Final Project: Statistical Models
A World of Warcraft Battlegrounds
Exploratory Data Analysis

Federico Montes de Oca Rueda
Statistical Modeling

April 26, 2018

Abstract

This document is a report of the EDA(*Exploratory Data Analysis*) and Statistical Models that I created based on the dataset "World of Warcraft Battlegrounds: *Details of some battlegrounds in World of Warcraft*" by *Carlos Blesa* wich I found on his Kaggle profile.

1 Description of the project and dataset

World of Warcraft is a **MMORPG** (*massively multiplayer online role-playing game*) video game created by **Blizzard Entertainment** back in 2004. It has a lot of content, but for this project we are just focusing only in player vs player content (*Battlegroubds*).

Battlegrounds are scenarios where the two playable factions of the game (Horde and Alliance) fight against each other to win. Some of them are capture the flag them, others are take control points them, some of them are gain resources them, and some of them are a mixture of several themes. Playing battlegrounds means a lot of fun, but you we will get a currency in game called "Honor", which can be traded for powerful gear of our characters. If you wish to know more about this, you can visit the official website. Content

In the original dataset "World of Warcraft Battlegrounds: *Details of some battlegrounds in World of Warcraft*" (*wowbgs.csv*) by *Carlos Blesa* you will 3726 rows and 14 colms, were each row represents the information of a single player who played on a particular battleground game. Common statistics (*columns*) in all files are:(*Note: The numbers in parenthesis represent the number of observations given for each variable*)

Code: code for the battleground(3726).
Faction: faction of the player (*Horde(1875) or Alliance(1851)*).
Class: class of the player (*warrior(376), paladin(324), hunter(341), rogue(310),*

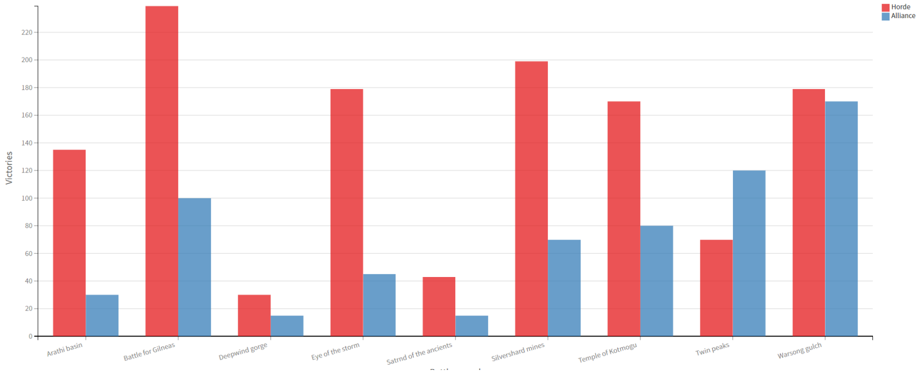
priest(303), death knight(227), shaman(396), mage(334), warlock(280), monk(167), druid(376), demon hunter(292)).
KB: number of mortal kills given by the player.(3726)
D: number of times that the player died.(3726)
HK: number of killings where the player or his/her group contributed.(3726)
DD: damage done by the player.(3726)
HD: healing done by the player.(3726)
Honor: honor awarded to the player. (3726)
Win: 1 if the player won.(1889)
Lose: 1 if the player lost.(1837)
Rol: dps if the player is a damage dealer(3000); heal if the player is focused in healing allies.(726) *Note that not all classes can be healers, just shaman, paladin, priest, monk and druid, but all classes can be damage dealers.*
BE: some weeks there is a bonus event, when the honor gained is increased. 1 if the battleground happened during that week.
Battleground: represent the kind of battleground: - *AB: Arathi basin. - BG: Battle for Gilneas. - DG: Deepwind gorge. - ES: Eye of the storm. - SA: Satrnd of the ancients. - SM: Silvershard mines. - TK: Temple of Kotmogu. - TP: Twin peaks. - WG: Warsong gulch*

With this given dataset we are interested in explore and see the different patterns and correlations that may influence the outcome of the battlegrounds. In this project we will focus in finding the right models to predict the outcome of almost each of our variables.
We are going to work with a supervised learning environment, as we are going to split the dataset into two one for training (*around 80% of the original dataset*) and one for testing(*around 20%*). As our variables are both numerical and categorical we are going to use both regression and classification methods.

