Using Pandas, Seaborn, and Scikit-learn's Machine learning tool to analyze Game of Thrones battle data.

FINAL PROJECT

TAHRIR IBRAQ SIDDIQUI

Introduction

Data Set

I have chosen to work with the data set containing information of battles fought in Game of Thrones taken from https://www.kaggle.com/mylesoneill/game-of-thrones/data. The data set, 'battles.csv', contains information on 38 battles separated into the following columns:

- name – Name of the battle.
- year – Year the battle was fought.
- battle_number – A unique ID number for the battle.
- attacker_king – The attacker's king. A slash indicates that the king changes in the course of the war i.e. "Joffrey/Tommen Baratheon" means Tommen took over from Joffrey.
- defender_king – The defender's king.
- attacker_1 – Major house attacking.
- attacker_2 – Second major house attacking. Similarly, attacker_3 and attacker_4.
- defender_1 – Major house defending.
- defender_2 – Second major house defending. Similarly, defender_3 and defender_4.
- attacker_outcome – The outcome from the perspective of the attacker.
- battle_type – A classification of the battle's primary type.
- major_death – If a major character died during the battle. 1 means yes and 0 means no.
- major_capture – If a major character was captured in the battle represented by 1 and 0.
- attacker_size – The size of the attacker's force.
- defender_size – The size of the defender's force.
- attacker_commander – Major commanders leading the attackers.
- defender_commander – Major commanders leading the defenders.
- summer – Was it summer?
- location – Location of the battle
- region – The region where the battle took place
- note – Notes regarding individual observations

Hypotheses

1. Lannisters always win pitched battles – The classification 'battle_type' has four categories and this hypothesis claims that Lannisters always end up in the winning side whenever the type of battle they fight is in the category 'pitched_battle'.
2. The region where the battle took place has a bigger influence in deciding the outcome of the attacking side than the type of battle. This test involves prediction of the outcome of the attacking side based on two features – 'region' and 'battle_type'.

Test Plan

Before carrying out any tests on the data, all the columns that are either irrelevant or unhelpful to both tests were dropped. The dropped columns and the reason for dropping them are as follows: name – irrelevant, year – irrelevant, battle_number – irrelevant, attacker_3 – too many missing values, attacker_4 – too many missing values, defender_2 – too many missing values, defender_3 – no values, defender_4 – no values, major_death – irrelevant, major_capture – irrelevant, attacker_commander – unhelpful because too many categories relative to the number of battles, defender_commander - unhelpful because too many categories relative to the number of battles, attacker_size – too many missing values which if replaced with 0 would be misleading, defender_size – too many missing values which if replaced with 0 would be misleading, location – too many categories relative to the number of battles, note – irrelevant.

Test 1

To test the first hypothesis, two catplot graphs were plotted. The first graph plotted the count of wins and losses of attacking houses in pitched battles, and the second graph plotted the count of wins and losses of defending houses in pitched battles. According to the hypothesis, Lannisters should have zero losses in both graphs.

Test 2

Part 1: To carry out tests for the second hypothesis, all columns with string values had to be converted to numerical values. This is because prediction algorithms and correlation maps can only process numerical data. After the conversion was done, a correlation heatmap was plotted with the following features included: attacker_king, defender_king, attacker_outcome, attacker_1, defender_1, battle_type, and region. According to the hypothesis, the correlation score between 'region' and 'attacker_outcome' should be higher than the correlation score between 'battle_type' and 'attacker_outcome'.

Part 2: Here, the hypothesis is tested using logistic regression algorithm. Since the feature being tested has binary outcomes, logistic regression was the most suitable algorithm to make predictions. First, 'battle_type' was dropped from the data set to predict attacker outcome based on region and other selected features. Then, 'region' was dropped from the data set to predict attacker outcome based on battle type and the same selected features as the previous prediction. Heat maps of the confusion matrix was shown in both predictions. According to the hypothesis, the accuracy score of the prediction should be lower when region is dropped than when battle type is dropped.

