DATA MINING APPLICATIONS  
  
  
ALY6040, WINTER 2020  
MIDTERM ASSIGNMENT

SUBMITTED BY: ANUPREETA MISHRA, SHIVANI ADSAR, NIKHIL SAKINAL

NUID: 001050752, 001399374, 001376444

SUBMITTED TO: PROF. JUSTIN GROSZ

DATE: 03/11/2020

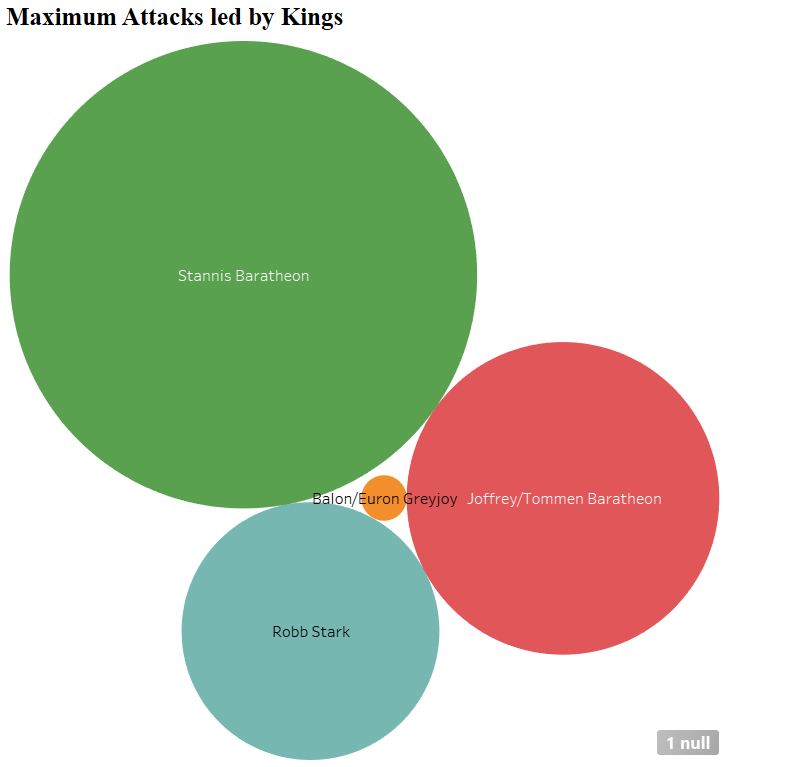
**Introduction**

In this assignment, we have used the “Game of Thrones” dataset for performing predictions on the death of kings. The dataset consists of the sources of data based on the information from the book series. Moreover, the “Battles” dataset gives insights about the kings who led the battles and their attacker army size. While, the “Character-Deaths” dataset gives information about the characters who died at various parts in the series. We have performed explanatory and exploratory analysis for better prediction about the deaths of kings in the battle using Tableau and R Studio.

**Analysis**

Part 1: Battles Dataset

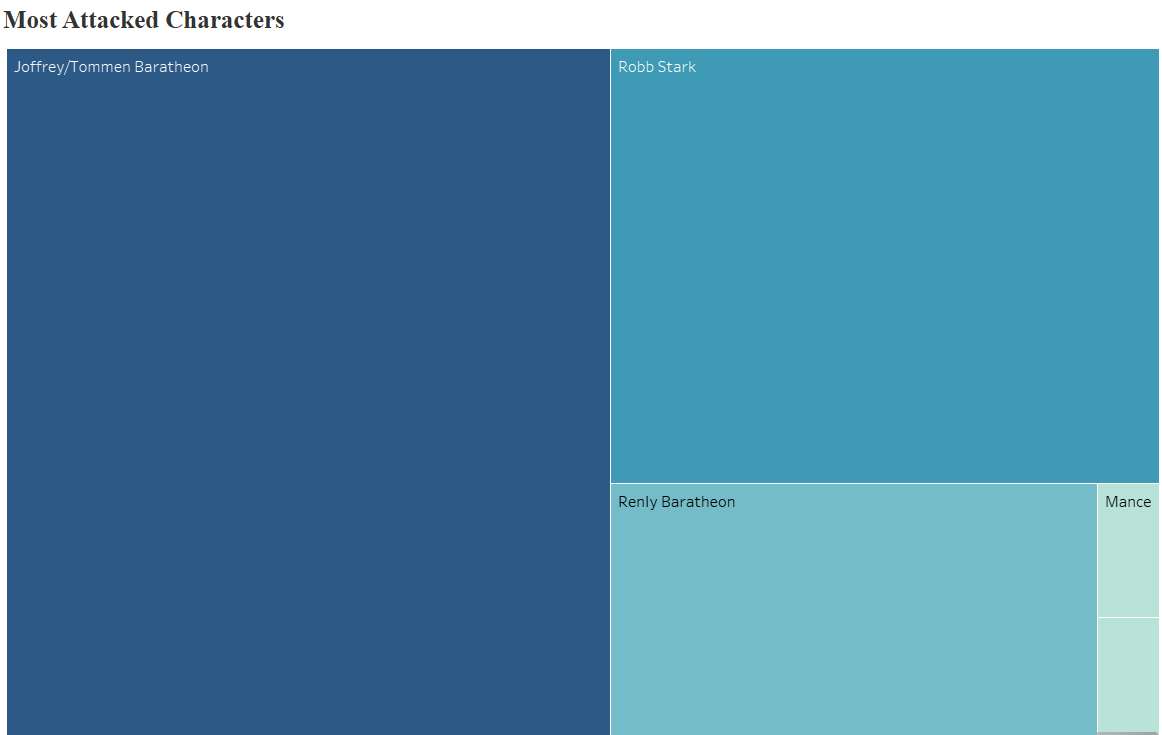
* We have performed visualizations using Packed Bubble Chart in Tableau which helps in understanding through the bubble sizes for getting to know inferences about the kings who lead the most attacks.