2 Exploratory Data Analysis

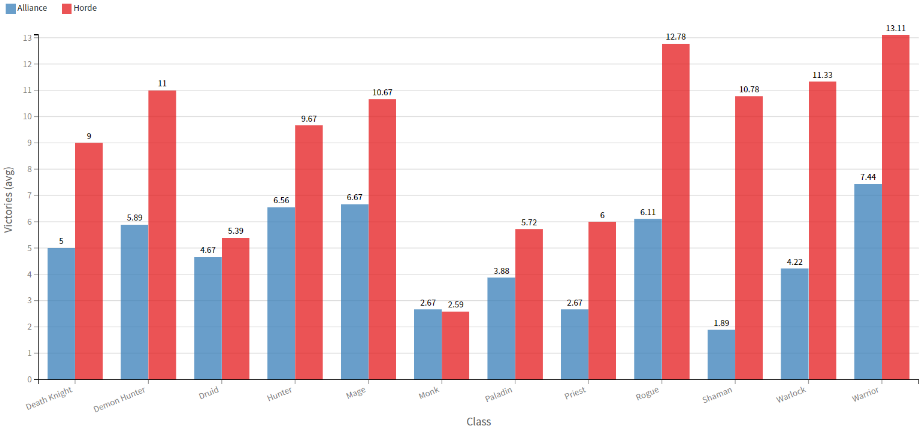
My first challenge was to determine how do I want to approach the dataset, which questions do I want to answer, and whether a variable is of meaningless or not for the analysis. With this in mind I decide that I don't care about either the code of the battleground nor the Honor gained, as I don't find of my interest to predict the amount of Honor that a player will won on a game. After that, I proceed to clean the dataset. I removed the columns: Code, Honer and BE. Then I changed the abbreviations in the Battleground columns to the actual name, which I found more useful. Finally I merge the Win and Lose columns into a single one called Win, which have either 1 (*for victory*) or 0 (*for deafeath*). With the new dataset (*wowbgsclean.csv*) I was able to performed an analysis on different charts comparing each of the variables.

Figure 1: Relationship between the *victories* and *battleground* for each *faction*



X-axis shows each one of the different types of *battleground*, Y-axis shows the total amount of *victories*, colors represent each *faction*(red for Horde, blue for Aliace)

Figure 2: Relationship between the *victories* and *battleground* for each *faction*



X-axis shows each one of the different of *classes*, Y-axis shows the average of *victories*, colors represent each *faction*(red for Horde, blue for Aliace)

3 Statistical Models

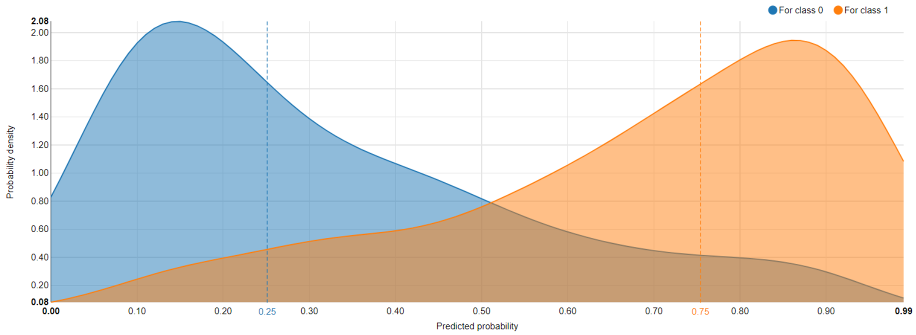
3.1 Predicting Victory using Logistic Regression

I used the Logistic Regression Algorithm to predict future game results (*victories*), aiming to have the best AUROC (Area Under the Receiver Operating Characteristic curve), without rejecting any of the feature(columns), and with a training set and test set that consist on 2994 and 732 observations respectively. The results with a AUROC of .85 (wich is great!—1=perfect model) were: That our model has a 78% **acurrancy** and a 79% **F1-score**(Harmonic mean between precision and recall.)

