

Steam Discount Thresholds: A Deep Dive with Survival Analysis

Student details

Name: Brian van der Steen
E-mail: b.vdrsteen@tilburguniversity.edu
Student number: U102291

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
DEPARTMENT OF COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
TILBURG UNIVERSITY

Thesis committee

Supervisor: dr. P. Hendrix
Second reader: dr. C. Emmery

Tilburg University
School of Humanities & Digital Sciences
Department of Cognitive Science & Artificial Intelligence
Tilburg, The Netherlands
<20 May> <2022>
<8653 words>

Abstract

This thesis aimed to find out which discounting patterns the visual novels and JRPGs on Steam followed and if these could be predicted by survival analysis. This type of analysis has not been used before to predict the time-to-event of reaching a discount. In addition this paper examined how well these models performed on data with various degrees of censoring, taking both the recall and discrimination of the Brier scores and concordance index into account. Each survival model was tested on four discount thresholds, namely 25%, 50%, 75% and 90%. Discount patterns and survival model predictions then lead to more detailed insight into how these games are discounted. As a result making Steam users able to make a more well-informed decision about whether they should wait for a better deal or take the current offer. This study is the first that examines the discount thresholds on Steam and which features affect discounting. Various patterns and discount strategies were found, including that high initial discount and mixed reviews resulted in the fast discounting. Additionally, seasonal Steam sales correlate with dates on which discount thresholds are reached. Finally, for a small dataset of 650 records it seems that the Cox model predicts similarly the Random Survival Forest, which has less assumptions that can be violated. For some discount thresholds the Cox model performs better, while for others the RSF performs better. However, the RSF performs better on unseen data.

Data Source Statement

The author of this thesis acknowledges that they do not have any legal claim to the data from Steam or the data collected from *steamdb.info*, which were the data sources for this thesis.

Page of contents

Abstract.....	2
Data Source Statement	2
1. Problem Statement.....	4
1.1 Research Goal.....	4
1.2 Findings	5
2. Literature Review	6
2.1 Discount Theory	6
2.2 Steam	6
2.3 Alternative Models	7
2.4 Survival Analysis.....	8
2.5 Methodology.....	8
3. Methodology & Experimental Setup.....	9
3.1 Cox Proportional-Hazards	9
3.2 Random Survival Forest	9
3.3 Evaluation Metrics.....	10
3.4 Data Collection.....	10
3.5 The Dataset.....	11
3.6 Descriptive Statistics	11
3.7 Preprocessing.....	12
3.8 The Pipeline.....	14
3.9 Software & Algorithms	15
4. Results	16
4.1 Survival Curves	16
4.2 Feature Selection & Importance	19
4.3 Model Comparison	20
5. Discussion.....	22
6. Conclusion.....	24
References	25
Appendices	28

1. Problem Statement

Most research that includes video games and discount theory is performed from the publisher's point of view, where increasing revenue or sales are common themes. One of such examples are the various articles discussing the best pricing strategies for the digital distribution platform Steam (Carless, 2020, 2021; Dealessandri, 2020). In Steam's case, although research has been done on the effects of discount pricing (Choi & Chen, 2019) and which qualities impact game sales (Choi et al., 2018), little research has been done about how these strategies are broadly applied on the platform. The culture around Steam is somewhat unique in that users like to collect games and bolster their collection, as mentioned by Balmford (2020), often waiting till the price is right. Johnson and Luo (2017) termed this phenomenon 'backlog purchasing', where games are bought cheaply but unlikely to ever be played.

As discounts and collecting games are a central part of Steam, insight into the discount patterns could be of substantial value for the userbase. Indeed, having an improved understanding of which features increase discount speed and when to expect specific discounts allows more informed decision-making. Particularly for cognitive biases, such as hyperbolic discounting and the fear of missing out. For instance, if a game is currently on sale, but is expected to reach a higher discount percentage within a month, the user may reconsider their purchase. Therefore, this study aims to identify discount patterns, such as features that increase discount rates, and attempt to predict when games will reach certain discount thresholds. These thresholds are divided into four levels, ranging from 25% to 90%, and indicate the duration till a discount was reached or breached. Incidentally, reaching a discount threshold can be considered an event. Therefore, survival analysis can be used. Additionally, survival analysis has not been utilized to predict discount thresholds before. Moreover, little research exists that compares the performance across several related models with increased levels of censoring. This thesis aims to fill those gaps in the literature.

In order to do so, data was collected and gathered in a new dataset. Although several Steam datasets exist, they do not include information concerning discount thresholds. Even the most extensive dataset by *steamspy.com*, which was used by the aforementioned related works (Choi & Chen, 2019; Choi et al., 2018), does not include this information. The scope for this study was therefore limited to two genres of games, considering that Steam contains more than 50.000 games (Statista, 2022). Namely, Japanese Role-Playing Games (JRPG) and visual novels. This decision is described in detail in the experimental setup section.

1.1 Research Goal

As the problem statement indicated, there is currently little research about discount thresholds of video games and how discount theory is broadly applied to Steam. Additionally, survival analysis – or time-to-event analysis in this case – has yet to be used to track the days till certain discount thresholds have been reached. In order to collect the information necessary, and fill the aforementioned gaps, the following research questions are in order, starting with the main research question:

Which patterns do visual novels and JRPG games follow regarding discount thresholds on Steam and how well can these thresholds be predicted by survival analysis?

The main research question consists of two parts. First, the exploration of patterns within the dataset. Second, the performance of the chosen survival analysis model. The patterns include general findings, hypotheses and important features. The survival analysis aspect considers the predictive performance of the models tested in light of scientific relevance. To answer the main question there are two sub-questions:

1. *Which features best predict quick discount thresholding in visual novels and JRPG games on Steam?*

This first research question will make use of the complete dataset. It will answer which features increase or decrease the speed in which games are discounted, and which are most significant. Furthermore, descriptive statistics and hypothesis testing will be performed to find patterns and discount strategies.

2. *Which of the selected survival analysis models can best predict the discount thresholds of visual novels and JRPG games on Steam?*

The second question compares the performance of the Kaplan-Meier, Cox Proportional Hazards and Random Survival Forests models on the four discount thresholds. Their performance will be measured with three metrics: the concordance index, Brier score and integrated Brier score. Additionally, the feature selection and feature importance of the models will occur in this stage. The model that performs best will be used on unseen data.

1.2 Findings

The features most likely to lead to quick discounting are a high initial launch discount and mixed reviews of the game. Popular and well-received games tend to be discounted slower as to get the most out of the standard price. Visual novels also do discount more conservatively than JRPGs. For this specific dataset and purpose the Random Survival Forest proved to most accurately predict the discounts, partly because it did not have any assumptions that were violated. The time when visual novels and JRPGs and be expected to reach sales is dependent on multiple factors, with the aforementioned ones and also the discounting strategy that the publisher uses for their games. If they tend to discount more aggressively that could very well occur for most of the games they publish. As for numbers, the 25% discount tends to have less than a 50% survival rate at around the 150 day mark, making the next seasonal sale around that time a good assumption.

2. Literature Review

This is the first study that examines discount thresholds for video games. As such, there are no comparable works. However, there are works based on discount theory that can be compared to this domain. In addition, the discounting documentation of Steam is discussed, including two related works based on discounts on Steam. Then, state of the art models and alternatives to survival analysis are discussed. Lastly, survival analysis studies will be reviewed to determine which models will be used and how to compare their performance.

2.1 Discount Theory

The paper of Grewal et al. (1998) argues the existence of low knowledge and high knowledge customers. The former is influenced by higher discounts and the latter by brand name. For instance, a high knowledge consumer would be a Steam user who had a positive experience with multiple titles of the same developer, resulting in them purchasing their new game near launch. In contrast, a low knowledge consumer would be wary and only purchase the game if it had an adequate discount or was bundled with various other games for an acceptable price.

Acceptable discount ranges differ by discount type. According to research by del Rio Olivares et al. (2018), initial discount values between 5-35% lead to positive customer retention. These values are remarkably similar to the allowed values of 10-40% for initial discounts on Steam (Steamworks, n.d.). Despite the fact that the setting is different, the result appears to generalize well to this domain. Moreover, Zuo & Iida (2018) found that acceptable discount values for online retailers are between 49-64%. Consequently, these results are consistent with the Steam userbase. On *isthereanydeal.com*, a third-party notification site that sends notifications when a Steam game reaches a user-set discount percentage, the average notification percentage for games falls within this range.

Luong and Slegh (2014) investigated the psychology of discount percentages and concluded that consumers were affected by a discount increase of 10% to 50% more so than an increase of 50% to 70%. These findings may indicate why there are significantly less games that reached the 75% and 90% discount thresholds in the dataset. Nevertheless, the results correlate with the most frequent user-set discount on *isthereanydeal.com*, which is 50%. Subsequently, Luong & Slegh also found that discounts for hedonic (pleasure-evoking) goods, such as games, were conceived as more important than other types of products. More importantly, they showed that discounts mattered most when the standard price of a product was perceived as high. This could affect the discount strategies of expensive games on Steam. For example, they might discount faster to 50% than cheaper games.

2.2 Steam

The Steam discounting documentation (Steamworks, n.d.) suggests three best practices. First, the usage of staircase discounting. This strategy consistently increases the discount percentage over time. Additionally, Steam advises to time discounts with content updates. Lastly, to take into account their tail sales delay curve – the amount of users that buy the game at full price – before engaging in discounts. The discounting limitations on Steam related to this study are described below. They indicate the possible discount values and how often discounts are allowed.

- ❖ Steam advises launch discounts to be around 10-15%, with a hard limit of 40%;
- ❖ A discount cannot be less than 10% or be higher than 90%;
- ❖ Discounts can not be run within 28 days of a prior discount, including a launch discount;
- ❖ Discounts can not be run for 28 days following a price increase in any currency;
- ❖ Custom discounts can not run longer than two weeks or shorter than one day.

The papers most similar to this study are those of Choi et al. (2018) and Choi & Chen (2019). These related works focus on discount pricing and use Steam datasets. They, however, did not include discount threshold progression or survival analysis methods. In addition, their data was gathered in 2016, when *steamspy.com* was still able to accurately infer sales data of Steam games (Good, 2018). This resulted in findings supported by accurate sales data, which cannot be achieved anymore. The 2018 study discussed the effects of various intrinsic (opinion) and extrinsic (factual) cues on sales. The findings relevant for this study are that low standard prices lead to increased sales, as long as the game does not get mixed reviews, and that high standard prices can hinder sales when the game is unpopular. Moreover, it appears product popularity and review ratings positively affect sales, whereas company reputation has no significant effect. This is in contrast to the previously mentioned high knowledge customers (Grewal et al., 1998), where brand name affected sales. Perhaps the developer reputation is insignificant when a game has positive reviews and is popular since it has then subverted expectations.

Choi & Chen (2019) confirmed that discounting increases sales. Additionally, they found that having too many competing same-genre products with the same discount percentage on sale decreases sales. For instance, a Steam sale where most visual novels are 50% discounted. Then, they showed that sales increase most at the 20-30%, 40-50% and 70-80% marks. This supports the most frequent-set discounts on *isthereanydeal.com*, being 25%, 50% and 90%. These three thresholds will therefore be three of the four discount thresholds used in this thesis.

2.3 Alternative Models

Survival analysis is a technique prevalent in medical research to predict the time until an event occurs. Unlike classification, which predicts if an event occurred, survival analysis calculates the survival probability in time intervals (George et al., 2014). Furthermore, it handles censored observations well. Particularly, right-censored data are observations that did not experience the event, such as not being discounted 25%. The time of these censored data is taken into account by survival analysis, as otherwise the survival probability is overestimated (Liou, 2021). If time is the variable of interest and the dataset contains censored data, survival analysis should be used (George et al., 2014; Lee & Go, 1997). On the contrary, if binary classification or counts are the response variable of interest, logistic regression or Poisson regression are more suited.

This study explores the time till discount thresholds are reached. Particularly which features affect this duration and how they correlate. Moreover, the data of the discount thresholds is increasingly censored. Therefore, survival analysis is chosen. Common techniques include the Kaplan-Meier, Cox Proportional-Hazards and Random Survival Forest models. Additionally, there are two state-of-the-art deep learning models. Namely, DeepSurv (Katzman et al., 2018) and DeepHit (Lee et al., 2018). The latter explores competing risks, which are not applicable to this study, therefore it is excluded. However, their study and that of Hadanny et al. (2022) found that the RSF model outperforms the DeepSurv model slightly. Contrarily, research by Lee & Lim (2019) and Spooner et al. (2020) reported that a Cox model outperforms a RSF. Therefore, these two best performing models are chosen for the model comparison.

2.4 Survival Analysis

Spooner et al. (2020) compared a large selection of survival analysis models with varying feature selection methods and found no significant difference between the performance of most models. They did conclude that the base Cox model has decreased performance with high-dimensional data. However, with feature selection the Cox model performed slightly better than the Random Survival Forest (RSF) model. This result is consistent with the findings by Lee & Lim (2019). Their research compared a Cox model, Support Vector Machine, RSF and Boosted Cox model and found that the Cox models outperformed the RSF to some degree. The Support Vector Machine performed significantly worse than the RSF and Cox models.

However, these papers only used the concordance index to measure performance. The C-index is a misclassification probability that “estimates the probability that, in a randomly selected pair of cases, the case that fails first had a worst predicted outcome” (Ishwaran et al., 2008, p. 7). As a performance indicator it focuses on discrimination. Alternatively, Mogensen et al. (2012) proposed the additional usage of prediction error curves, known as Brier scores, to get a cumulative rank probability score and compare model performance. Then, both discrimination and calibration could be measured.

