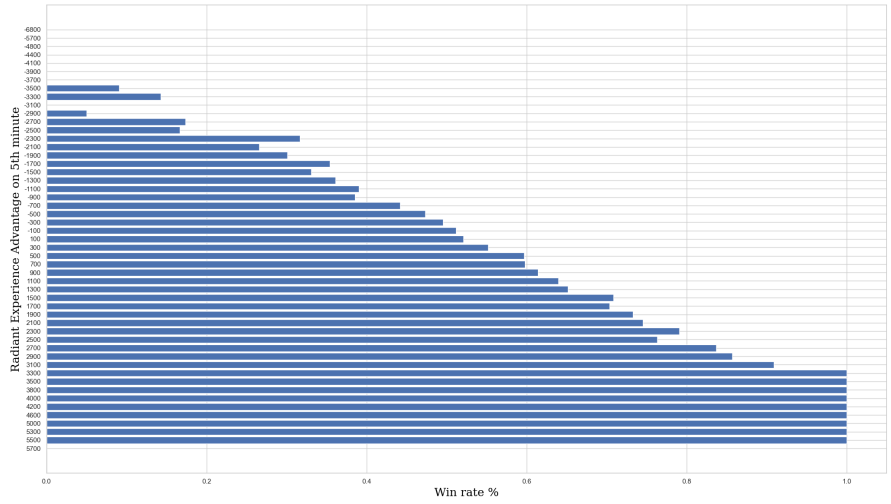
Radiant's gold advantage

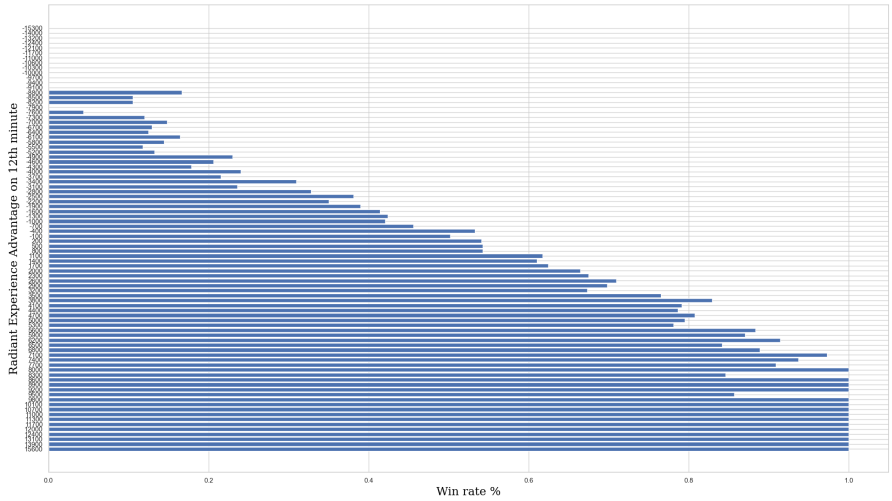


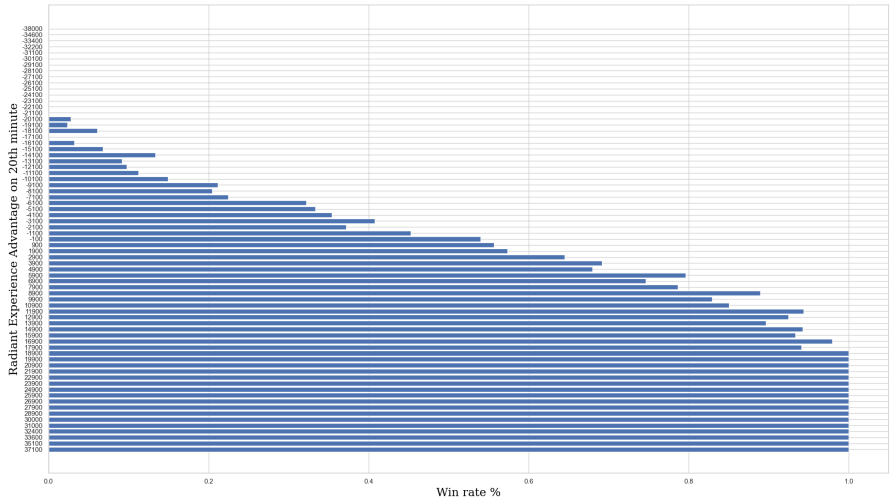




Radiant's experience advantage







Besides some outliers, it can be said that the bigger the advantage was the larger amount of games was won. This means that gold/experience advantage may be a decent predictor of the match outcome.

Methodology

As it was previously mentioned, during my research I was dealing with the match outcome, which is a binary variable in this case. If you look at the data I provided earlier, you will notice that there is specific data on whether Radiant won the match (TRUE / FALSE).

