ASOIAF Survival Analysis

Christian Stone

May 13, 2021

Main Objective:

The fictional world of A Song of Ice and Fire (ASOIAF) is notorious for the large amount of characters in its books as well as the unexpected ways in which many characters are killed off. Because of this, the series is a great fit for a Survival Analysis of its characters.

As this analysis deals with data about a fictional universe, there aren't many useful business applications that this analysis can provide. The main stakeholders of this analysis would instead be fans of the ASOIAF series or people with an interest in the fantasy genre that the series falls in. While the series involves a lot of supernatural and fantasy elements, many elements of the story are also based off of historical events and settings and so it is possible that analysis could provide some insight to traits that may have improved the fates of people in feudal times.

The series has been popular since its debut in 1996, but within the last decade has become a cultural phenomenon due to its adaptation into the HBO series "Game of Thrones". While the TV adaptation has ended, the book series is still incomplete after its fifth installment was released in 2011 and will progress the story in a different way than in the show. There is a large community of people who discuss and theorize about the future of the series, and it is possible that these datasets could provide insight on the way author George R. R. Martin writes and may give an idea to whether the characters in his series will survive.

The goal of this report is to take this data and present its findings as if it were of scientific or business importance. Overall, however; this is just a fun experiment based on data from a beloved book series.

Data:

The dataset used in this analysis was created by Myles O'Neill and posted on kaggle.com. The set was one of three built from data from the ASOIAF series. He created datasets about the battles that occurred in the series and predictions for characters, but the data needed for Survival Analysis came from his character deaths dataset. Included in this dataset were a list of over 900 characters, their allegiances, the year, book, and chapter they died (if any), the chapter they were introduced, their gender, whether they were a noble or commoner, and finally what books they appeared in.

The three necessary parts of Survival Analysis data are having a time element, a survival indicator, and attributes which should correlate with survivability in some way. Since there is no precise timeline of events in the series, time was instead measured by the amount of chapters that had progressed in the story. Since the analysis dealt with the actual survival of characters, the survival indicator was simply an indicator of whether the character had died or not. The survivability attributes were additional information that the dataset kept track of such as nobility status, gender, and the major house or group that the character had allegiance to, if any.

Data Exploration and Analysis:

As is usually the case, the first step in this analysis was to clean and manipulate the data in a way that could be used to generate the models. The chapter data provided was given relative of the book that the character was introduced/killed in and so functions had to be created to convert this to a global chapter number. Once converted, the difference in the character's introduction chapter and the chapter of their death (if it had occurred) could be used to determine the character's "lifespan".

One important aspect of the data was to see how balanced it was over the different features. The breakdowns of each of the following features were measured: gender (82.65% male – 17.35% female), nobility (46.85% nobles – 53.15% commoners), status (67.29% alive – 32.71% dead), and allegiance (12.38 Night's Watch – 11.93% Stark – 11.05% Lannister – 8.29% Greyjoy – 6.85% Baratheon – 4.42% Wildling – 4.09% Martell – 3.87% Targaryen – 3.31% Tully – 2.87% Tyrell – 27.62% None).

A Kaplan-Meier plot was created to show how survival probability was related to chapter lifespan for each character:

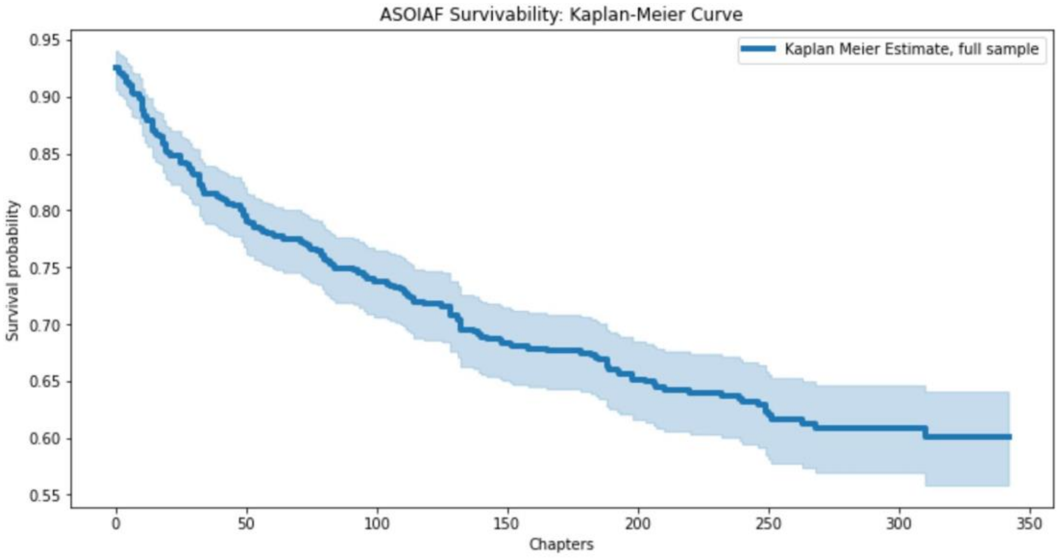
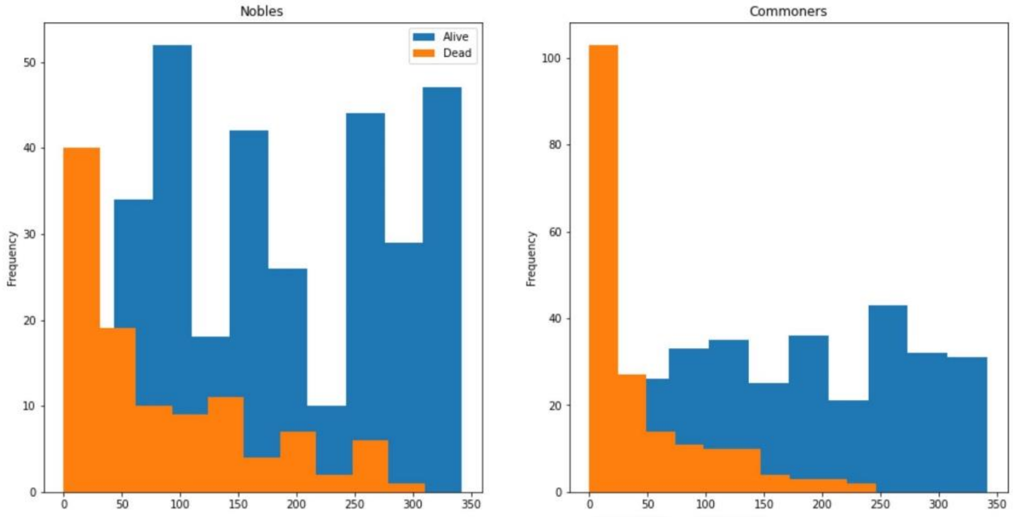


Figure 1: Kaplan-Meier plot showing survival probability at each "chapter lifespan" for all characters

A lot of different information can be determined from this plot. The shape shows that the longer a character is in the story, the more likely they are to survive longer. Normally, the curve should start at 100% survival probability, but it starts below 95% percent due to the amount of characters that die at time 0 (characters killed in the chapter they are introduced). The plot also shows that the confidence interval for survival probability increases over time. This makes sense as there will be less data as the chapter count increases since few characters in the series have lasted for 100s of chapters.

A Kaplan-Meier plot and histogram were created to show the effect that nobility could have on a characters chances of survival:



Figures 2 & 3: Histogram of character deaths of nobles and commoners

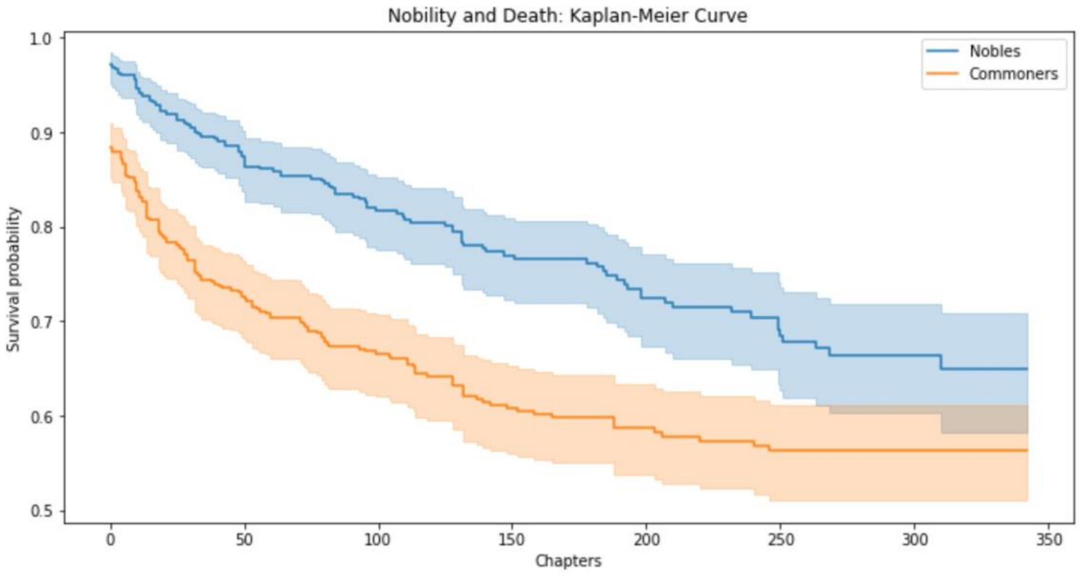


Figure 4: Kaplan-Meier curves for the survival probabilities for Nobles and Commoners

The statistic that stands out with the histogram is the likelihood for a commoner character to be killed off soon after being introduced. Otherwise, the likelihood that characters will die decreases as time goes on. The Kaplan-Meier curve shows that at any point in time in the story, a character of nobility has a greater chance of survival than a commoner.

Model Training:

3 different models were created in this analysis, each used a different set of data. The first model just used nobility status as the only attribute apart from lifespan and death condition. The 2nd model was the same as the first, but instead included the allegiance and gender attributes as well. The final model was similar to the first model, but included the data on the book in which the character was first introduced to the story.

Each of the three models was fitted to a simple Cox Proportional Hazards Model with a Breslow baseline estimation. 609 out of the total 905 character data points ended up being right-censored as these characters are still alive by the end of the fifth book in the series.

Recommended Model:

As all 3 models function very similarly to each other, the recommended model will depend on the information that is desired. The first model only features the nobility attribute, and so will be the most useful if this is the sole focus of analysis. The second model is less focused, but has more uses than the first. Due to having more feature sets, the second model can provide information on the impact of survivability on gender as well as a character's allegiance. The third model is similar to the second, except that it also provides information on when the character was introduced to the story.

Overall, the second model would be ideal for most cases. While the third model is similar, a few issues arise from the new data it provides. For one, the book that a character is introduced does not mean much within the universe of the story. Additionally, the data this provides is skewed as characters introduced later in the story will inherently be less likely to be killed as they will be involved in less situations which could result in their death.

Summary:

The main information derived from this analysis is information the factors that may impact the likelihood of a character dying at a certain point in the story. The following figure shows the data features and their effect on the likelihood of survival in the story:

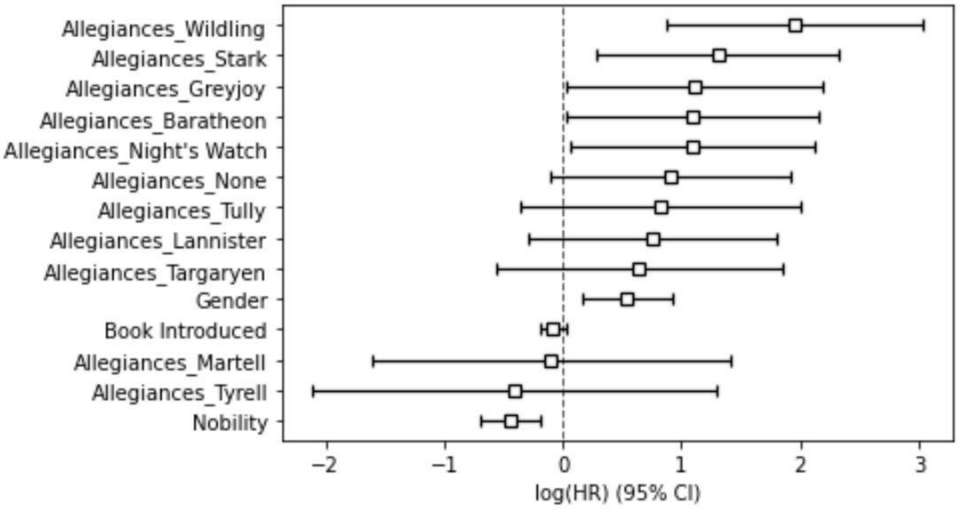


Figure 5: CPH plot showing the effect of each feature on survivability

The previous data analysis showed how having a higher status in society lead to a higher likelihood of survival, and this shows that it is the most significant stat in determining this.