Additionally, it must be noted that the findings of these papers used survival analysis in a medical context. It is unknown how these models perform specifically for discount data and how well they can react to severe censoring. Consequently, the data used in this thesis includes features with more than three groups. To determine if multiple survival curves are significantly different for one feature, such as for release year, a log-rank test will be used. The default test of Kaplan-Meier extends to more than two groups (Bland & Altman, 2004), ensuring accurate results.

2.5 Methodology

This thesis bridges several gaps within the current literature. First, it will explore the discount strategies on the Steam platform for visual novels and JRPGs. For this, the 25%, 50% and 75% discounts thresholds discussed in the literature will be tracked. Each survival model will be trained, evaluated and tested on the four separate discount thresholds. This allows the model performance and patterns for each discount threshold to be compared. Additionally, the 90% discount threshold will be added. It is the maximum discount percentage on Steam and should therefore be the most censored model of the four, which is necessary to determine the increasing effect of censoring.

Secondly, the chosen models consist of Kaplan-Meier, Cox Proportional-Hazards and Random Survival Forest. The latter two because they have proved to be the most accurate models in model comparisons. The Kaplan-Meier model will be used to compare survival curves for the whole dataset and test for differences within features. Moreover, it will function as the baseline model performance for the Brier scores. This is because the most important benchmark for their performance is a null model, which ignores all predictor variables, and the Kaplan-Meier is such a model (Mogensen et al., 2012).

3. Methodology & Experimental Setup

The chosen models and performance measures will be further explained in this chapter, starting with the Kaplan-Meier model. This is the most basic survival model and plots a single survival curve for the whole dataset or multiple curves of the groups in a single feature. These curves illustrate the survival function, which shows the probability of the event of interest – in this case, if a game has reached a discount threshold – for each time point, such as days. The Kaplan-Meier model does not show confounding factors between variables, as it is a univariate model. It, however, is a nonparametric model and therefore has no assumptions that can be violated. The model will compare the survival curves of features, such as release year, to determine if the survival curves for all years are the same or if there is a significant difference. The log-rank test will calculate the Chi-square value for each group with the null hypothesis that each survival curve is identical. It will be rejected when a p -value $< .05$ is found, as this indicates that it would be very unlikely at the 5% significance level for the survival curves to be the same. In addition, the Kaplan-Meier model will be used as a baseline expected value for the Brier- and Integrated Brier scores. It cannot be used as a baseline for the concordance index because it is incompatible.

3.1 Cox Proportional-Hazards

The Cox Proportional-Hazards is a semi-parametric survival analysis model and the industry-standard (Graphpad, n.d.). Instead of modelling a survival curve, it estimates the hazard function. For instance, the hazard function for a JRPG that has not reached a 25% discount at 100 days would be the estimated discount rate on that particular date (day 100), given that it ‘survived’ up until that point. The main assumption, proportional hazards, assumes that the hazard ratio is proportional over time. For instance, two survival curves should remain consistent and not abruptly differ in course, otherwise they are non-proportional. If the proportional hazard assumption is met, additional features can be added to the model. This allows Cox to model covariates and their effect values.

The hazard function is estimated for each feature and the groups within, including their confidence interval, in addition to a log-rank p -value for the whole model. The first value of a feature becomes the reference value with a hazard ratio of 1, for example the genre JRPG. The hazard ratio of visual novels, for example 0.8, would then indicate that they discount slower than JRPGs. The included confidence interval then determines the significance of this result. The further from 1 the confidence interval is, the more significant the result.

3.2 Random Survival Forest

When the proportional hazard assumption is violated the performance metrics cannot be used for inference as the Cox model is no longer robust. However, it can still be used for prediction (Graphpad, n.d.; Lifelines, n.d.). In this case, the Random Survival Forest model is the best alternative. There are no assumptions that can be violated and the model is robust to outliers. The cumulative hazard function is averaged amongst all trees, therefore the resulting ensemble hazard cumulative hazard error has a low risk of overfitting (Ishwaran et al., 2008). Further, Nasejje & Mwambi (2017) found that a RSF model that includes non-proportional covariates outperforms a RSF model that only includes proportional covariates. The main parameters of RSF models include the number of nodes and trees created and the maximum grow depth.

3.3 Evaluation Metrics

The concordance index, or C-index, is a performance indicator frequently used to evaluate survival analysis models. It is a measure of discrimination; a ranking metric which allows the correct determination of orders (Scikit-survival, n.d.). Further, it estimates the misclassification probability. The values of the C-index range between 0 and 1. A prediction of 0.5 indicates a prediction identical to random chance. A higher score indicates that the model can discriminate well between games that reach the discount event and those which do not.

Conversely, the time-dependent Brier score also measures calibration. Instead of the ranking of the scores, it uses the actual values of the predicted risk scores to determine if predicts are consistent with the actual observed frequency in the data. Moreover, the Brier score is the extension of the mean squared error for right-censored data (Scikit-survival, n.d.). Furthermore, the Brier score is calculated for a singular time point chosen by the user, while the integrated Brier score calculates the calibration over all time points in the dataset. Both Brier scores result in a value between 0 and 1, where a low score indicates that the model is well-calibrated.

3.4 Data Collection

The dataset for this thesis was created based on data from *steamdb.info* and the Steam store pages of the respective games. The scope includes partially complete populations of two genres instead of random sampling. This leads to more concrete results and relevance. Visual novels and JRPGs genres are chosen as they are of manageable size after some filtering, namely:

- ❖ A price of €4.99 or more, as cheap games are frequently bought without discounts;
- ❖ All release years except 2022, as those are too recent for meaningful data;
- ❖ English support and either Japanese support, Japanese Audio or both;
- ❖ Games that can be viewed without an account, so no adult-only games.

These filters were applied with the instant search engine of *steamdb.info* as it allows advanced filtering, a feature lacking on Steam. The filters meant that only localized games from Japan or popular games that included the language made it into the dataset. In turn this means that the results of this paper could potentially be biased towards discounting strategies of Japanese localized works. However, these two genres have some interesting properties. Both have a wide price range and are either translated or localized. Additionally, they are mainly single-player story-driven games that can either be relatively short or take longer than 50 hours to complete. Visual novels are a niche genre and tend to be discounted conservatively. Contrarily, JRPGs are a more well-known market and tend to be discounted more aggressively. Below is a brief description of these genres:

- ❖ JRPGs are Role-Playing Games in Japanese style. Most often heroic adventures with party members met along the way, such as the Final Fantasy and Neptunia franchises.
- ❖ Visual novels can be seen as (choose-your-own-path) audiobooks with interactive character sprites, background music and background art. Can include gameplay, such as the Danganronpa franchise, or be entirely text-based such as STEINS;GATE.

3.5 The Dataset

The raw dataset consists of 650 rows, of which 368 (57%) visual novels and 282 (43%) JRPGs. The data was collected from the 11th to 21th of April, 2022. Therefore, the cut-off point of the data is the 10th of April, 2022. The dataset consists of 19 columns that are classified under five categories. This classification is shown in Table 1. First is general data that was collected from the Steam store. These consist of the title, developer, publisher, genre, release date, standard price, amount of reviews and the percentage of positive reviews. In addition, followers and discount data was gathered from *steamdb.info*. Most features will be explored for patterns and their effect on discount speed, such as release year, genre, standard price and initial discounts. In contrast, the highest discount percentage dates will be used to determine the occurrence of flash sales and when games last reached an ‘historical low’ discount. The thresholds record when a game first reached or breached that particular threshold. Lastly, each standard price change that occurred for a game was recorded in the price change column. These are color coded to indicate the severity in which they affected the data. For more information about each label in the raw dataset see Appendix A. For discount strategies found during data collection see Appendix B and for examples of the color codes see Appendix C.

Table 1: Labels in the raw dataset.

Type of data	Labels
General data	Game title, developer, publisher, genre, release date and standard price in euro.
Continuous data	Amount of reviews, the percentage of positive reviews, followers.
Discount statistics	Launch discount, highest discount percentage, first date of highest discount percentage, last date of highest discount percentage.
Discount thresholds	25% threshold, 50% threshold, 75% threshold and 90% threshold.
Other	Price changes.

3.6 Descriptive Statistics

Time tends to not be normally distributed. Likewise, this occurred in this dataset as well. This section provides a brief insight into the distribution of the data. For instance, various features such as reviews number, review rating and follower amount are highly skewed. As they have values greater than 1 or lower than -1. Furthermore, these have kurtosis values considerably higher than 3. The review ratings of the JRPG and visual novel genres tend to be biased towards positive ratings, therefore they are skewed. Moreover, popularity of games varied significantly in the dataset, resulting in the skewness of review and follower amount. The descriptive statistics in Table 2 indicate two trends. First, both genres reached 25% and 50% discounts at a somewhat equal rate. However, JRPGs reached 75% discounts more often than visual novels, despite being smaller in sample initially. Second, when viewing the mean time till a discount is reached, it may appear that JRPGs reach the 25% and 50% discount faster, while they reach the 75% and 90% discount at nearly the exact rate of visual novels. The high standard deviation indicate that this is, however, likely coincidental.

From the 650 games in the dataset, 600 (92%) reached a 25% discount, 447 (69%) reached 50%, 155 (24%) reached 75%, and 26 (4%) reached the maximum discount on Steam, 90%. Moreover, a Pearson’s correlation test was conducted and found a correlation of 0.833 between reviews and followers, indicting a strong positive relationship. In addition, there appears to be a correlation of 0.380 between followers and standard price, indicating a weak positive relationship.

Table 2:
Descriptive statistics divided by visual novels and JRPGs.

Labels	Features				
	N	Mean	SD	Skew	Kurtosis
price_eur	368 / 282	19.10 / 27.70	10.6 / 16.3	1.16 / 0.70	1.3 / -0.8
reviews	368 / 282	955 / 2218	2805 / 6685	5.45 / 7.04	33.8 / 60.9
percentage	368 / 280	0.89 / 0.80	0.13 / 0.15	-2.72 / -1.43	9.9 / 3.2
followers	368 / 282	6175 / 19920	11766 / 39180	4.31 / 6.20	21.6 / 56.9
ini_discount	368 / 282	0.13 / 0.12	0.10 / 0.12	1.66 / 1.33	5.6 / 2.0
highest_threshold	368 / 282	0.54 / 0.56	0.19 / 0.22	-0.09 / -0.23	-0.6 / -1.0
tf0_25	342 / 258	179 / 114	160 / 103	1.89 / 1.90	6.0 / 5.1
tf0_50	256 / 191	434 / 314	327 / 254	1.29 / 2.72	2.0 / 13.6
tf0_75	73 / 82	727 / 690	514 / 412	0.56 / 0.55	-0.4 / -0.5
tf0_90	12 / 14	909 / 922	828 / 642	0.66 / 0.37	-1.3 / 1.4
tf25_50	256 / 191	286 / 218	274 / 236	1.75 / 3.46	4.5 / 20.9
tf50_75	73 / 82	477 / 481	486 / 363	0.86 / 1.02	0.02 / 0.7
tf75_90	12 / 14	328 / 338	702 / 464	2.10 / 1.47	3.4 / 0.9

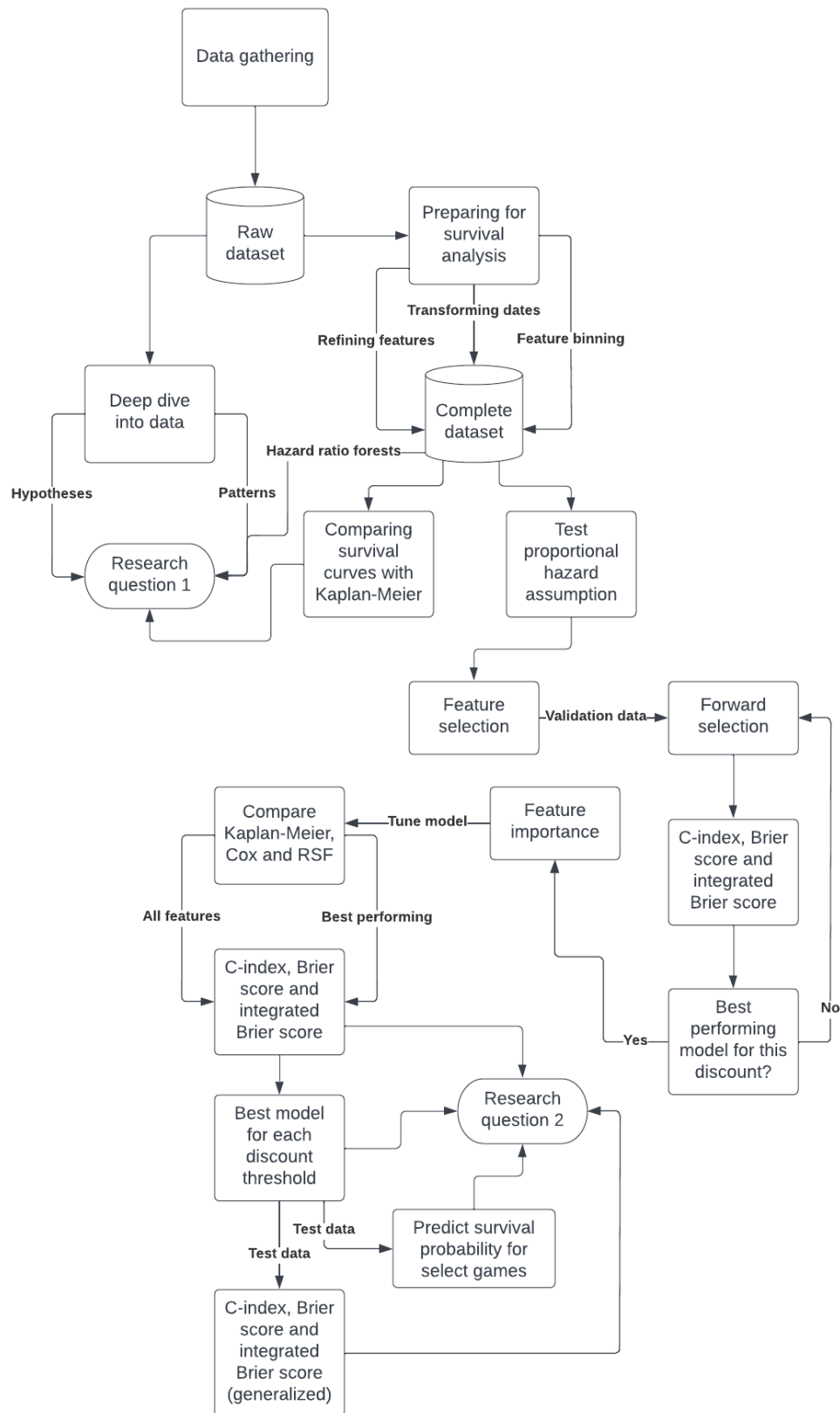
Note. The first number in each column is that of visual novels, the second of JRPGs. In order to display the numbers on a single row they are rounded to the nearest number.