Thus, the logistic regression can be applied to create and train the model for predicting match outcomes.

The logit will look like this:

$rwin = \beta_0 + \beta_1 rga_t + \beta_2 rxpa_t + \beta_3 rwr$

rwin - 1 if Radiant won, 0 otherwise

rga_t - Radiant's gold advantage at time t

$rxpa_t$ - Radiant's experience advantage at time t

rwr - Radiant's heroes average win rate against the enemy heroes

In fact, there will be 3 separate models, since t is 5th, 12th or 20th minute of the game.

How it all works:

1. A function for reading data from csv and merging it is called.
2. The merged dataset is being passed to the logistic regression function
   (a) Model is created
   (b) The whole dataset is randomly split into train and test subsets (test subset = 30%)
   (c) Model is being trained
   (d) Making predictions with the model
   (e) Obtaining results of different forms

Results

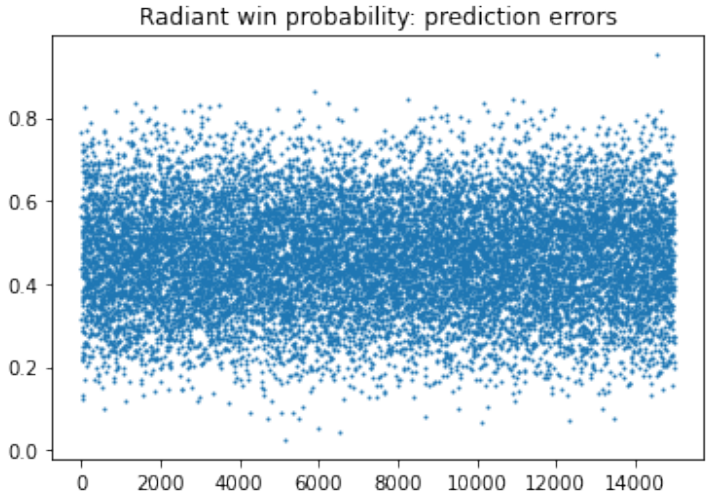
$t = 300$, Radiant's advantage at 5th minute

Results: Logit

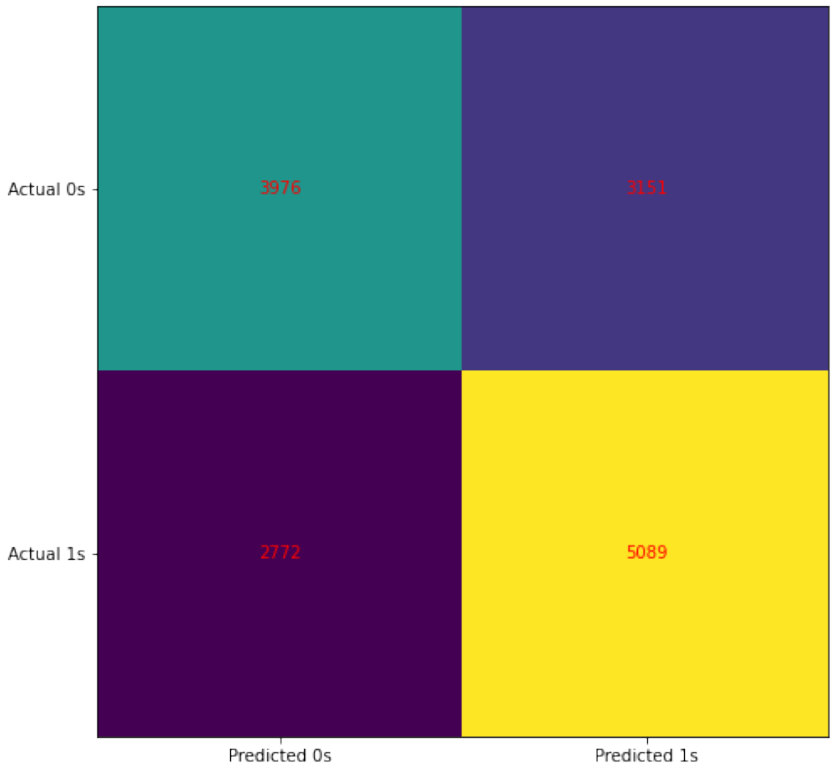
```
==================================================================
Model:              Logit            Pseudo R-squared: 0.052
Dependent Variable: radiant_win
No. Observations:   49959            Log-Likelihood:   -32808.
Df Model:           2                Df Residuals:       49956

------------------------------------------------------------------
                Coef.   Std.Err.     z     P>|z|   [0.025  0.975]
------------------------------------------------------------------
rad_gold_adv    0.0004    0.0000  32.9821  0.0000  0.0004  0.0004
rad_xp_adv      0.0001    0.0000   8.9565  0.0000  0.0001  0.0002
rad_win_rate    0.1783    0.0186   9.5942  0.0000  0.1419  0.2148
==================================================================
```

```
Accuracy on test set: 0.6048171870829997
```