A character's gender was shown to have a significant impact on a character's survivability:

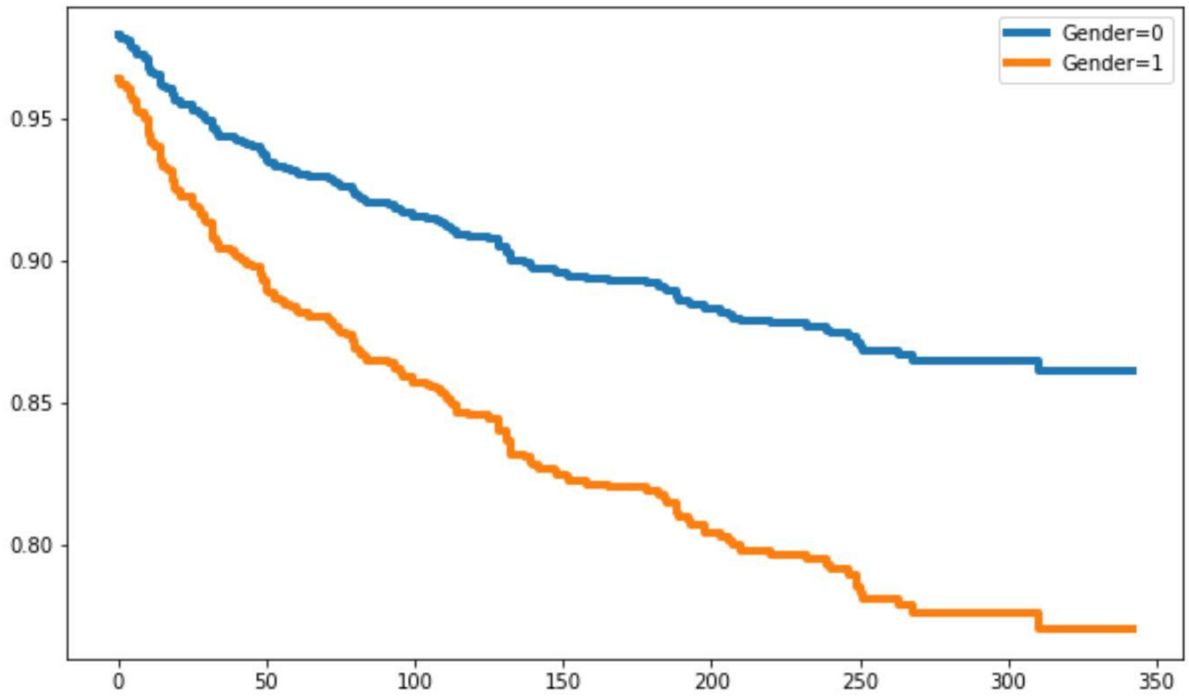
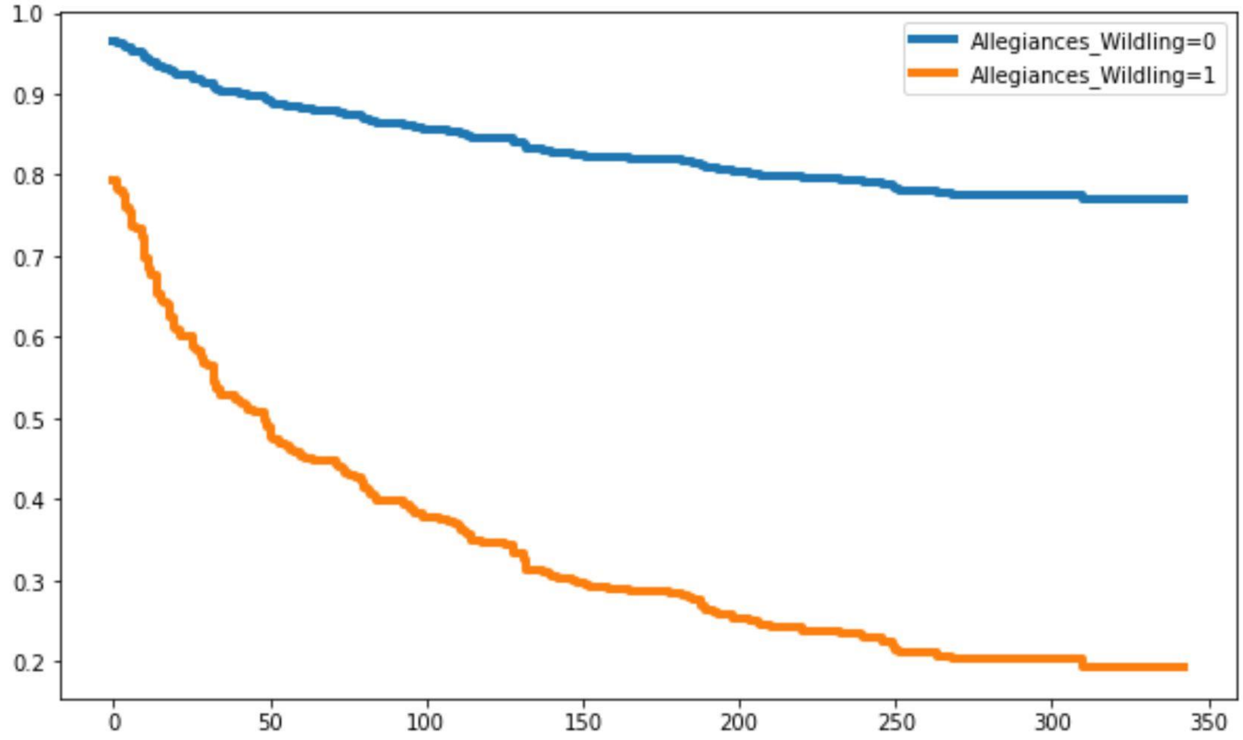
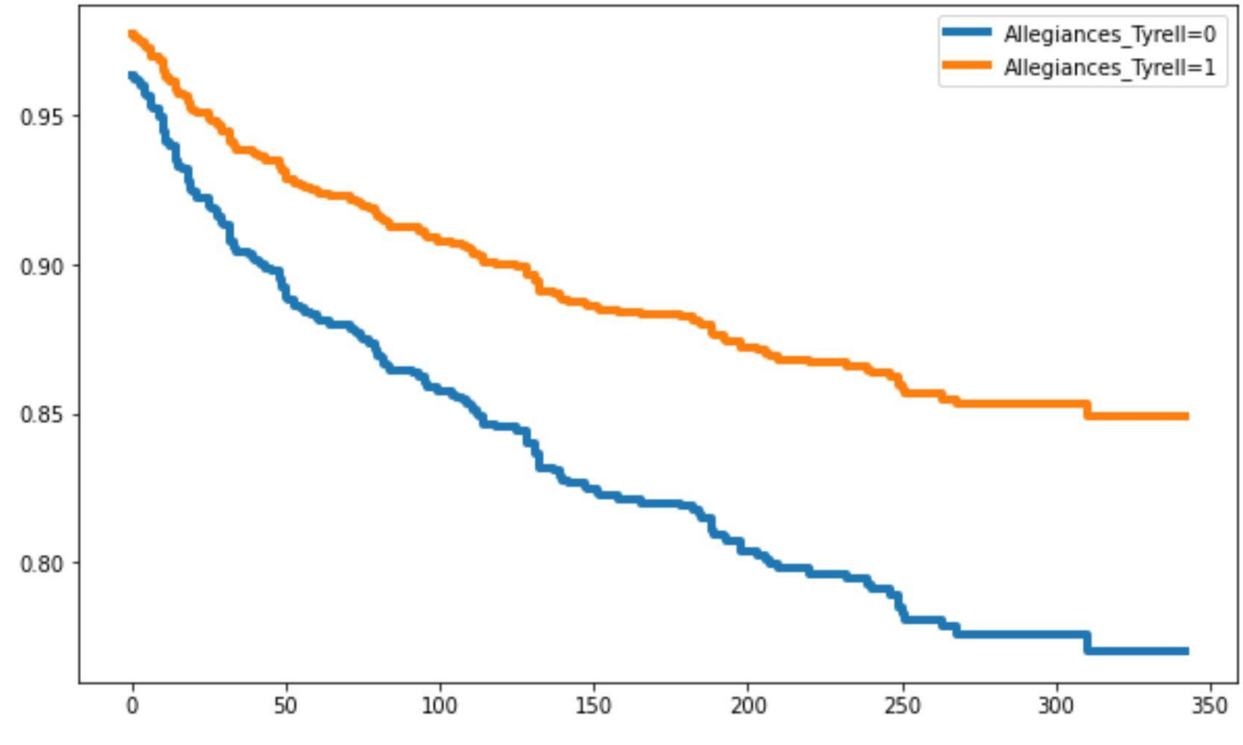