3.7 Preprocessing

This section describes the process from the raw dataset to the survival analysis suitable dataset. This required a time till event, event indication and binned features. The duration till a discount was tracked in days and was achieved by subtracting the threshold date by the release date. The NA values – which resulted from games that did not reach the discount threshold – were used to classify if a game reached an event (1) or was censored (0). This is the event indicator for the survival models. The NA values were then transformed by subtracting the cut-off point of data selection – 10 April 2022 – by the release date.

Then, relevant numerical features were binned to ensure that they could be compared by the survival analysis models. First, standard price was roughly divided into its quantiles. Reviews ratings were binned based on Steam’s review categories, such as ‘very positive’. Groups that had at least 30 observations were preferred. For instance, when a feature such as release year had two observations in 2012, it would be combined with the following levels till this condition was met, resulting in a ‘2012-2015’ level. Release year was binned twice, once by year and once by quantiles to compare their performance. Followers were binned by quantiles and so was highest discount percentage. Initial discount had enough observations for each discount that occurred and was left as is. A new feature, releases, was created by calculating the releases of all publishers in the dataset. Each publisher was binned into one of five levels, such as ‘2-4’, resulting in each game from that publisher to have that release level. Lastly, Brier scores do not function with events that are reached instantly, therefore all values of 0 were changed to 1.

Figure 1: Flowchart of the pipeline of this thesis.



3.8 The Pipeline

The first research question that must be answered includes finding patterns and features that affect the time till discount. In addition, patterns for three topics will be explored. First, whether there is a pattern in standard price changes. Second, whether discount thresholds primarily occur during Steam sales, and if so which ones. Third, whether the fear of missing out on historical low prices is grounded in reality. These findings could further benefit Steam users in their decision-making, increasing the relevance of this study. To determine which features affect discounting for the games in the dataset, first the survival curves of the isolated features will be compared with the Kaplan-Meier models. The resulting p -values of these comparisons are placed in a table for each tested feature and each discount threshold. Moreover, the survival curves for each isolated feature and discount threshold will be shown. They provide valuable information concerning which levels of features have an increased discount rate and therefore are included. In addition, Cox hazard ratio forests are included as they provide this information based on multivariate analysis, indicating the significance of various features or levels within. Additionally, the proportional hazards assumption of the Cox model is tested with Schoenfeld residuals. Aalen additive hazards are shown to reveal the differences in hazard ratio over time.

The second research question compares the performance of the selected survival analyses and identifies which model can best predict discount thresholds. For each discount threshold the Cox and RSF models will be calibrated with forward selection; a feature selection technique where the best performing feature is added, then the next best performing feature and so on. For example, the performance of a model is first compared with one feature. When there are six features, the one that results in the highest concordance index is chosen. Then, the remaining five features will be compared in a two feature model that includes the first best performing feature. This process continues until a model decreases in performance by adding more features. For this thesis, both the concordance index and integrated Brier score will be used to determine this performance. Ultimately this results in the best performing Cox and RSF models for each discount threshold. Then, the feature importance of the RSF models is calculated and located into a table. This indicates the effects the individual features have on the model. The hyperparameters of the RSF are partially automatically tuned. The number of variables to possibly split after each node and minimum size of the terminal node are selected that result in the lowest error rate for the model. Additionally, RSF sizes of 100 trees, 500 trees and 1000 trees were considered. The error rate remained lowest with 500 trees, therefore it was chosen as the default.

The dataset is split into 70% train, 15% validation and 15% test sets and equally distributes the event variable, so that each subset contains an equal amount of censored and uncensored data. Feature selection and feature importance are tuned with the validation data. Additionally, the Cox and RSF models are compared on the validation data. The first comparison will use the full models of the Cox and RSF to determine how they perform with the same features. Further, this allows accurate interpretation of the baseline Kaplan-Meier. Then, the best performing models of Cox and RSF – which use their own best performing features – are compared to determine whether Cox or RSF is the better technique. As these models may be overfitted to the validation data, they will also be used on the test data to determine how well they can generalize. Lastly, these models will be used to predict the survival probability of a collection of five unseen visual novels or JRPGs that were not included in the dataset. These numbers are then compared to the actual values to determine how well these results can be interpreted.

3.9 Software & Algorithms

To fit the Kaplan-Meier model and Cox models I use the Survival package in R and use the functions `surv_fit()` and `coxph()` in addition to `aareg()`. To test for proportional hazards I use the `cox.zph()` function from the `Survminer` package. Additionally, I fit the RSF model with the `rfsrc()` function and tune it with the `tune()` function in the `randomForestSRC` package. Feature importance will be calculated with the `vimp.rfsrc()` function from the same package. The dataset is split with the `partition()` function from the `splitTools()` package. For the evaluation metrics I use the `Cindex()`, `Brier()`, `IBS()` functions from the `SurvMetrics` package. Lastly, to predict survival probabilities I use the `predictSurvProb()` function from the `pec` package.

4. Results

Several patterns were found regarding the topics that could be beneficial to Steam users. First, standard price changes proved not entirely coincidental. All games published by ‘Idea Factory International’ ($n = 24$) have had price reductions. In contrast, of their games that were published by another company, only one changed price ($n = 10$). Additionally, ‘ROSEVERTE’ ($n = 10$) had a price reduction ratio of 60%, ‘Fruitbat Factory’ ($n = 8$) 50%, ‘NIS America, Inc.’ ($n = 19$) 42%, ‘JAST USA’ ($n = 10$) 40% and ‘XSEED Games’ ($n = 14$) 36%.

Second, a significant amount of discount thresholds occurred on the same date that a seasonal sale started. This accounts for at least half of all 25% discount thresholds and progressively increases as the threshold got higher, resulting in a percentage of 77% for the 90% threshold. In addition, thresholds appeared to happen most during the three main seasonal sales: summer, winter and autumn; being responsible for 71% of all occurrences. A frequency table of these numbers is showcased in Table 3.

The last pattern regards fear of missing out. Out of 596 cases in the dataset – as orange-labeled entries had to be removed for this – 384 (64%) had their highest threshold discount in 2022 and 109 (18%) in 2021. Flash sales, defined as reaching the highest discount on the same first and last date in addition to having this occur before 2021, occurred 62 (10%) times. The remaining 41 entries had their highest discount last before 2021, but were not considered flash sales.

Table 3:

Frequency table of discount threshold dates that overlapped with seasonal sale start dates.

Seasonal sales	Discount threshold				
	25%	50%	75%	90%	Total
Summer [14]	109	84	33	8	234 (33%)
Winter [14]	63	57	15	6	141 (19%)
Autumn [7]	51	56	30	2	139 (19%)
Halloween [4]	40	24	6	2	72 (10%)
Lunar New Year [7]	34	29	5	2	70 (10%)
Golden Week [7]	29	22	6	0	57 (8%)
JRPG [7]	1	2	3	0	6 (1%)
Total	327 (55%)	274 (61%)	98 (63%)	20 (77%)	719

Note. The horizontal percentages are based off the total number of games that reached that threshold, the vertical percentages are based off the total occurrences in the table. The brackets after the sale name indicate their most recent duration in days.

4.1 Survival Curves

The Kaplan-Meier survival curves indicate that around 160 days there is a 50% survival probability that a game has been discounted to 25%. This takes around 420 days for a discount of 50% and for both 75% and 90% has not been reached. The lowest survival probability for a 75% discount lies around 60% and that of 90% at the same percentage, 90%.

The comparison of the isolated features indicated that most groups, regardless of the discount threshold, had at least one factor that was significantly different compared to the others. Of the 40 combinations tested, only 6 (15%) resulted in a non-significant p -value. Noteworthy results include genre, which does not seem to affect survival significantly for 90% discounts, as do review ratings. Instead, publishers and launch discount are the best indicators. However, initial discounts can be 25% or higher, and highest discount contains data leakage, therefore these five results are biased. For the full list of p -values, see Table 3.

Table 4:
Survival curve log-rank test p -values

Label	Discount threshold			
	25%	50%	75%	90%
Genre	<.0001	<.0001	<.0001	.14
Price	.031	.18	.11	.019
Review rating	<.0001	<.0001	.00036	.083
Releases	.00035	.00042	.11	.025
Followers	.002	.0012	.13	.21
Release year	<.0001	<.0001	<.0001	.019
Release year (Q)	<.0001	<.0001	<.0001	.01
Launch discounts	<.0001	<.0001	<.0001	.00059
Highest discount	<.0001	<.0001	<.0001	<.0001
Publisher	<.0001	<.0001	<.0001	.0026

Note. Log-rank $p < .05$, Bonferroni correction $p < .00125$. Italic: significant in log-rank test, bold: significant after Bonferroni correction.

Two examples of significant survival curves are shown in Figure 2 and Figure 3. The publisher survival curves indicate that most publishers discount at around the same time, as can be seen by the steep declines of the curves, especially for the 50% model. The steep decline at in the 25% model implies that most games by those publishers launched with a discount of 25% or more. Furthermore, the release year survival curves illustrate well how games within the dataset have varying amounts of time to remain censored. It also seems that games from 2016 or earlier discounted faster than recent releases. In addition, games from 2019 seem to be more likely to reach a 25% near their tail-end compared to games from 2017 and 2018 by a significant margin, even though they had equal survival probability around the 450 day mark. Appendix D contains an overview of all survival curves. Additionally, the hazard ratio forests in Appendix E include the feature effects as covariates, which more accurately indicates which levels of features are significant and lead to reduced and increased discounting speed. However, the proportional hazards assumption of the Cox model was violated, therefore these results might not be entirely robust. Appendix F illustrates the non-proportionality in hazard ratio of the features with the Aalen additive model.

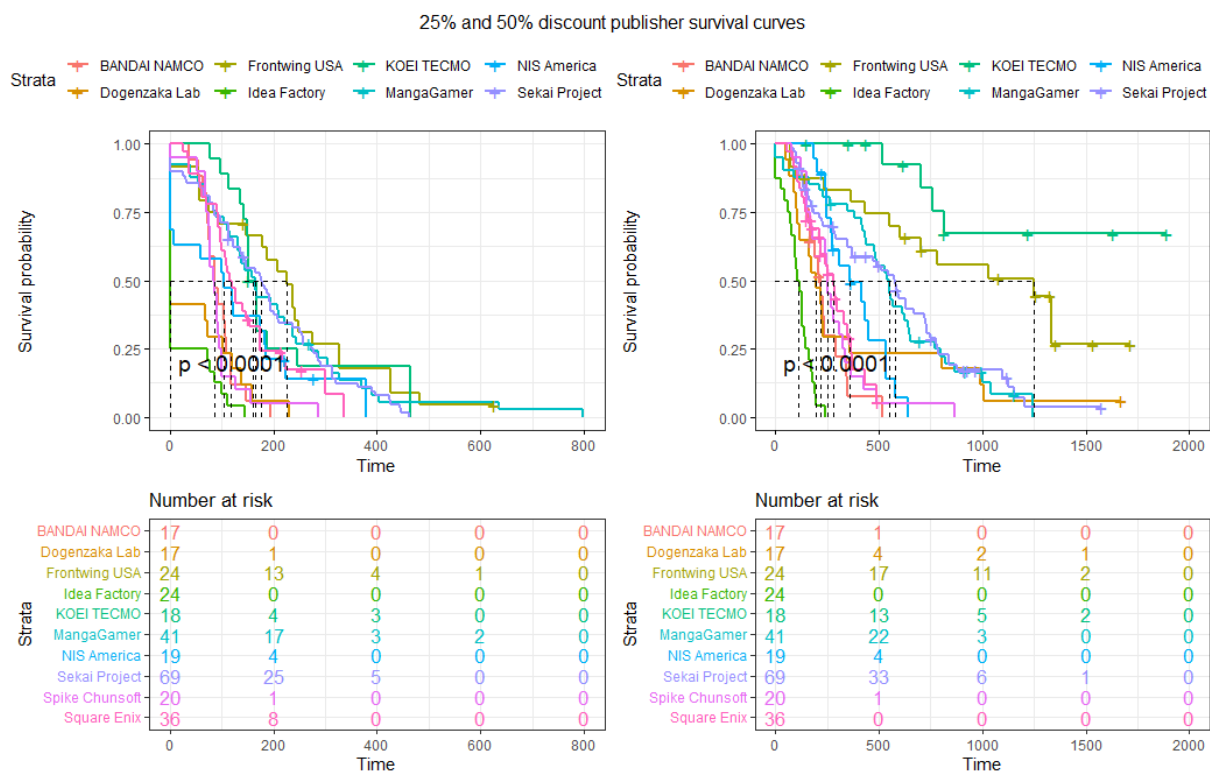


Figure 2: Two survival curve models of the top five publishers of each genre.