Results

Test 1

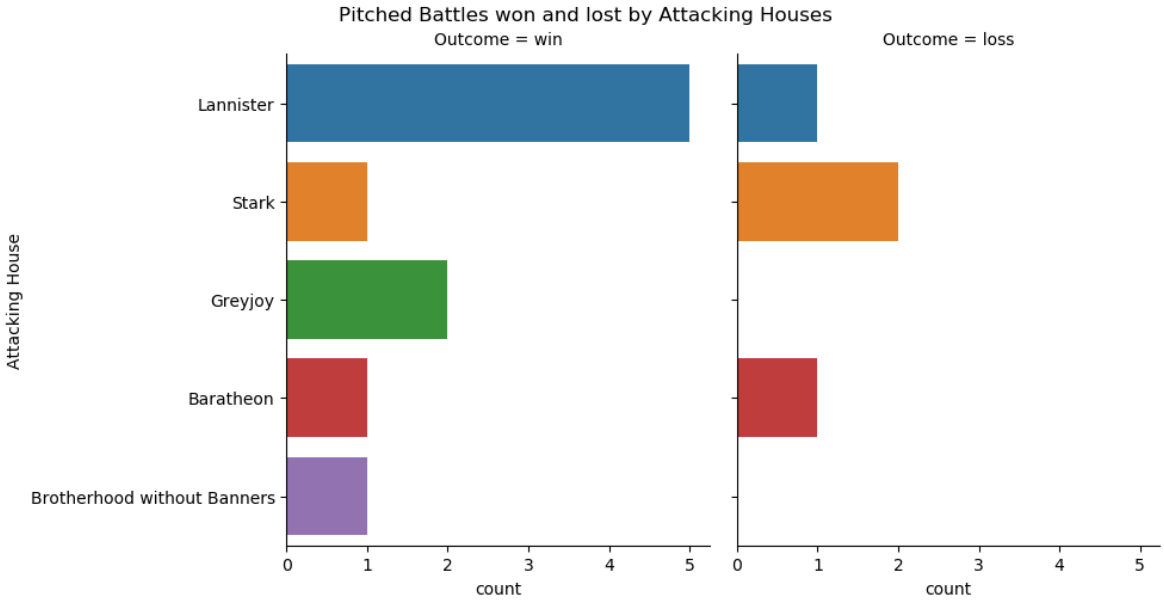

Figure 1: Catplot of Attacking Houses.

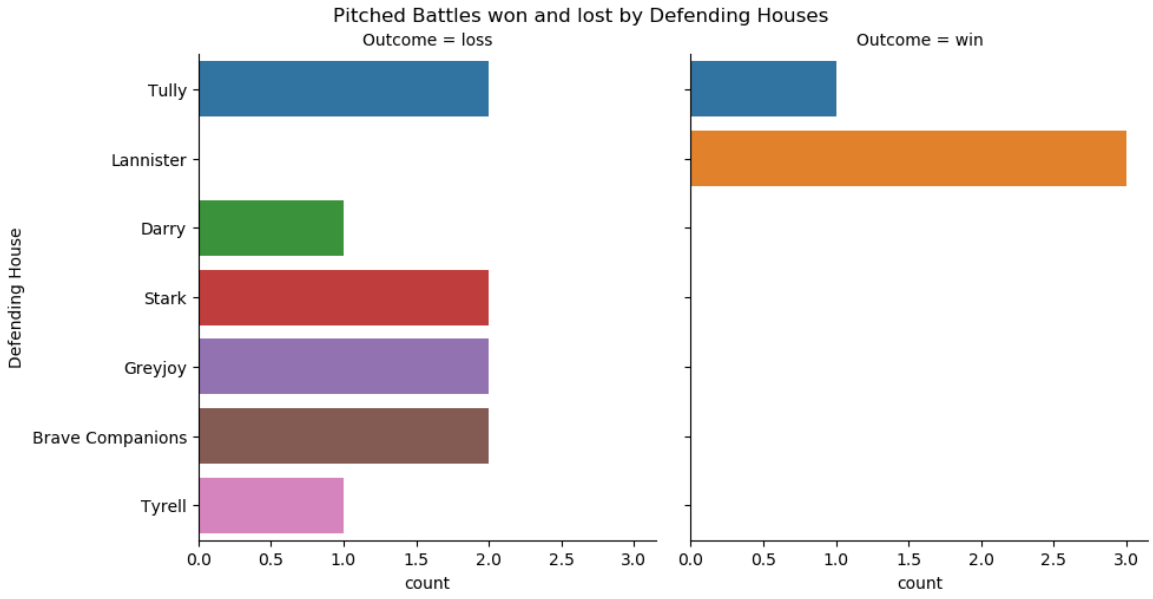

Figure 2: Catplot of Defending Houses.