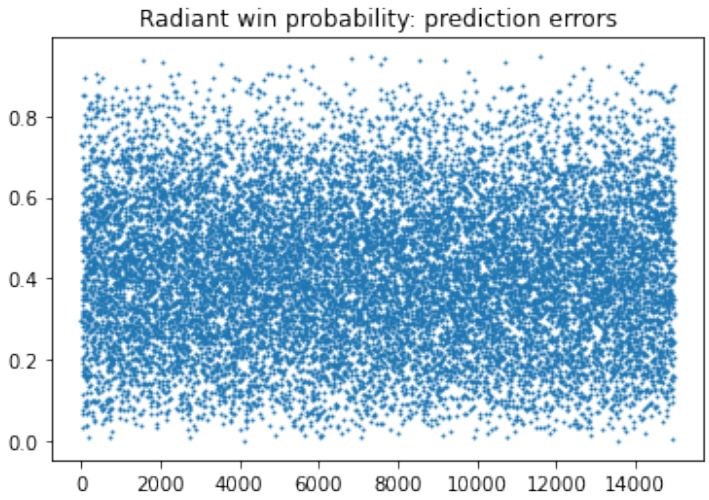

```
Predicted radiant win probability average error: 0.4646702805309118
```



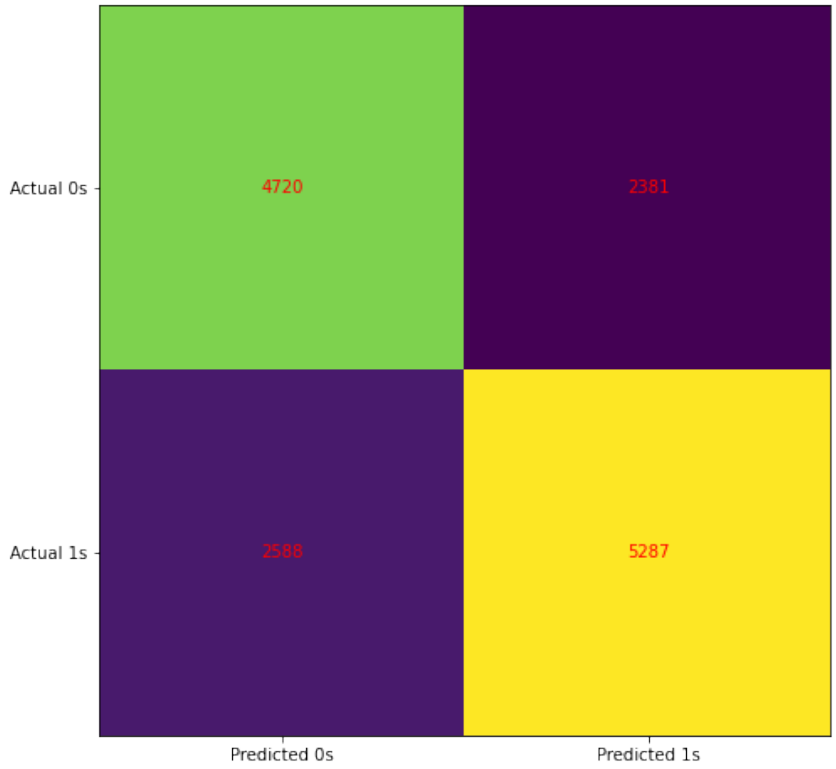
$t = 720$, Radiant's advantage at 12th minute

```
                         Results: Logit
================================================================
Model:              Logit            Pseudo R-squared: 0.127
Dependent Variable: radiant_win
No. Observations:   49917            Log-Likelihood:   -30175.
Df Model:           2            Df Residuals:       49914
----------------------------------------------------------------
                Coef.   Std.Err.     z     P>|z|   [0.025  0.975]
----------------------------------------------------------------
rad_gold_adv    0.0002    0.0000  33.3258  0.0000  0.0002  0.0002
rad_xp_adv      0.0001    0.0000  15.1879  0.0000  0.0001  0.0001
rad_win_rate    0.1286    0.0196   6.5698  0.0000  0.0903  0.1670
================================================================
```

```
Accuracy on test set: 0.6682024572649573
```