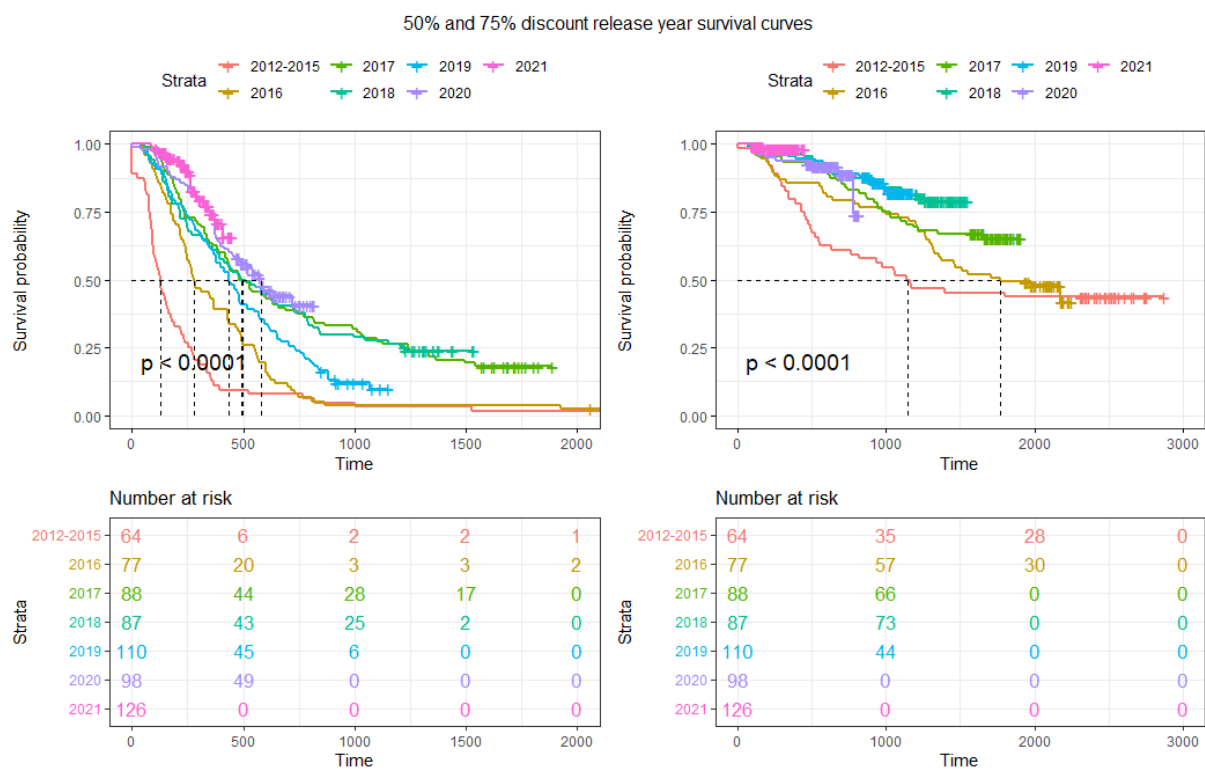


Figure 3: Two survival curves models of the release year feature.

4.2 Feature Selection & Importance

To prepare for model comparison the features that maximized model performance are selected by forward selection. Table 5 illustrates an example of the 50% Cox model. Release year proved to be the best performing feature, but overall performed best with 5 features included. Furthermore, the 50% discount threshold was the only model where the Cox and RSF selected the same features, albeit in a different order. For the 25% discount model both selected three features, including release year. Contrarily, for the 75% discount model the RSF performed best on all six features, while the Cox model did so on three. Primarily, review ratings, initial discount and followers. The latter two are then not included in the 90% discount model, which contains the remaining four. The RSF model selected three features for this last model. This indicates that both models differ in their selection procedure. Moreover, it appears that each discount threshold has different features that best predict it. Between models and within them. The full feature selection for both models is located in Appendix G.

Table 5
Feature selection subset of Cox for the 50% threshold on evaluation data.

Label & model	Performance metric			
	C-index R	C-index	Brier	IBS
Release year	0.649	0.648	0.208	0.358
Review ratings	0.687	0.707	0.194	0.344
Initial discount	0.735	0.721	0.231	0.272
Price bin	0.742	0.725	0.226	0.239
Genre	0.743	0.726	0.223	0.236

Note. The C-index R was the reference model and based on the whole dataset. For the C-index a higher score is better. For the Brier and IBS score lower is better.

Furthermore, feature importance concluded that the chosen parameters of the RSF models all had positive predictive power and should remain in the model. In addition, the performance error progressively decreased when the discount threshold went up. The certainty of the model also increased at the 90% threshold. Perhaps because there was little censored data to misclassify. Release year and initial discount seemed to have the largest effect across discounts.

Table 6
Feature importance

Features	Discount thresholds			
	25%	50%	75%	90%
Genre	0.145	0.057	0.068	
Price bin	0.116	0.015	0.012	0.353
Review ratings		0.073	0.138	
Release year	0.218	0.176	0.099	0.109
Initial discount		0.173	0.140	0.506
Followers			0.032	

Requested performance error	0.310	0.259	0.194	0.010
--------------------------------	-------	-------	-------	-------

4.3 Model Comparison

Table 7 illustrates the performance of the Cox and RSF models across discount thresholds. In the validation data the RSF model outperforms the Cox model in the 25% and 75% discounts in concordance. However, the Cox model outperforms the RSF model greatly for the 50% discount threshold. In addition, it consistently has a better integrated Brier score than the RSF or one close to it. This trend continues for the test data, except for the 75% discount threshold, where it is outperformed by a large margin. Furthermore, the Cox model had a decreased performance on the test data for all thresholds except 50%. In contrast, the RSF model increased in performance for the 50% and 75% thresholds. Both models decreased in performance for the 25% and 90% thresholds, indicating that they may have been overfit for the validation data. In addition, the Brier scores are reasonable for the 25% discount, decrease for the 50% one, then become high at 75% to ultimately become very small. One explanation is that the increase in censoring in the discount thresholds decreases the calibration performance until there are very few censored rows, where ranking the censored data becomes simpler.

In contrast, when the models which contained six features were compared the Cox model performed better than the RSF model in the 25% and 50% discount thresholds, whereas the RSF model had increased performance for the 75% and 90% model. The lowest integrated Brier score of the 75% discount threshold, however, was achieved by the Kaplan-Meier model. Furthermore, when the predicted survival probabilities were compared to the actual values, the RSF predicted these closer to the actual values for releases dates it had learned from than the Cox model. Contrarily, it estimated more conservative estimates than the Cox model for games released in 2022. An overview of the comparison of the fully featured models is shown in Appendix H. The survival probability predictions table can be found in Appendix I.

Table 7
Model performance comparison with best performing features on validation data and test data.

Metric	Models			
	Cox val	RSF val	Cox test	RSF test
25%				
C-index	0.687	0.698	0.616	0.614
Brier	0.190	0.174	0.221	0.225
IBS	0.125	0.124	0.104	0.113
50%				
C-index	0.726	0.696	0.728	0.732
Brier	0.223	0.282	0.176	0.170
IBS	0.236	0.273	0.192	0.205
75%				
C-index	0.719	0.736	0.678	0.758
Brier	1.290	2.114	0.094	0.097
IBS	0.372	0.448	0.728	0.561
90%				
C-index	0.972	0.957	0.839	0.729
Brier	0.019	0.017	0.010	0.007
IBS	0.021	0.024	0.014	0.016

5. Discussion

In this section, the main research question *Which patterns do visual novels and JRPG games follow regarding discount thresholds on Steam and how well can these thresholds be predicted by survival analysis?* and the subsequent sub-questions *Which features best predict quick discount thresholding in visual novels and JRPG games on Steam?* and *Which of the selected survival analysis models can best predict the discount thresholds of visual novels and JRPG games on Steam?* will be elaborated.

The first sub-question aimed to find discount patterns that could prove beneficial to Steam users and determine which features had a significant increasing or decreasing effect on discount duration. First is the permanent price reduction of games of several publishers, including ‘Idea Factory International’. These findings indicate that publishers may use permanent price reduction as a strategy to increase sales. Furthermore, some publishers do this for more of their games than others. When this pattern of a publisher is known, users may consider to wait longer to purchase games from them for the chance of a price reduction. Contrarily, the games of developer and publisher ‘Idea Factory International’ that were published by other publishers did not decrease in price, whereas those published by themselves had all reduced in price. This might indicate that permanent price reductions are determined by the publisher and not the developer. In addition, Steam users can expect discount thresholds to be reached or breached most often during seasonal sales. Games that did not reach a discount yet might therefore reach the desired one during the upcoming sale. In particular in the Summer, Winter or Autumn sale. The higher the discount percentage, the more likely that it is only reached during one of the seasonal sales.

The features that best indicate increased discounting speed across all discount thresholds are launch discount percentage and review rating. Mixed reviews and a high initial discount are the most significant predictors of earlier discount. In contrast, recent games with a positive review that are the first game of that developer and had a low initial discount, discount the slowest. Furthermore, visual novels tend to discount more conservatively than JRPGs, which reach discount thresholds faster, except for the 90% discount threshold. In addition, the higher the user rating for a game, the more slowly it is discounted.

In continuation of the research of Spooner et al. (2020) and Lee & Lim (2019), the Cox model and Random Survival Forest were compared. Unique in this study were the observations of the same event at multiple levels and various amounts of censoring. In contrast to these studies not only the concordance index, but also the Brier scores were measured. These do tend to perform, or either well together, or poor together. The Cox and Random Survival Forests performed similarly as was concluded in prior research. Their performance varied between discount thresholds however, where the Cox model outperformed the RSF or likewise. This indicates that discount thresholds significantly differ, and that there is no one-model solution when predicting discounts thresholds. The RSF predictive power did however improve on generalized data, and therefore would be the better one of the two in this case.

The two main limitations of this thesis were the scope and the robustness of the Cox model results. The results may be biased towards games that were translated or localized in Japanese and therefore might not be reproducible for other genres. Additionally, the proportional hazard assumption of the Cox model was violated, leading to comparisons that might not be robust. However, the generalized predictions conclude that the RSF model is the better predictive

model for unseen data, when prediction should not be affected by the violation. Therefore the violation might not have affected the results significantly.

To answer the main research question, there are several discount strategies used by publishers on Steam. These include permanent price reduction, the frequent breach of discount thresholds during seasonal sales and the occurrence of historical low prices that might incur fear of missing out. In addition, it was confirmed that more well-received and popular games discount slower than those which receive mixed reviews are unpopular. Furthermore, survival analysis as a technique best indicates which features lead to slower discounting or faster discounting. Therefore, for this task they perform well. They also perform well in comparing the discount speed of several features and variables. However, they do not perform as well for prediction purposes, such as an accurate time till an event for this dataset. This should be further tested for a larger dataset. Nonetheless, the Random Forest Survival forest outperformed the Cox model for generalization, and should therefore be considered for survival analysis in this domain.

6. Conclusion

In this section a brief overview is given of the thesis. First, it was proven that games that are well received discount slower than games which are not. In addition, high initial discounts and mixed reviews resulted in games that discounted the fastest. Contrarily, the pricing of games, the amount of followers and the genre affected discount speed less. Furthermore, consumers can expect discount thresholds to be reached during seasonal sales, with higher discounts having the more likely occurrence. Second, the Cox and Random Survival Forest models were compared. These models chose differing features to predict discount thresholds for the discount thresholds models and outperformed each other on the varying thresholds. The Random Survival Forest model, however, generalized the best of both models. Further research should be done a larger dataset that uses random sampling or a larger variety of genres as to not bias the results towards localized titles.

References

- Balmford, W. (2020). Collecting, displaying and not playing: Steam sales and digital game collections. *Proceedings of DiGRA Australia 2020*. https://digraa.org/wp-content/uploads/2020/02/DiGRAA_2020_paper_15.pdf
- Bland, J. M., & Altman, D. G. (2004). The logrank test. *BMJ*, 328(7447), 1073. <https://doi.org/10.1136/bmj.328.7447.1073>
- Carless, S. (2019, August 14). *Steam's "Follower" counts - hidden in plain sight*. Game Developer. Retrieved May 12, 2022, from <https://www.gamedeveloper.com/business/steam-s-follower-counts---hidden-in-plain-sight->
- Carless, S. (2020, October 2). *Switch & Steam discount strategies? We got 'em*. Gamediscover. Retrieved May 12, 2022, from https://newsletter.gamediscover.co/p/switch-and-steam-discount-strategies?utm_source=url&s=r
- Carless, S. (2021, January 7). *Discount strategy: your key post-launch game discovery tool 2*. Game Developer. Retrieved May 12, 2022, from <https://www.gamedeveloper.com/business/discount-strategy-your-key-post-launch-game-discovery-tool-2>
- Choi, H. S., & Chen, C. (2019). The effects of discount pricing and bundling on the sales of game as a service: An empirical investigation. *Journal of Electronic Commerce Research*, 20(1). https://libres.uncg.edu/ir/asu/f/Chen_Charlie_2019_Effects%20of%20Discount%20Pricing%20Bundling.pdf
- Choi, H. S., Ko, M. S., Medlin, D., & Chen, C. (2018). The effect of intrinsic and extrinsic quality cues of digital video games on sales: An empirical investigation. *Decision Support Systems*, 106, 86–96. <https://doi.org/10.1016/j.dss.2017.12.005>
- Dealessandri, M. (2020, June 15). *How to price your game and thrive during Steam sales*. GamesIndustry.Biz. Retrieved May 12, 2022, from <https://www.gamesindustry.biz/articles/2020-06-10-how-to-price-your-game-and-thrive-during-steam-sales>
- del Rio Olivares, M. J., Wittkowski, K., Aspara, J., Falk, T., & Mattila, P. (2018). Relational Price Discounts: Consumers' Metacognitions and Nonlinear Effects of Initial Discounts on Customer Retention. *Journal of Marketing*, 82(1), 115–131. <https://doi.org/10.1509/jm.16.0267>
- George, B., Seals, S., & Aban, I. (2014). Survival analysis and regression models. *Journal of Nuclear Cardiology*, 21(4), 686–694. <https://doi.org/10.1007/s12350-014-9908-2>
- Good, O. S. (2018, April 27). *Steam Spy isn't dead, but it isn't as accurate as it was, creator says*. Polygon. Retrieved May 12, 2022, from <https://www.polygon.com/2018/4/28/17292228/steam-spy-privacy-law-shut-down#:~:text=On%20April%2011%2C%20Valve%20announced,that%20owner%20actively%20opted%20in.>