Test 2

PART 1

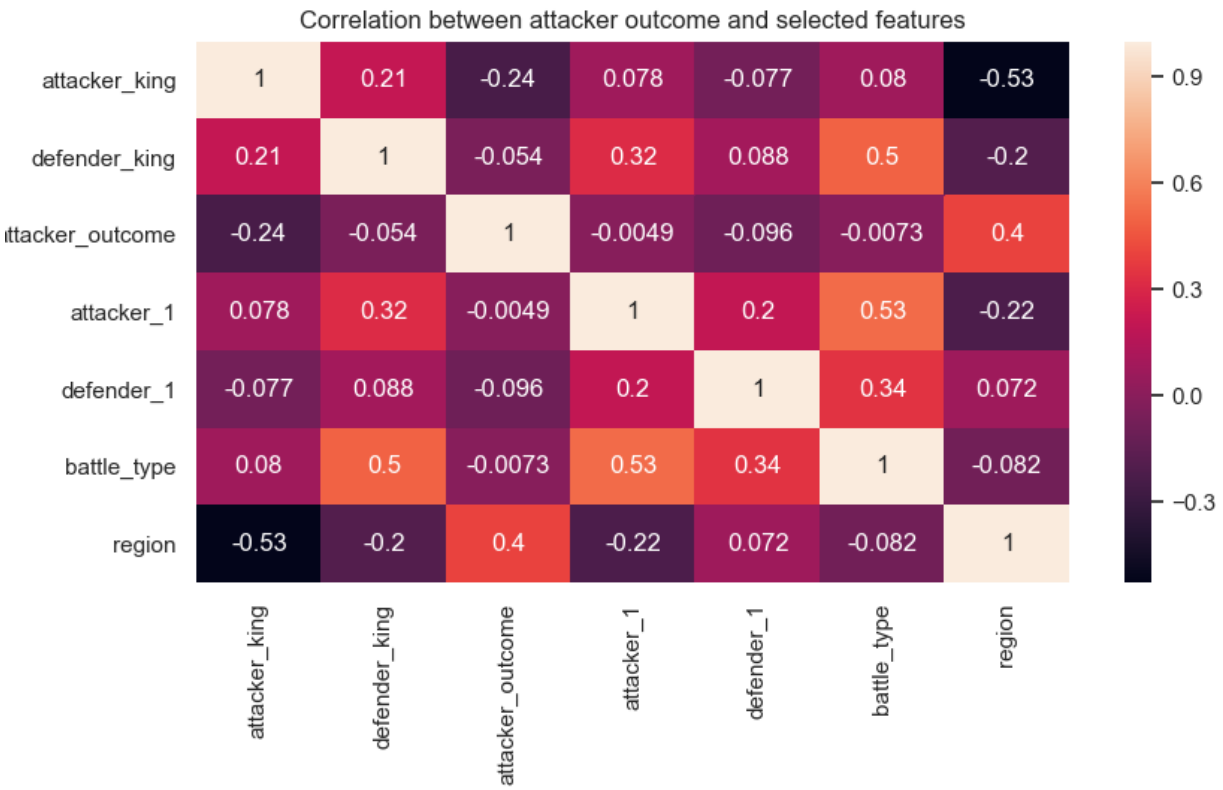

Figure 3: Correlation hetmap.

PART 2

```
Predicting attacker outcome based on region and selected features:
  FutureWarning)
The accuracy score is: 0.8666666666666667
              precision    recall  f1-score   support

           0       0.00      0.00      0.00         2
           1       0.87      1.00      0.93        13

   micro avg       0.87      0.87      0.87        15
   macro avg       0.43      0.50      0.46        15
weighted avg       0.75      0.87      0.80        15
```

Figure 4: Classification report of outcome predictions based on region and selected features.