Feature	**Coefficient**
HK	0.091141761
Faction is Horde	0.981325001
Battleground is Satrnd of the ancients	-0.743215515
D	-0.3100825
Battleground is Battle for Gilneas	0.549240761
Battleground is Eye of the storm	-0.456341641
Battleground is Twin peaks	0.275488791
Battleground is Temple of Kotmogu	-0.274146785
Class is Mage	-0.233731978
Class is Priest	-0.222011967
Rol is dps	-0.194685956
Battleground is Arathi basin	0.192797702
HD	-4.97E-06
Battleground is Warsong gulch	0.185773973
DD	-5.22E-06
Class is Death Knight	0.15542441
Class is Hunter	-0.148550622
Battleground is Silvershard mines	0.102734003
Class is Shaman	0.091317225
Class is Demon Hunter	0.090266174
Class is Warrior	0.07883633
Class is Rogue	0.057039158
KB	0.009956534
Class is Paladin	0.030778919
Class is Druid	0.007635181
Class is Warlock	0.002577684
Intercept	-0.297448957

Table 1: Features that affect our model with their corresponding coefficients

Figure 3: Density char for our model



The density function of 0 should be entirely on the left
The density function of 1 should be entirely on the right

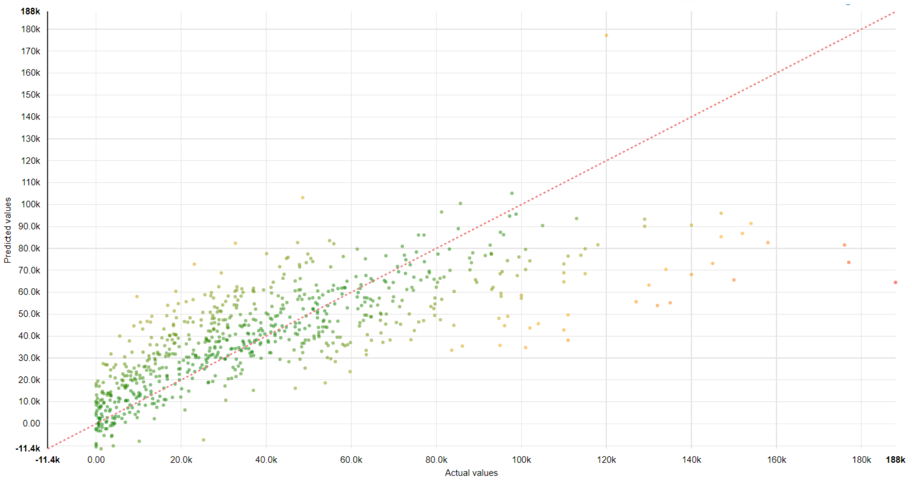
3.2 Predicting Damage done using Ridge (L2) Regression

I used the Ridge (L2) Regression Algorithm to predict the Damage Done(DD) for a player in a new game, aiming to have the best R2 Score (Coefficient of determination) regression score function, for this particular model I only take in consideration the *class* and the *KB* (Kills on Battleground) variables, rejecting all others (*Faction, Deaths, HK, HD, Rol*).
I used the same amount of observations for bouth the training(2994) and the test(732) sets that I used for the case above.
The result wasnt as good as I hope for, even with several attemps to train our model taking difrent considerations as training size and correlations the best model that I got had a R2=0.45218 a Mean Average Percentage Error= 301% and a MSE=5.943e+8!.
So I made several more atemps takinginto consideration the variables that I original decide to reject, for the best model I only rejected the *Faccion*. The result was a better model with a **R2**=0.571(still prety low) a **Mean Average Percentage Error**=148% and a **MSE**=4.65e+8.

Feature	**Coefficient**
Rol is dps	29143.49514
Class is Priest	7196.721158
Class is Mage	6526.28581
Class is Death Knight	6328.412272
Class is Rogue	-5737.195729
Class is Warrior	-4783.564104
Class is Hunter	4630.284165
Class is Warlock	4504.321274
D * HD (computed)	4152.031902
KB+HD (computed)	3439.24552
Class is Demon Hunter	3286.417466
KB	945.5707898
HD (computed)	2979.347384
KB+D (computed)	2874.048203
D * HK (computed)	-2688.971428
Class is Paladin	2586.930596
KB+HK (computed)	2421.692292
Class is Druid	2006.32052
KB-HK (computed)	1762.062891
D+HK (computed)	1627.603594
HK+HD (computed)	1627.397587
D+HD (computed)	1603.333589
Class is Shaman	1450.991398
KB * HD (computed)	1443.745413
KB-D (computed)	1397.7503
D (computed)	1307.573473
HD	0.034143499
HK * HD (computed)	-1295.684377
HK	75.41341256
HK (computed)	1206.771955
KB-HD (computed)	1068.862553
KB (computed)	929.8943624
D	422.6520403
D-HD (computed)	-263.618483
D-HK (computed)	-183.5019458
HK-HD (computed)	-79.37018355
KB * HK (computed)	58.75390712
KB * D (computed)	14.0060452
Intercept	15703.42558