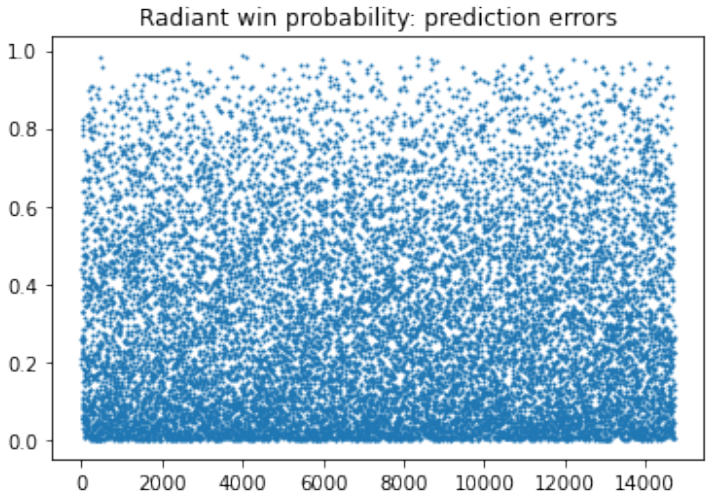

```
Predicted radiant win probability average error: 0.41800437608086927
```



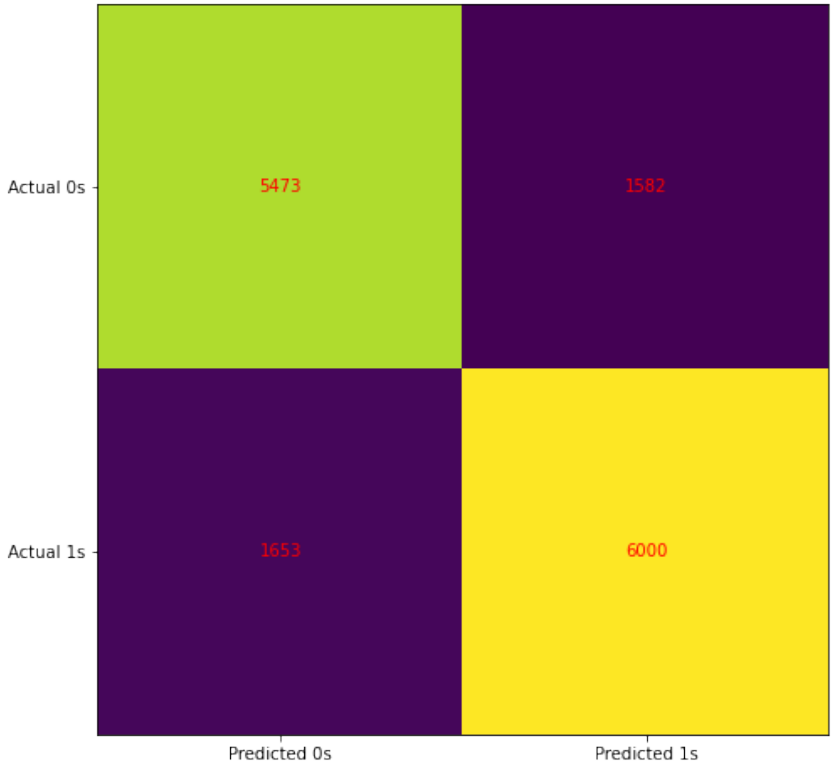
$t = 1500$, **Radiant's advantage at 20th minute**

```
                          Results: Logit
==================================================================
Model:              Logit            Pseudo R-squared: 0.351
Dependent Variable: radiant_win
No. Observations:   49026            Log-Likelihood:   -22038.
Df Model:           2          Df Residuals:       49023
------------------------------------------------------------------
                Coef.   Std.Err.     z     P>|z|   [0.025  0.975]
------------------------------------------------------------------
rad_gold_adv    0.0001    0.0000  33.0198  0.0000  0.0001  0.0001
rad_xp_adv      0.0001    0.0000  25.6805  0.0000  0.0001  0.0001
rad_win_rate    0.0959    0.0236   4.0636  0.0000  0.0496  0.1421
==================================================================

Accuracy on test set: 0.7800516725591515
```




```
Predicted radiant win probability average error: 0.295385347299292
```



Comments: It is quite obvious that as t increases (the deeper we go into the game and the less chances there are to make a comeback) the model becomes more accurate and the average error gets smaller. By looking at p-values it can be observed that all of the explanatory variables are statistically significant. Unfortunately, Pseudo R-squared is < 0.5 in all cases, which means that this model is far from perfect. Though it could be improved with the access to more data.

Conclusions and limitations

To sum up everything said and shown, in this particular topic the amount of data available plays a big role in the quality of predictive model. Not to mention the human factor, `Dota 2` is a pretty complicated game with numerous aspects in place, so the main limitation of this project was the lack of publicly available more detailed match data.