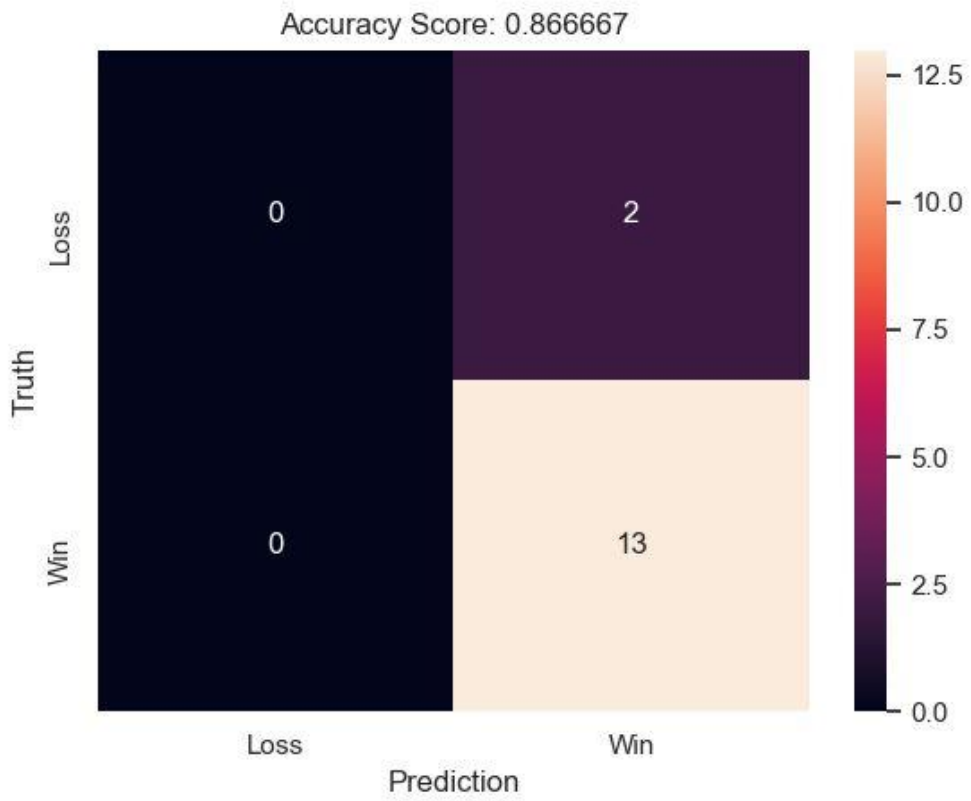

Figure 5: Confusion matrix of outcome predictions based on region and selected features

```
Predicting attacker outcome based on battle type and selected features:
  FutureWarning)
The accuracy score is: 0.8
              precision    recall  f1-score   support

           0       0.00      0.00      0.00         2
           1       0.86      0.92      0.89        13

   micro avg       0.80      0.80      0.80        15
   macro avg       0.43      0.46      0.44        15
weighted avg       0.74      0.80      0.77        15
```

Figure 6: Classification report of outcome predictions based on battle type and selected features.