Table 2: Features that affect our model with their corresponding coefficients The (computed) Features are computed thanks to Pairwise linear and Pairwise combinations that generate A+B,A-B and A*B features

Figure 4: Scatter Plot of our Model



The X-axis represent the Actual Values of our test dataset. The Y-axis represent the Predicted Value by the Model. Our model is represented by the red line

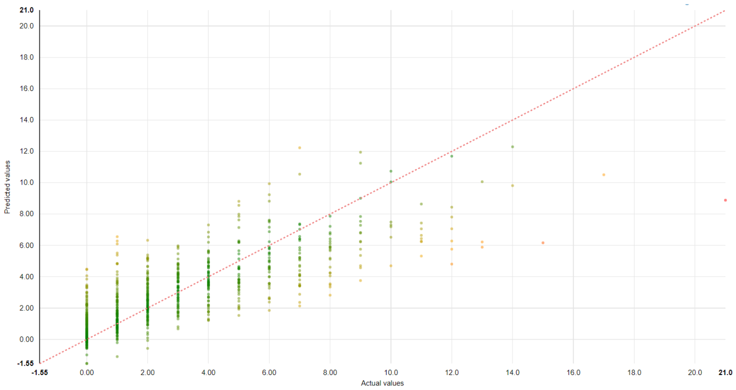
3.3 Predicting "Kills" using Lasso-Lars

Finally, I used Lasso-Lars Algorithm to predict how many kills would a player have on a particular game. Similar to the case above the test and training sets consist in 732 and 2994 observactions respectivly, only rejecting the *faccion*.
In this model we are also aiming to have the best R2 as possible, I got an above average result of **R2**=0.592 (wich is pretty okay). **Mean Average Percentage Error**= 55.0% and **Mean Squared Error**=3.9222.

Feature	Coefficient
DD+HK (computed)	4.46E-05
Class is Warrior	1.004407433
Class is Mage	-0.659812221
Class is Shaman	-0.546449648
Class is Druid	-0.53862003
Class is Hunter	-0.512431919
Class is Death Knight	-0.444899218
D-DD (computed)	-0.172415176
HK-HD (computed)	0.019756687
DD-HD (computed)	6.98E-06
HK * HD (computed)	-0.218023044
DD * HD (computed)	0.182335873
Class is Demon Hunter	-0.173597545
D * DD (computed)	-0.135051713
Class is Paladin	0.115822671
DD * HK (computed)	0.094210803
D * HK (computed)	0.088917752
D * HD (computed)	-0.086768836
Class is Priest	-0.053661924
Class is Rogue	-0.049326054
Class is Warlock	-0.026936212
Intercept	3.190988814

Table 3: Features that affect our model with their corresponding coefficients The (computed) Features are computed thanks to Pairwise linear and Pairwise combinations that generate A+B,A-B and A*B features

Figure 5: Scatter Plot of our Model



The X-axis represent the Actual Values of our test dataset. The Y-axis represent the Predicted Value by the Model. Our model is represented by the red line

4 Conclusion

After working on this models and other that I dint include do to the lack of good results with them, I learn how meticulus a datascientist should be while working on Statistical Models, and that there are always side to improve a model, a model should keep changing constantly as you are getting new data and changing your training set. In some of my models I realized that I was underfeeding my model and I needed to add more data to the training set or more features to the model. The model wich achive our goals the best was the Logistic Regression Algorithm.

5 References and Tools

https://www.kaggle.com/cblesa/world-of-warcraft-battlegrounds : Original dataset by Carlos Blesa
https://www.dataiku.com/ : Datascience tool, all the preprosessing and models were done on their IDE.