- Graphpad. (n.d.). *The Ultimate Guide to Survival Analysis*. Retrieved May 12, 2022, from <https://www.graphpad.com/guides/survival-analysis>
- Grewal, D., Krishnan, R., Baker, J., & Borin, N. (1998). The effect of store name, brand name and price discounts on consumers' evaluations and purchase intentions. *Journal of Retailing*, 74(3), 331–352. [https://doi.org/10.1016/s0022-4359\(99\)80099-2](https://doi.org/10.1016/s0022-4359(99)80099-2)
- Hadanny, A., Shouval, R., Wu, J., Gale, C. P., Unger, R., Zahger, D., Gottlieb, S., Matetzky, S., Goldenberg, I., Beigel, R., & Iakobishvili, Z. (2022). Machine learning-based prediction of 1-year mortality for acute coronary syndrome ☆. *Journal of Cardiology*, 79(3), 342–351. <https://doi.org/10.1016/j.jjcc.2021.11.006>
- Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. *The Annals of Applied Statistics*, 2(3). <https://doi.org/10.1214/08-aos169>
- Johnson, M. R., & Luo, Y. (2017). Gaming-value and culture-value: Understanding how players account for video game purchases. *Convergence: The International Journal of Research into New Media Technologies*, 25(5–6), 868–883. <https://doi.org/10.1177/1354856517743667>
- Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., & Kluger, Y. (2018). DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1). <https://doi.org/10.1186/s12874-018-0482-1>
- Lee, C., Zane, W., Yoon, J., & van der Schaar, M. (2018). DeepHit: A Deep Learning Approach to Survival Analysis With Competing Risks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2314–2321. <https://doi.org/10.1609/aaai.v32i1.11842>
- Lee, E. T., & Go, O. T. (1997). Survival analysis in public health research. *Annual Review of Public Health*, 18(1), 105–134. <https://doi.org/10.1146/annurev.publhealth.18.1.105>
- Lee, S., & Lim, H. (2019). Review of statistical methods for survival analysis using genomic data. *Genomics Inform*, 17(4), e41. <https://doi.org/10.5808/gi.2019.17.4.e41>
- Lifelines. (n.d.). *Survival regression — lifelines 0.27.0 documentation*. Readthedocs.Io. Retrieved May 12, 2022, from <https://lifelines.readthedocs.io/en/latest/Survival%20Regression.html>
- Liou, L. (2021, February 4). *How well do you “really” know survival analysis?* Medium. Retrieved June 11, 2022, from <https://towardsdatascience.com/how-well-do-you-really-know-survival-analysis-221885b89a8e>
- Luong, A., & Slegh, D. (2014). Hedonic product discounts: When is the price right? *Nankai Business Review International*, 5(4), 356–364. <https://doi.org/10.1108/nbri-03-2014-0018>
- Mogensen, U. B., Ishwaran, H., & Gerds, T. A. (2012). Evaluating Random Forests for Survival Analysis Using Prediction Error Curves. *Journal of Statistical Software*, 50(11). <https://doi.org/10.18637/jss.v050.i11>

- Nasejje, J. B., & Mwambi, H. (2017). Application of random survival forests in understanding the determinants of under-five child mortality in Uganda in the presence of covariates that satisfy the proportional and non-proportional hazards assumption. *BMC Research Notes*, 10(1). <https://doi.org/10.1186/s13104-017-2775-6>
- Scikit-survival. (n.d.). *Evaluating Survival Models* — *scikit-survival 0.17.2*. Readthedocs.Io. Retrieved May 12, 2022, from https://scikit-survival.readthedocs.io/en/stable/user_guide/evaluating-survival-models.html
- Spooner, A., Chen, E., Sowmya, A., Sachdev, P., Kochan, N. A., Trollor, J., & Brodaty, H. (2020). A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. *Scientific Reports*, 10(1). <https://doi.org/10.1038/s41598-020-77220-w>
- Statista. (2022, February 25). *Number of games released on Steam 2004–2021*. Retrieved May 12, 2022, from <https://www.statista.com/statistics/552623/number-games-released-steam/>
- Steamworks. (n.d.). *Discounting*. Retrieved May 12, 2022, from <https://partner.steamgames.com/doc/marketing/discounts>
- Zuo, L., & Iida, H. (2017). An Analysis of Sales Promotion ‘Discount’ Using Game Refinement Measurement. *Entertainment Computing – ICEC 2017*, 487–491. https://doi.org/10.1007/978-3-319-66715-7_69

Appendix A: The Dataset

<i>Column / Label</i>	<i>Description</i>
<i>title</i>	The name of the game on the Steam store.
<i>developer</i>	The name of the developer credited on the Steam store page. In case of multiple developers, the first one mentioned or one better suited was chosen.
<i>publisher</i>	The name of the publisher, which for translated works often is the localization company.
<i>genre</i>	Either JRPGs (Japanese Role-Playing Games) or VNs (Visual Novels).
<i>release_date</i>	The release date as shown on the Steam store page. Sometimes these use the original release date of the game instead of the Steam release, which is then corrected.
<i>price_eur</i>	The standard price of the game in euros. If a price change occurred at any point it indicates the chosen standard price for the data.
<i>reviews</i>	The amount of reviews the game has received. This number includes reviews of any language, but by default excludes reviews by users who activated a Steam key instead of purchasing the game directly through Steam. For example, when buying via a third-party seller – like Humble Bundle – or receiving product keys for beta testing or reviewing purposes (see Packages on a steamdb.info game page).
<i>percentage</i>	The percentage of total reviews that were a positive review.
<i>followers</i>	The amount of users who followed the game on Steam. When a user follows a game, they will see announcements of the game in their community activity feed.
<i>ini_discount</i>	The ‘launch discount’ the game had on Steam. If a game had no launch discount this will be 0%.
<i>highest_discount</i>	The highest discount percentage the game has had.
<i>highest_discount_date_f</i>	The date on which the highest discount percentage first occurred.
<i>highest_discount_date_l</i>	The date on which the highest discount percentage last occurred.
<i>thr_25p</i>	The date on which the 25% discount percentage was first reached or breached.
<i>thr_50p</i>	The date on which the 50% discount percentage was first reached or breached.
<i>thr_75p</i>	The date on which the 75% discount percentage was first reached or breached.
<i>thr_90p</i>	The date on which the 90% discount percentage was first reached or breached.
<i>price_changes</i>	Only filled in if a game has changed the standard price once or more. It includes the price changes and a color code to indicate how it affects the data (see Appendix D).

Appendix B: Discount Strategies

Below are several discount strategies that were found during data gathering. The graphs shown are from *steamdb.info*, which was the primary source for the price history of Steam games. Each figure has a short explanation of the pattern and hyperlinks to one or more games that followed it, where the first game listed is showcased in the corresponding figure.

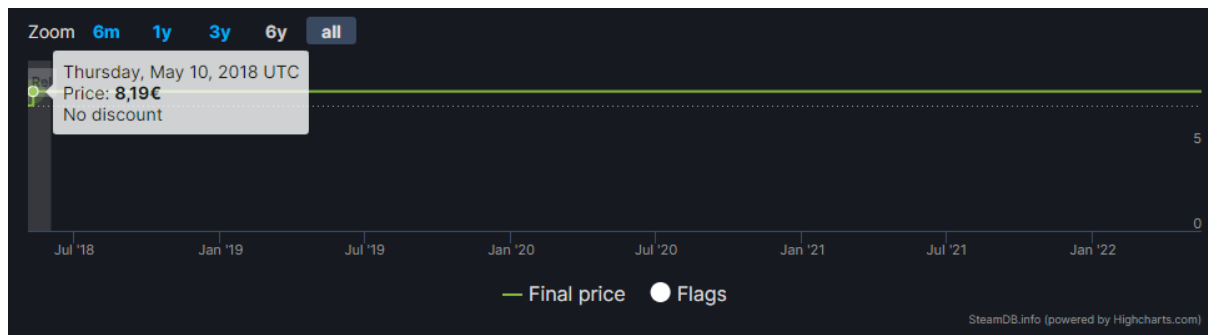


Figure 4: A launch discount, then forever standard. *Kemonomichi-White Moment-*, *My Maid Girlfriend* and *Stand by you*.



Figure 5: An example of perfect staircase/progressive discounting. *Nightshade*, *Tokyo Xanadu eX+* and *Omega Labyrinth Life*.



Figure 6: Heavy discounting (~50%), a flash sale of 90%, and then alternating discounts. *eden** and *Highway Blossoms*. The flash sale for both corresponds with a Humble Bundle that was active at the same time.

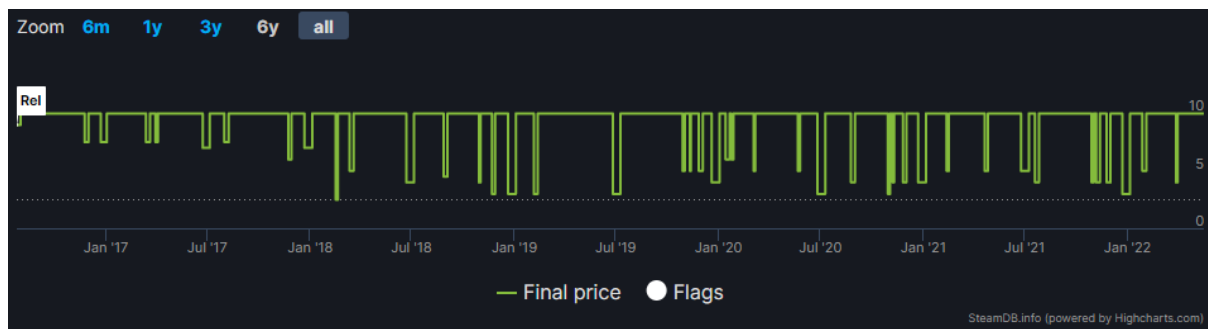


Figure 7: A sudden heavy discounting breach into alternating pattern. *Just Deserts* and *A Kiss for the Petals – Remembering How We Met*. The flash sale for both corresponds with a Humble Bundle that was active at the same time.



Figure 8: A long-lasting staircase discount pattern. *If My Heart Had Wings*, *Mugen Souls* and *A Sky Full of Stars*.



Figure 9: Little discount club. *WILL: A Wonderful World*, *Palais de Reine* and *Grisaia Phantom Trigger Vol.7*.

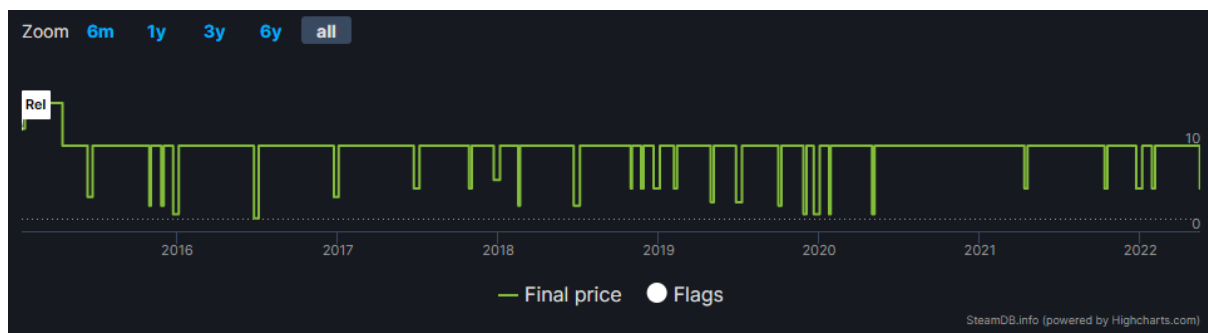


Figure 10: Experimenting with high discounts, but most likely did not increase sales much. *Alphadia Genesis* and *Dragon Sinker*.

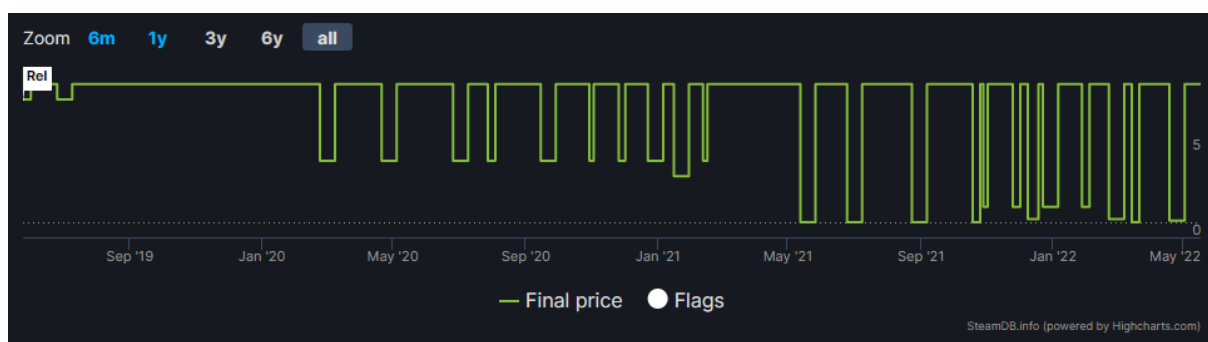


Figure 11: Calm before the storm strategy, ending at frequent 90% discounts. *WINDSHIFT*.