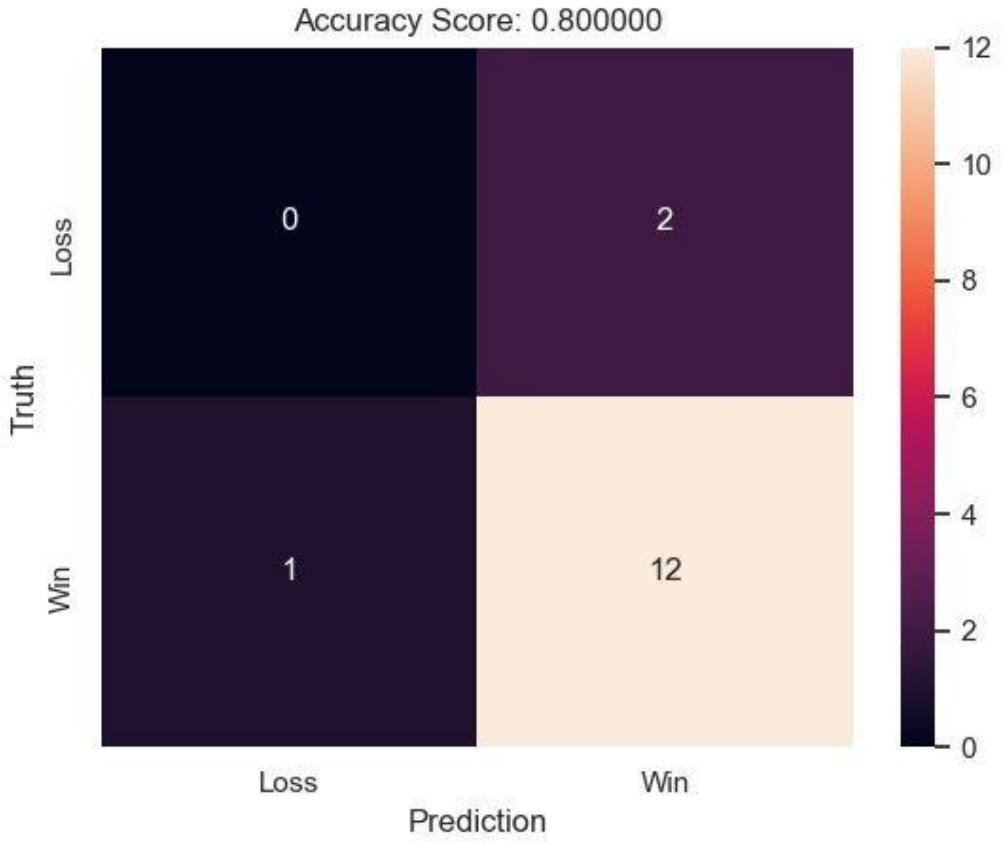

Figure 7: Confusion matrix of outcome predictions based on battle type and selected features

Analysis of results

Test 1

We can see from figure 1 that the Lannisters won the highest number of pitched battles among the attacking houses with 5 wins, whereas no other house won more than 2. However, the loss count shows that the Lannisters lost 1 pitched battle when attacking. Figure 2 shows that the Lannisters did not lose any pitched battles when defending and won 3. Although that single loss disproves the hypothesis that they always win pitched battles, the fact that they won 8 out of the 9 pitched battles they fought strongly testifies to their strength when it comes to this battle type. This is further emphasized by the fact that no other house comes close to this ratio.

Test 2

Part 1: From the correlation heatmap in figure 3, we can see that 'region' has the strongest correlation with 'attacker_outcome' with a score of 0.4 while the score with 'battle_type' is much lower at -0.0073. Thus, the hypothesis is supported by this graph. The other features also score much lower in comparison to region, with 'attacker_1' being the second highest at -0.24. One possible explanation is that the other features are a lot more interdependent in deciding the outcome than region is.

Part 2: As seen in the classification reports in figures 4 and 6, the attacker outcomes predicted by logistic regression with a test size of 0.4 gave an accuracy score of around 87% when based on region and selected features, and 80% when based on battle type and selected features. The confusion matrices in figures 5 and 7 show that the first test got 13/15 predictions correct while the second test got 12/15 predictions correct. Although a lower accuracy score for the second test supports the hypothesis, it is caused by a difference of just one prediction. Since the data only has 37 samples, and subsequently 15 test samples, a difference of one prediction represents a relatively large percentage difference of around 7%. Therefore, the prediction algorithm cannot support the hypothesis definitively.

It should also be noted that the accuracy scores vary with multiple runs and these are the results from one run. The accuracy scores for both tests were equal for many of the runs and in some rare cases, the score was greater for test 2. This is mainly due to two reasons:
1) The data set has a small number of samples.
2) 86% of the outcomes are wins, which makes it easier to predict correct wins when training both data sets.
The algorithm was throwing 'UndefinedMetricWarning' for the same reasons, despite stratifying the target data. However, the fact that most of the runs gave a higher accuracy score for predictions based on region does provide solid support in favor of the hypothesis.

CONCLUSION

Other than the correlation heatmap, none of my tests were able to prove the hypotheses conclusively. However, the results show strong support in favor of the relations being tested by both hypotheses. Maybe next time I'll test a feature with data that is not highly skewed towards one outcome so that a prediction algorithm can give more meaningful results. The biggest takeaway from this project is learning how to munge and process a data set to make it suitable for running tests that can reveal meaningful relations between features. Finally, it was nice to work with a data set that is relevant to the current hype, especially after the EPIC battle of Winterfell in the last Game of Thrones episode!