

Game of Thrones

Based on a given dataset giving information about approximately 2,000 characters of the TV series Game of Thrones the task was to predict if a character was dead or alive. Ideally the algorithm trained should be usable to predict the likelihood of death for future characters that are not yet part of the series.

The dataset on which this algorithm was to be trained included information about:

- Name
- Title, Culture and social Status
- Relatives and them being alive
- Appearances in the books behind Game of Thrones
- Birth year and age
- Popularity

It is important to mention that some of these features were not available for most of the character. This is most likely because of the origin of the dataset not from the author of the series but created by fans that only had the information given in the series which for most minor characters is not complete.

For characters that have information about birth year and age it is easy to add these two up which if this gives a date in the past is a clear indicator of the person being dead.

Unfortunately, this is not very useful for predicting the death of future characters as it is a retrospect indicator.

Another important indicator is a character's popularity. From the data dictionary it is not clear how the popularity is measured but clearly a very low popularity score makes a character likely to be alive.

The number of books in which a character is mentioned is another important factor.

Characters mentioned in three or four books are likely to be dead whereas characters mentioned in five books are likely to be alive.

Persons that have dead relatives are likely to be dead as well and this likelihood is greater the greater the number of dead relatives.

In general, being mentioned in the later books makes it more likely for a character to be still alive, which of course is not surprising.

There are many other insights that due to space constraints can't be mentioned but to generalize from the model the ideal set of features for a character with a high likelihood to be alive would be:

Being of House Stark, female, having a wife that is still alive and no dead relatives and being part of the latest book and ideally all others. It is important that the character is not popular. The age plays no role besides the fact that the character should not be older than the normal age span (up to 100 in GoT).

A character with highest chance of being deceased would have the following attributes:

Being of House Night's Watch, male, and having a large number of relatives deceased and holding the title of "Ser". The character should occur in the least number of books possible and be very popular.

The final model's highest mean AUC value after cross-validation, rounded to three decimal places (this being the H2O model I couldn't use sklearn for the cross validation but had to use H2O's integrated cross-validation – parameter: n-folds which I set to 5) was 0.900.

The highest mean AUC value after cross-validation for the XGBoost model was 0.895.