Observations: We can see that the highest number of attacks were lead by “Stannis Baratheon”. Whereas, “Balon/Euron Greyjoy” lead least number of attacks during the battle.

Fig.1: Packed Bubble Chart for Maximum Attacks led by Kings

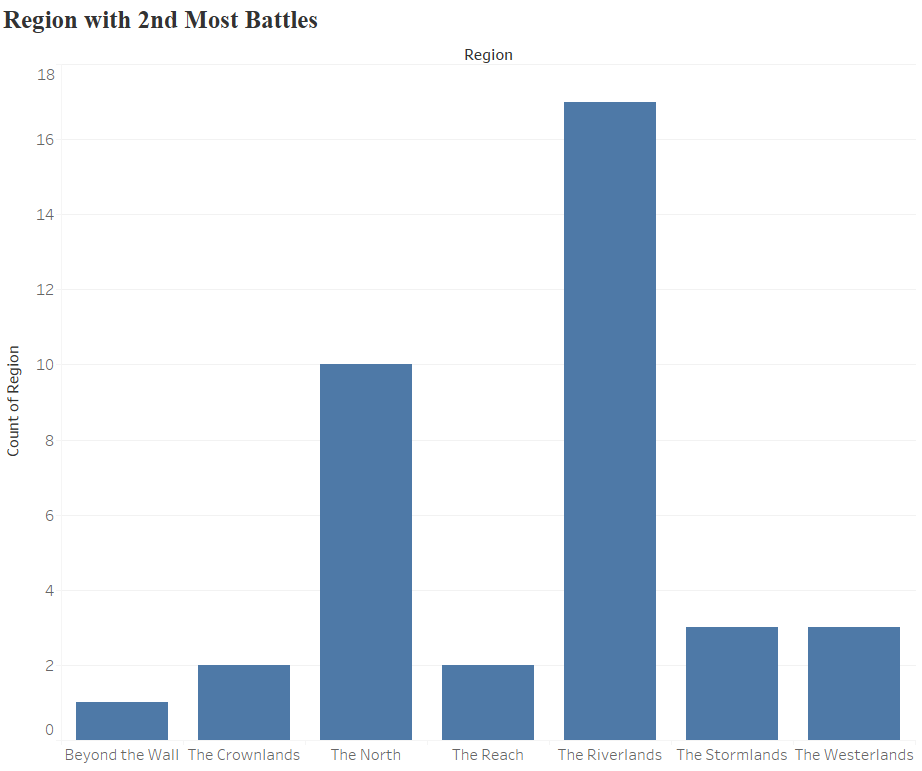
* Using Treemaps, we have visualized the highest number of attacked characters in series.



Observations: We can see that, “Joffrey / Tommen Baratheon” was the highly attacked character, while, “Balon/Euron Greyjoy” was the least attached character.

Fig.2: Treemap for Most Attacked Characters

* In order to get information about Regions with 2nd Most Battles, we have used a Histogram.



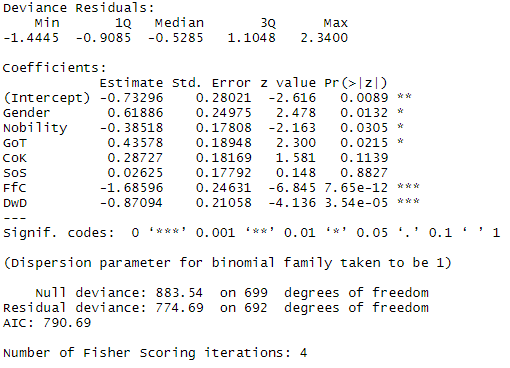
Observations: As we can see that, the Regions with 2nd Most Battles are “The Crownlands” and “The Reach” regions.

*Fig.3: Histogram for Regions with 2nd Most Battles*

* We can eliminate the null values in the “Attacker Size” and “Defender Size” columns by considering the ratios of attackers and defenders as per the “Battle Types” like pitched battle, ambush, siege, razing etc. and filling up the null values accordingly.