Appendix C: Coding Examples

Certain games were not included within the dataset or included with a caveat because of their price changes. This caveat comes in the form of a color code, indicating its severity level.

A green code indicates a price change which does not affect the data. These consist mainly of a quickly corrected launch price or very recent price change. A price change of about 10% is also labelled green, as it is reasonable to assume that this price change did not significantly affect the discounts prior to or after it occurred.

While for the green code it is relatively simple to select the standard price, this becomes more challenging in the case of multiple price changes. In order to accurately represent the data in these scenarios the orange and red codes were made. The orange code consists of the games that did not reach one of the tracked discount thresholds after their price changes. By selecting the earliest standard price where all relevant thresholds were reached, the time till thresholds remains accurate and usable. The caveat is that only the data of the chosen standard price is used, so data after a price change – the cut-off point – is not taken into account. This makes the highest discount percentage reached, including the first and last occurrence flawed. When doing calculations with the dates of when highest discounts first and last occurred, rows with the orange label should be excluded.

A red code was given to games that reached a new tracked threshold, such as 50% or 75%, after a price change. There was no cut-off point. Instead, the last threshold reached did not technically occur for the chosen standard price in the data. However, as a 90% discount on €20 gives less profit than a 90% discount on €30, it was argued that the discount percentage would have been reached anyway.

Green [29 games]

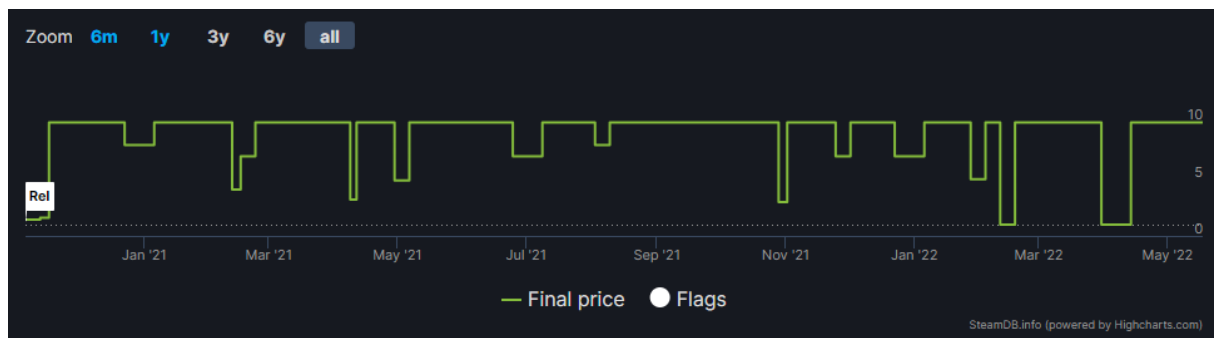


Figure 12: An extreme example of correcting the starting price. As no/few discounts are in that short phase, and afterwards it stays the same price, we will pretend it started at that later price instead. The launch discount stays. The Remission of Sins.



Figure 13: The price was changed only recently and does not affect the data in any way, as no new discount percentage was reached. The price change happened when a remastered version released. How to Fool a Liar King.

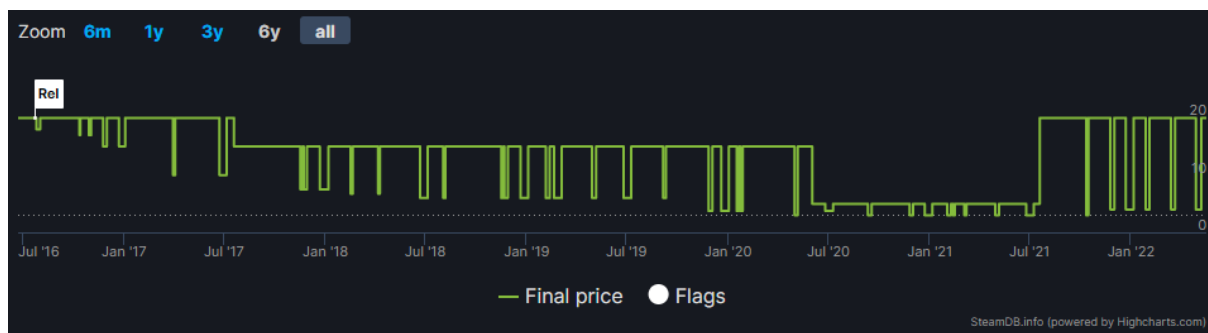


Figure 14: Although drastic price changes occurred, it ends at the same price as it started. Since a 75% discount was reached at the lower price, it would be reasonable to assume it would have occurred with the higher price as well, as that would be a discounted price of €4.99 instead of the €3.74 reached. So we will pretend the price only ever was €19.99. Phantom Brave PC.

Orange [54 games]



Figure 15: All threshold milestones were reached with the original price point, so after the first price reduction in late 2018 the data is cut-off. Hyperdimension Neptunia Re;Birth1.



Figure 16: A price change from €19.99 to €16.79. Could supposedly be green, but has more than a 15% change in price. Since the price in the dataset is the €19.99, a cut-off point felt suitable. Miniature Garden.

Red [8 games]



Figure 17: Decent data from the original price, but breached 75% around 2022. So the 75% and these four 80% discounts were added to the dataset. Valkyria Chronicles 4 Complete Edition.

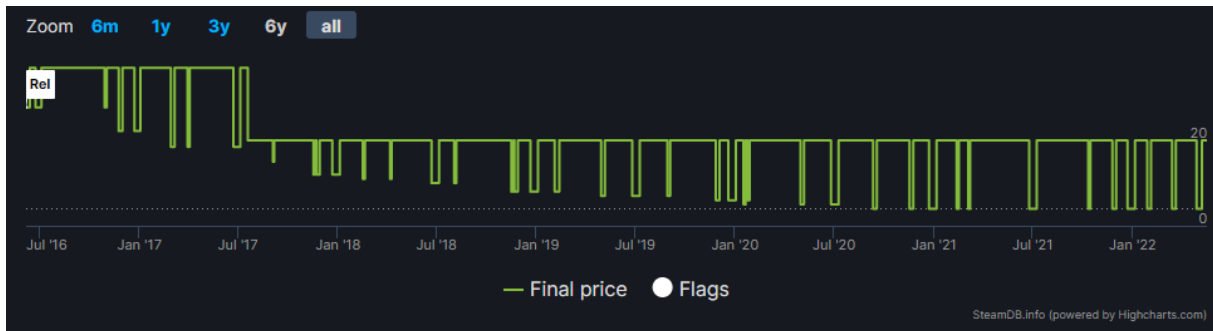


Figure 18: Changed from €36.99 to €19.99. The problem here is that it discounted fast on the higher price and slow on the newer price, so we cannot assume discounts to have been the same. So the price in the dataset is €36.99, on which it reached 50%. The new price, however, reached 75% so these had to be included. Since this was breached it was fine to change the highest threshold from 50% to 80% as well, as the highest threshold of 75% would be incorrect. Stranger of Sword City.

Not included

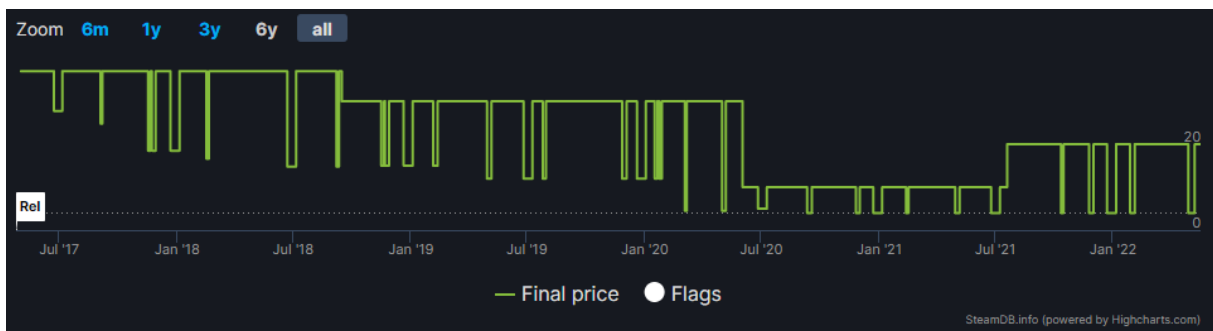


Figure 19: This one is a case of too many changes and new thresholds on these changes. It cannot be orange, and red felt off as well as the discount percentages are different yet lead to the same price point at last three price changes. Choosing a price here felt like invalid data for any choice, so this game was not included. PSYCHO-PASS: Mandatory Happiness.

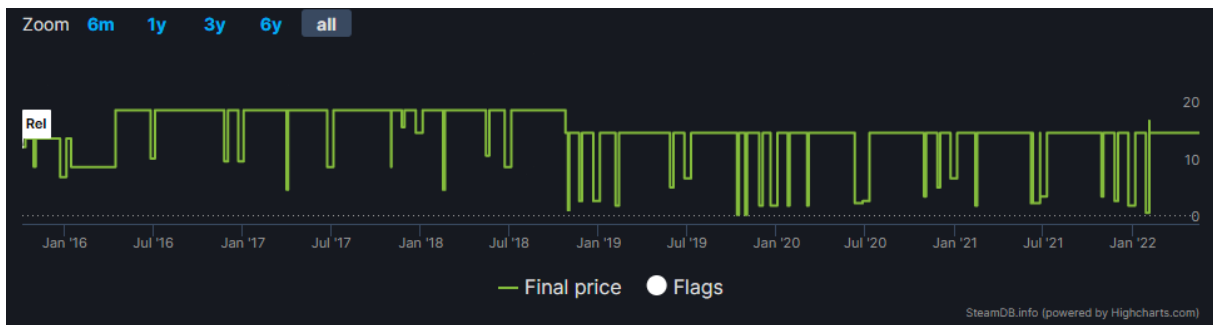


Figure 20: A tad too many frequent price changes to make assumptions about the standard price. Ignoring the first two and starting at the highest price felt wrong, as a 45% discount was already reached on the first one. Hard to correctly interpret the data for this one, so it was not included. 2064: Read Only Memories.

Appendix D – Survival Curves for each feature

25% and 50% discount developer survival curves

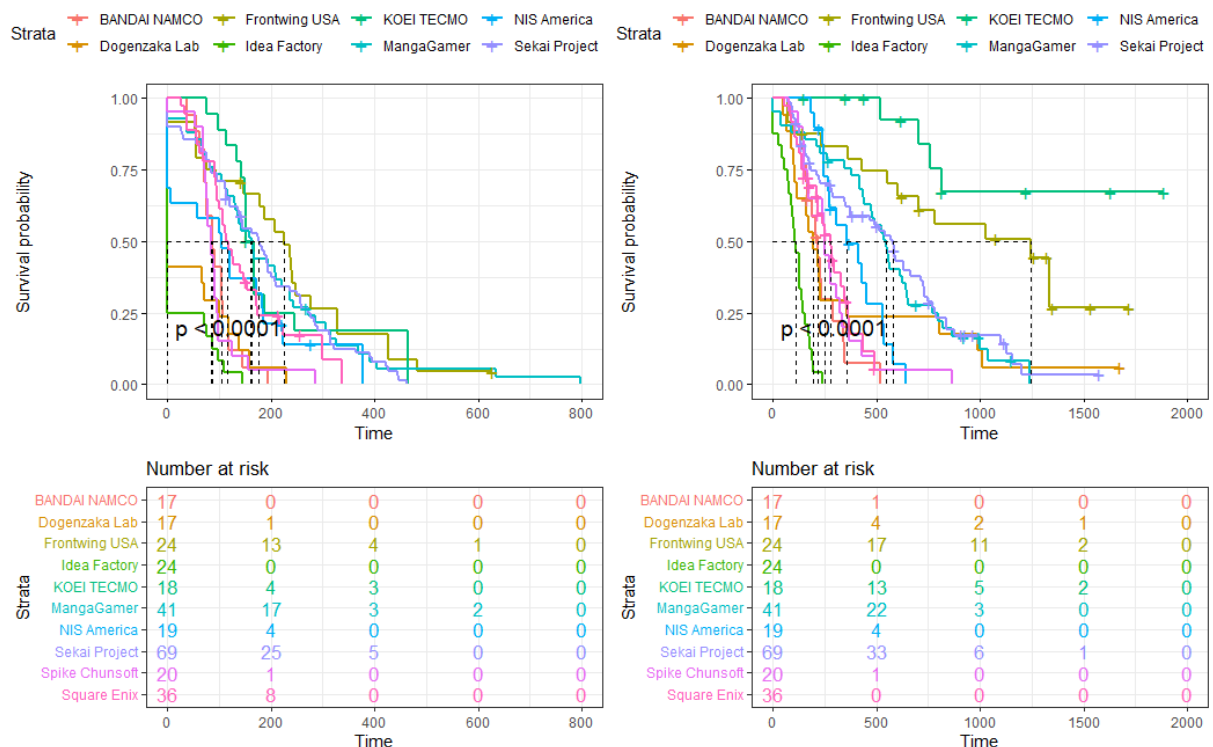


Figure 21: the 25% and 50% discount survival curves of the developer feature.