Figure 6: CPH plot of gender and survivability

The plot shows that a female character has a much higher chance out outliving a similar male character. This makes sense both in the story and in historical context as armies were entirely made up men, and very few women characters are involved with fighting in the story.

One of the most interesting findings was the effect a character's allegiance had on their likelihood of survival. The following plots show the difference between the two groups with the starkest contrast in survivability, the wildlings and the Tyrells:





Figures 7 & 8: CPH plots showing survivability for wildlings and those allied with House Tyrell

The plots show that by 350 chapters after being introduced into the story, nearly 80% of wildlings would be expected to die to around only 15% of Tyrells. This makes sense within the story as few characters involved with that house are central to the story and, by the end of the fifth book, have managed to avoid most major conflicts that have occurred. The Wildlings, on the other hand, experience many hardships in the story as they have to battle nature, supernatural beings, and even other wildlings.

The last model was able to show the impact that the book that the character was introduced in has on their survivability:

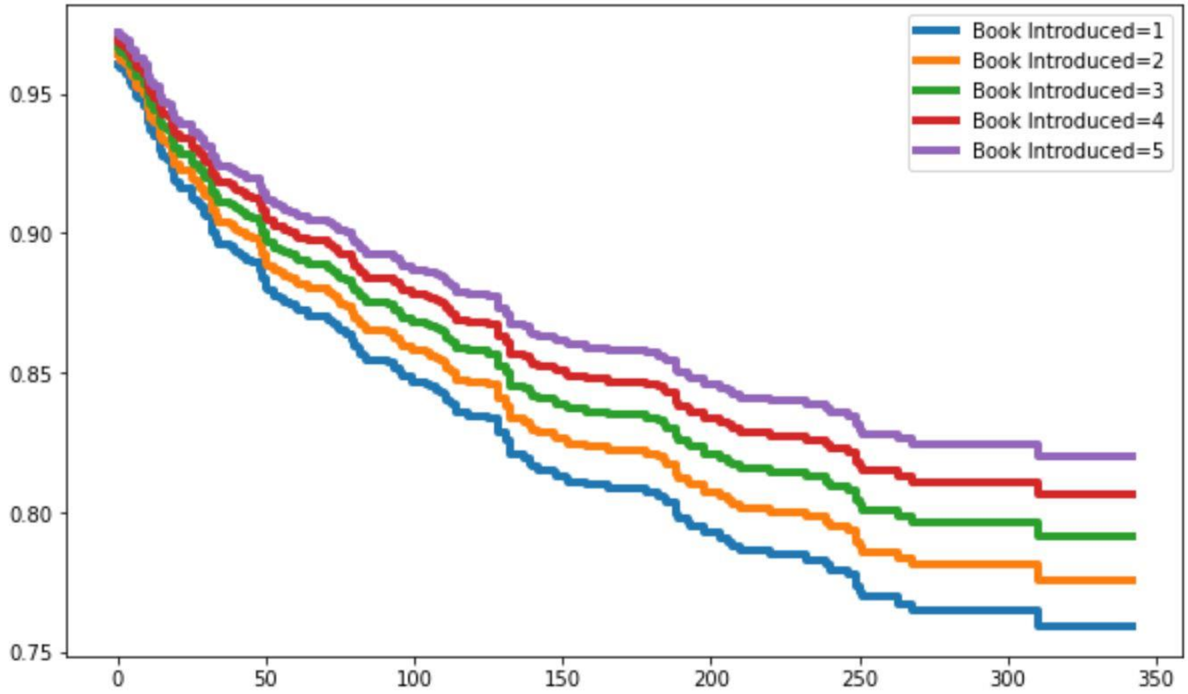


Figure 9: CPH model of how book introduction effects survivability

This data may be misleading, however; due to the censoring of the data. All characters introduced in book five have only been around for less than 100 chapters of the series while characters introduced earlier have experienced more events that could end in their death.

Next Steps:

There are a variety of different steps that can be taken to expand on the steps taken in this report. While the dataset only included a few attributes for each character (gender, allegiance, and nobility status), but any number of different attributes could be added to the dataset as well. Characteristics such as age, role, faith, significance to the story, or location. The issues with adding any new attributes are that information may be difficult to find for many characters and with over 900 characters, it could be time consuming.

Another way that this analysis could be taken further is by trying different survival analysis models. While this report only used the Cox Proportional Hazards Regression Model approach to survival analysis, there are a variety of other model types such as parametric survival models, survival trees, or survival random forests. It is unlikely that these approaches would offer different results in terms of how attributes affected the survival probabilities of characters, however; they could provide different numerical or visual results that should change how these results are interpreted.