Part 2: Character Deaths Dataset

* Data Cleaning: Considering the “Death Year” column, we have performed categorization into 0 and 1, where the dead characters are replaced by 1 and the null values in that column are replaced by 0 for the characters that are alive.
* Logistic Regression: In order to perform better predictions, we have split the data into Train and Test models. Since we have to predict if the character was dead or not, we have used logistic regression on the cleaned column that shows if the character was alive or dead, using parameters like Gender, Nobility, GoT, CoK, SoS, FfC, DwD on the Train Data.



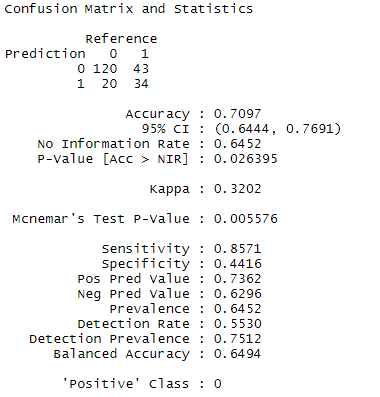
On performing Logistic Regression, we can infer form the Coefficients, the significant parameters are the books named, “Dance with Dragon” and “Feast of Ice and Fire” while the less significant parameters are Gender, Nobility and GoT.

Fig.4: Summary of Logistic Regression

Confusion Matrix Interpretations:

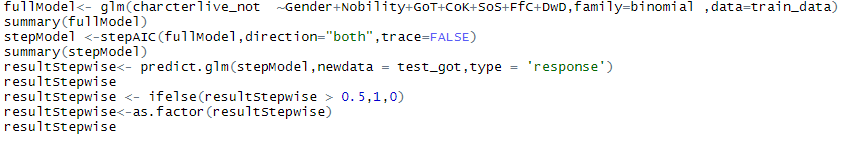
1. True Positive- Actual values were predicted correctly for 120 characters
2. False Negative- 43 characters who were not dead were predicted incorrectly
3. True Negative- 20 characters were correctly predicted to be alive
4. False Positive- It is observed that 34 characters were alive but the model predicts they were dead

The accuracy of the model is observed to be 70.97%



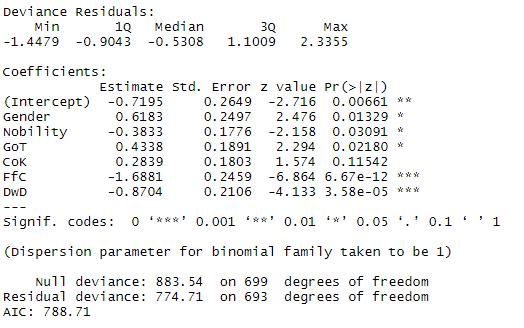
*Fig.5: Confusion Matrix*

* Optimisation & Comparisons: In order to perform further optimisation, we have performed the Stepwise Regression, which is a method of fitting the variables based on the Forward selection, Backward Elimination and Bidirectional elimination by testing variables at each step for variables to be included or excluded.



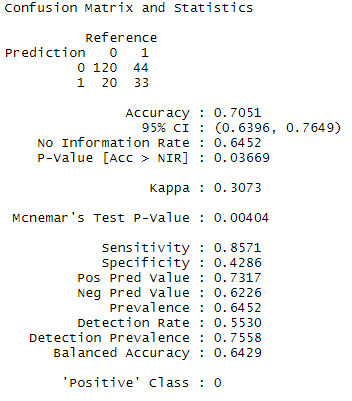
*Fig.6: R Code for Stepwise Regression*

* Considering, “Allegiances” as our dummy variable for better prediction, we have performed stepwise refression on the logistic regression of the model, this uses both the forward and backward selection methods for considering the optimised parameters for the model.



After performing the stepwise regression method, we can note that, the significant parameters remain constant. However, the AIC has reduced to 788.71 which shows that the model has improved, as lower the AIC, better is the prediction. The AIC of a model helps us to understand the error of the model and quality of samples with respect to other models.

*Fig.7: Summary of Stepwise Model*



We can compare that the AIC of the stepwise model has reduced to 788.71 from previous logistic model having AIC as 790.69. Hence, stepwise shows better prediction for the model due to lower AIC.

We can interpret from the confusion matrix, that the accuracy of the model has become 70.51%, and has not decreased drastically.

*Fig.8: Confusion Matrix for Stepwise Model*

**Conclusion**

* Age, heart disease, and glucose level are highly statistically significant factors which affect the stroke levels of patients.
* The accuracy of the model is 75.72% which summarizes the performance of the classification algorithm.
* Clearly, there are a lot of things left to analyze using algorithms like, the K-Means Clustering, Decision Trees and Random Forest

**References**

1. Eckerson, W. (n.d.). Secrets of Analytical Leaders. Retrieved from https://learning.oreilly.com/library/view/secrets-of-analytical/9781935504344/chap13.xhtml
2. The Internet Stroke Center. (n.d.). Retrieved from <http://www.strokecenter.org/patients/about-stroke/stroke-statistics/>