75% and 90% discount developer survival curves

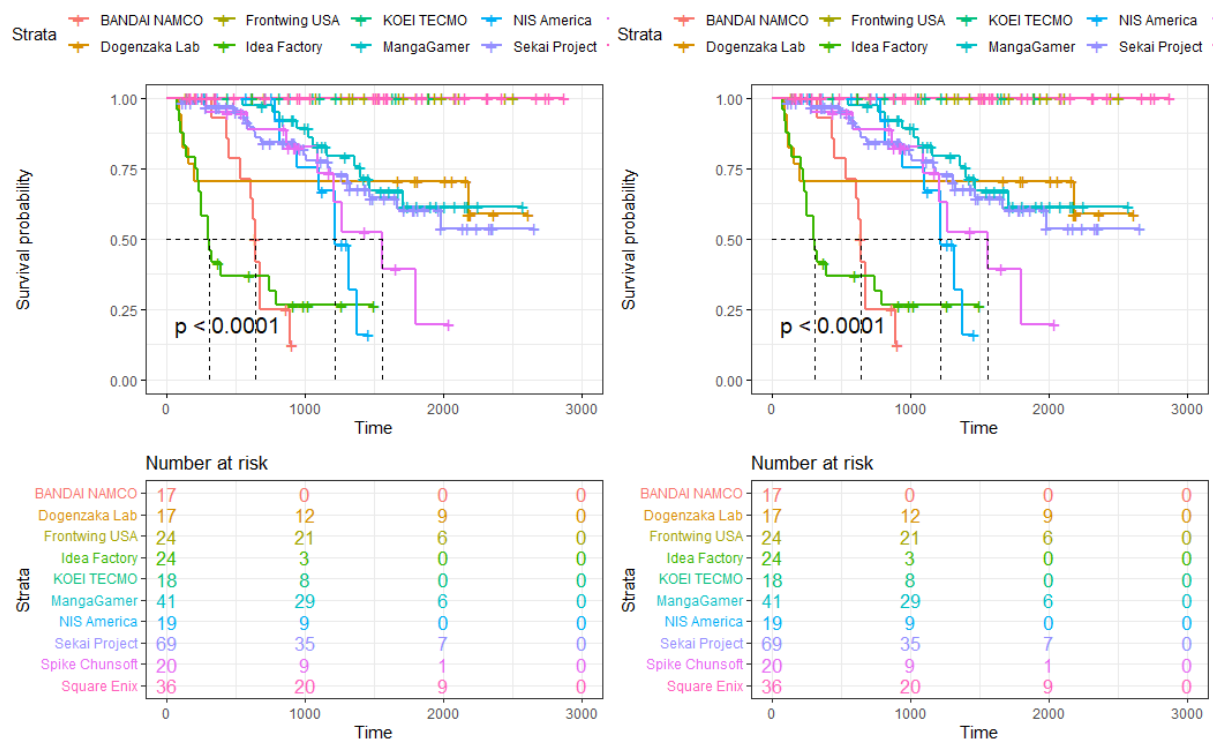


Figure 22: the 75% and 90% discount survival curves of the developer feature.

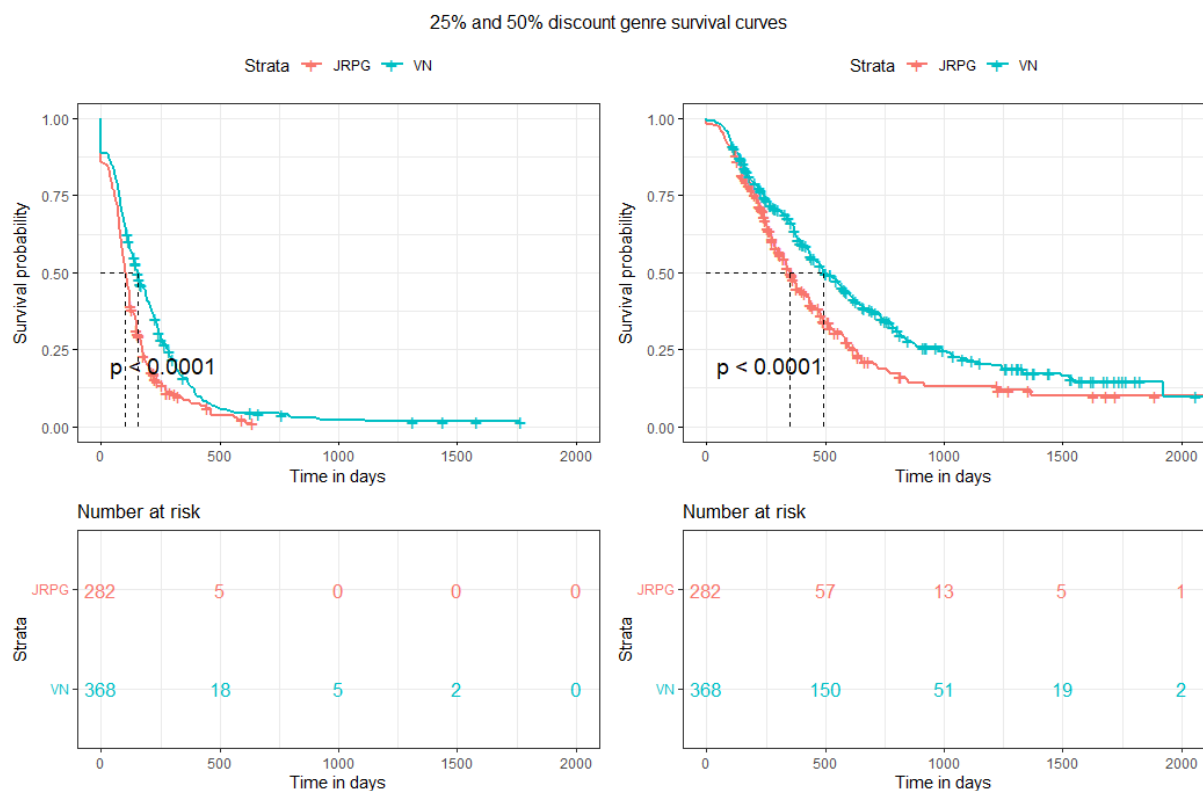


Figure 23: the 25% and 50% discount survival curves of the genre feature.

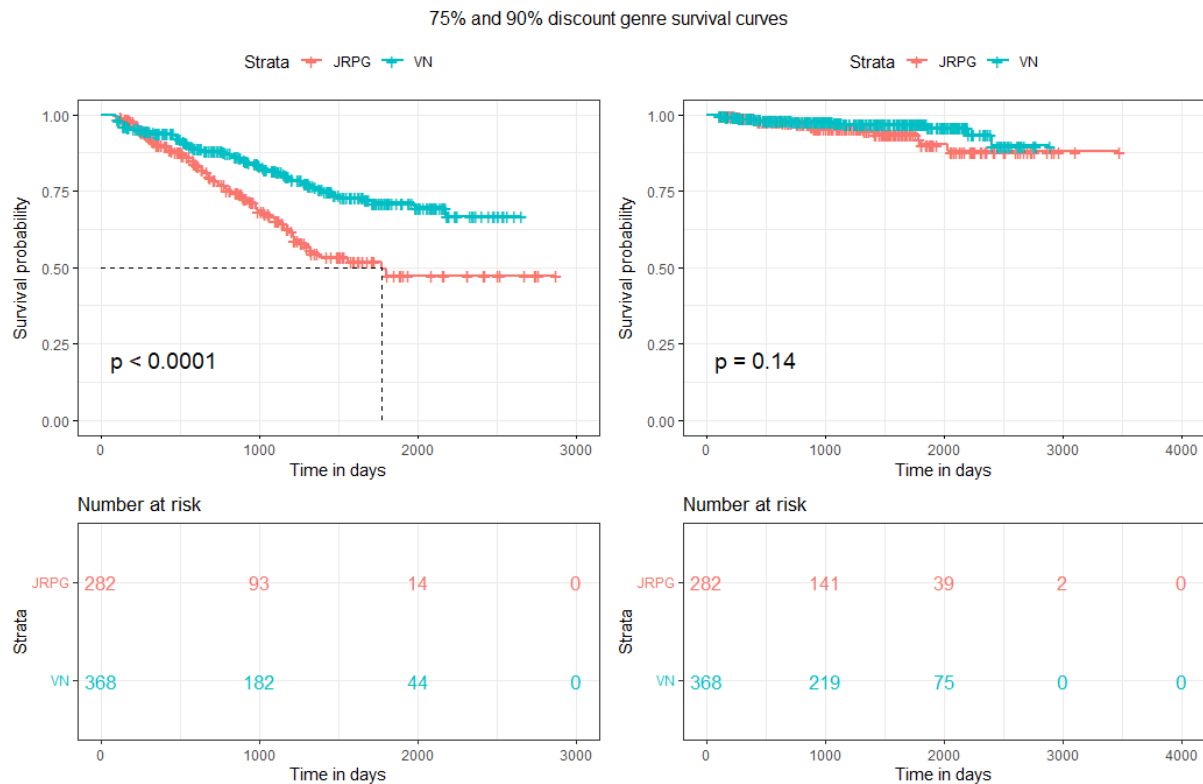


Figure 24: the 75% and 90% discount survival curves of the genre feature.

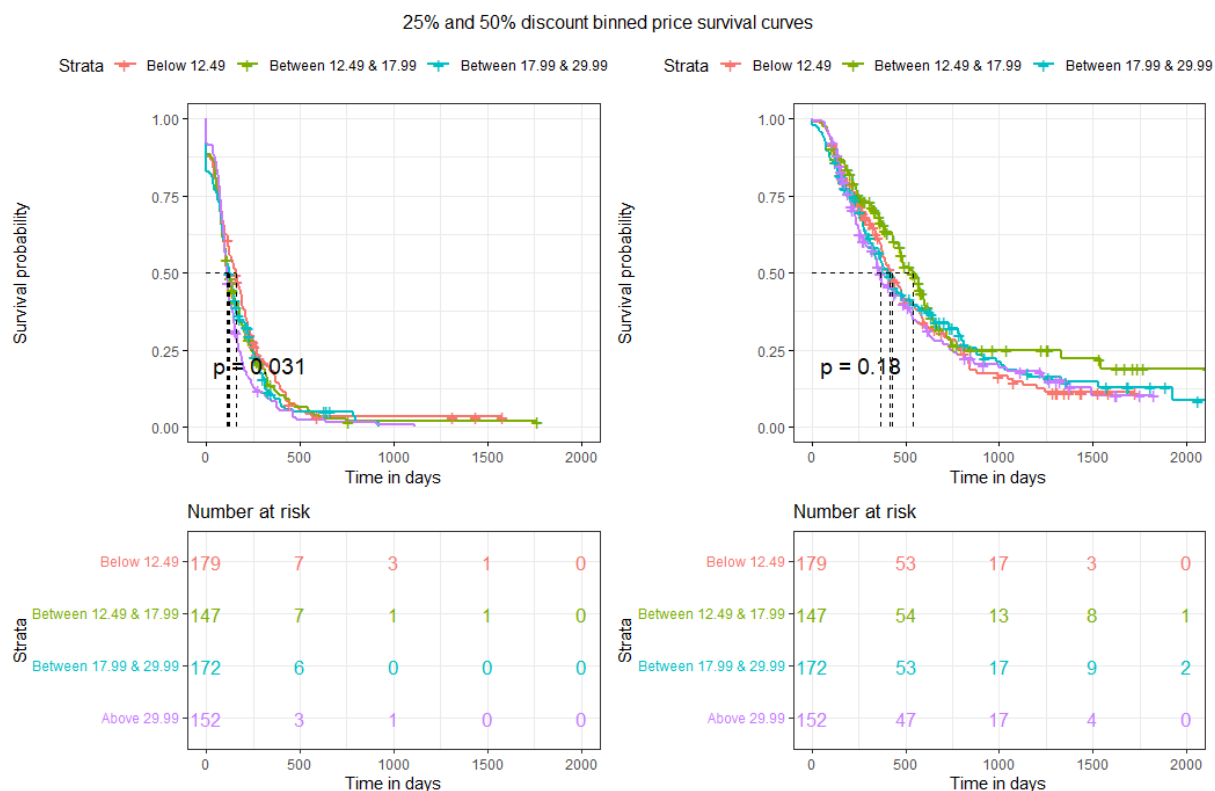


Figure 25: the 25% and 50% discount survival curves of the price feature.

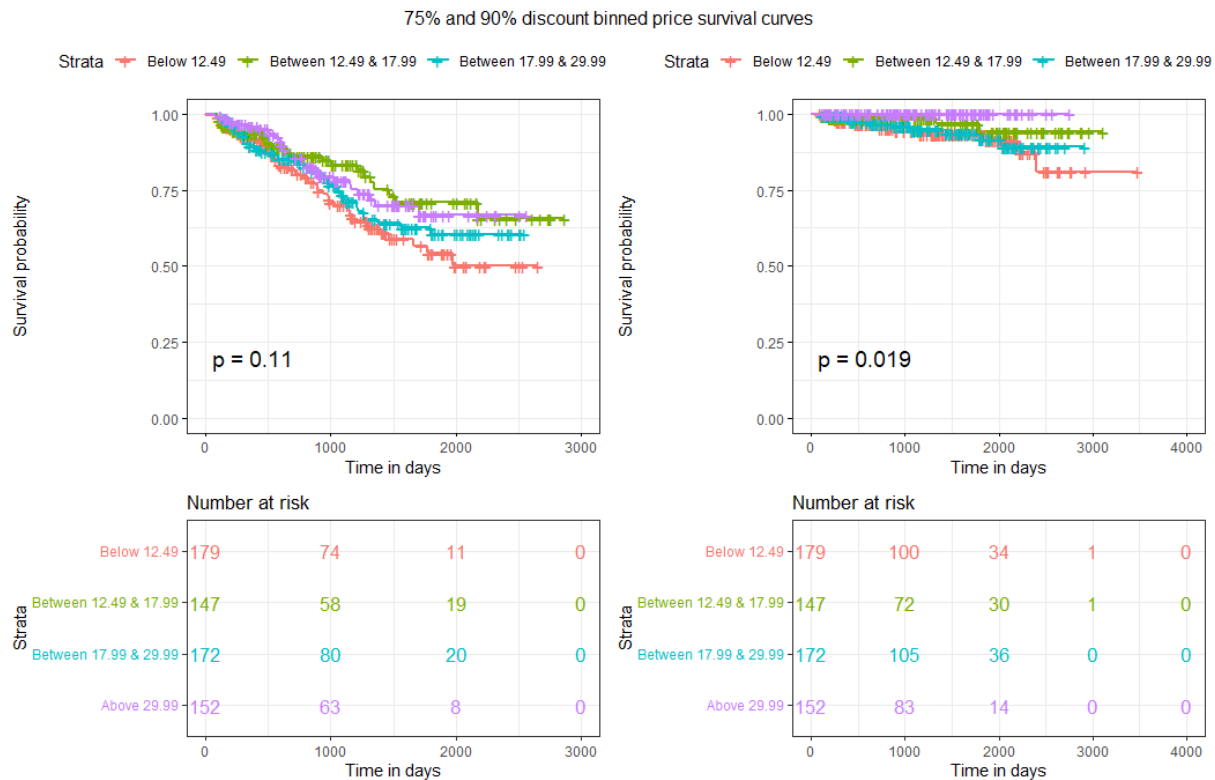


Figure 26: the 75% and 90% discount survival curves of the price feature.

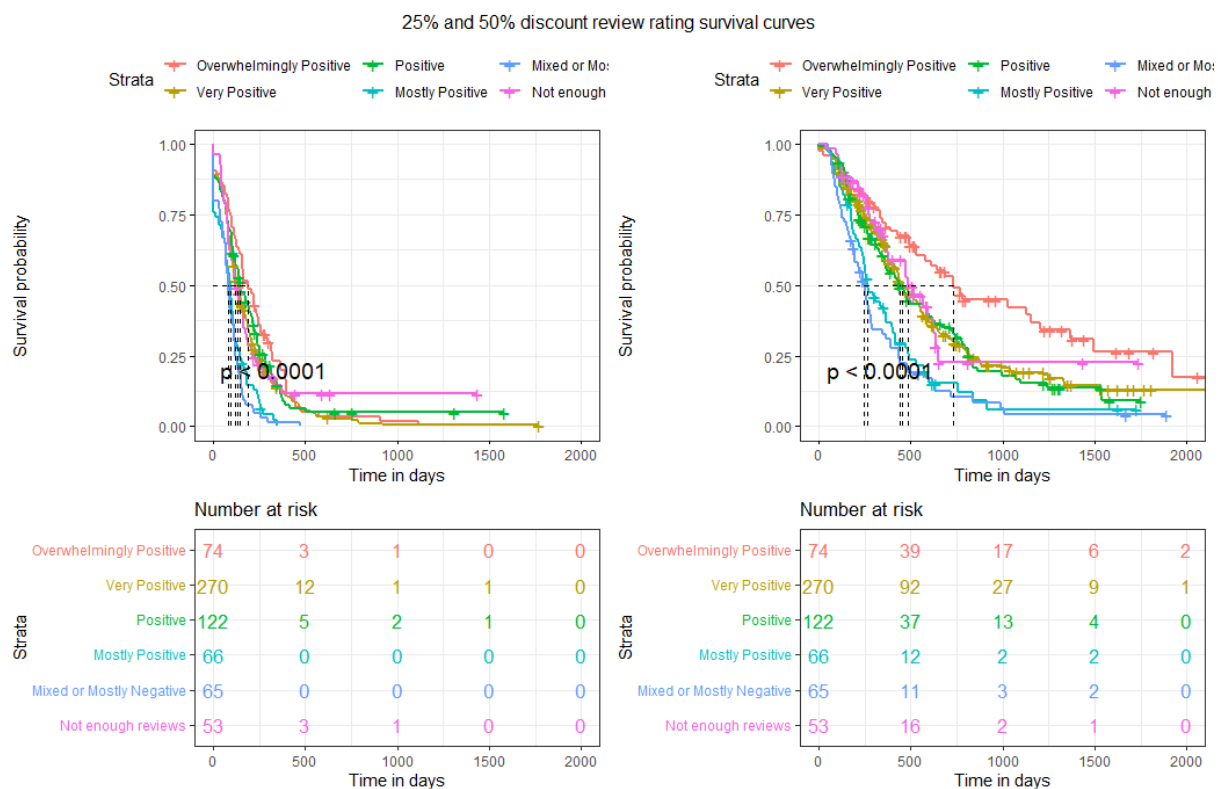


Figure 27: the 25% and 50% discount survival curves of the review ratings feature.

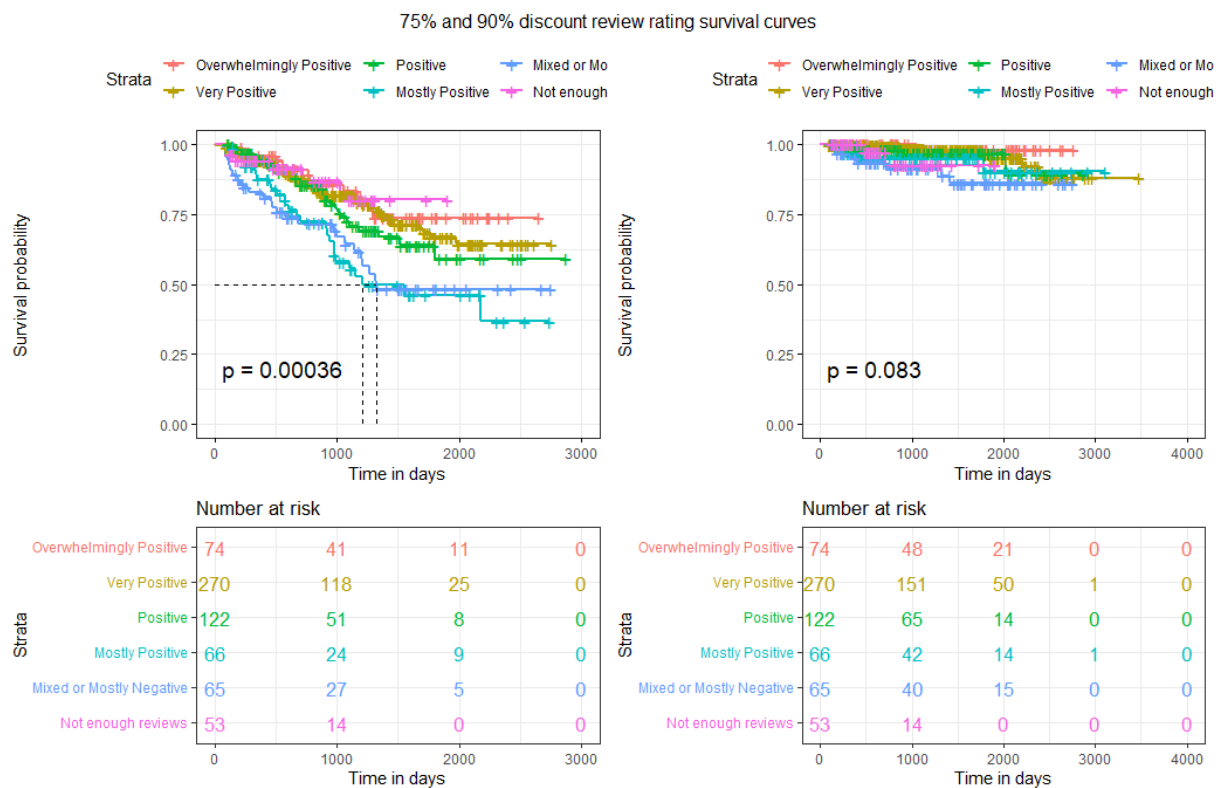


Figure 28: the 75% and 90% discount survival curves of the review ratings feature.

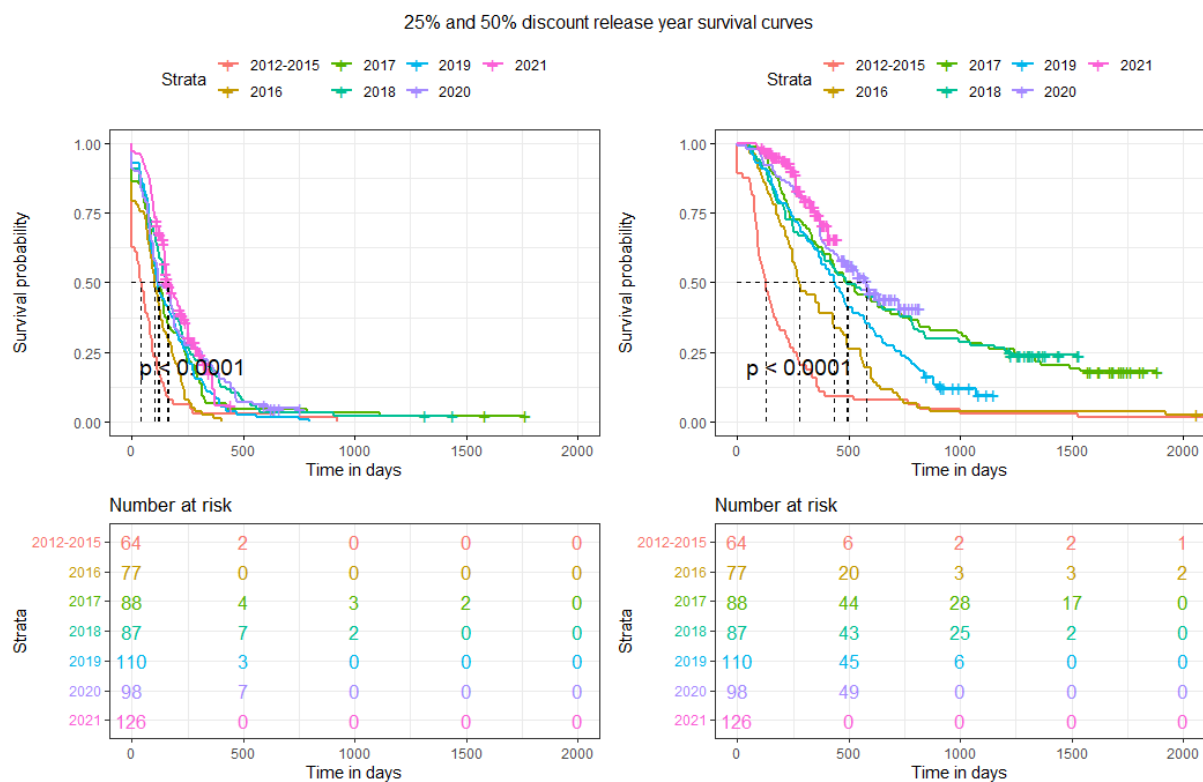


Figure 29: the 25% and 50% discount survival curves of the release year feature.

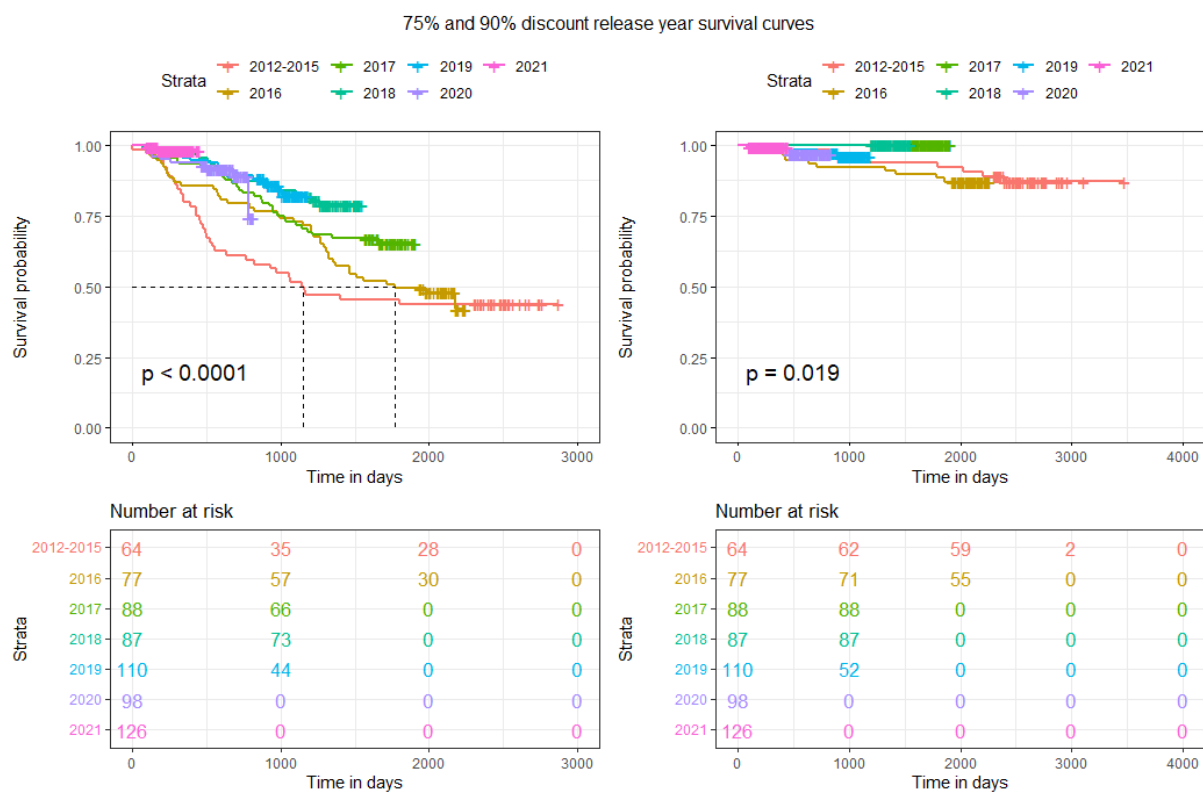


Figure 30: the 75% and 90% discount survival curves of the release year feature.

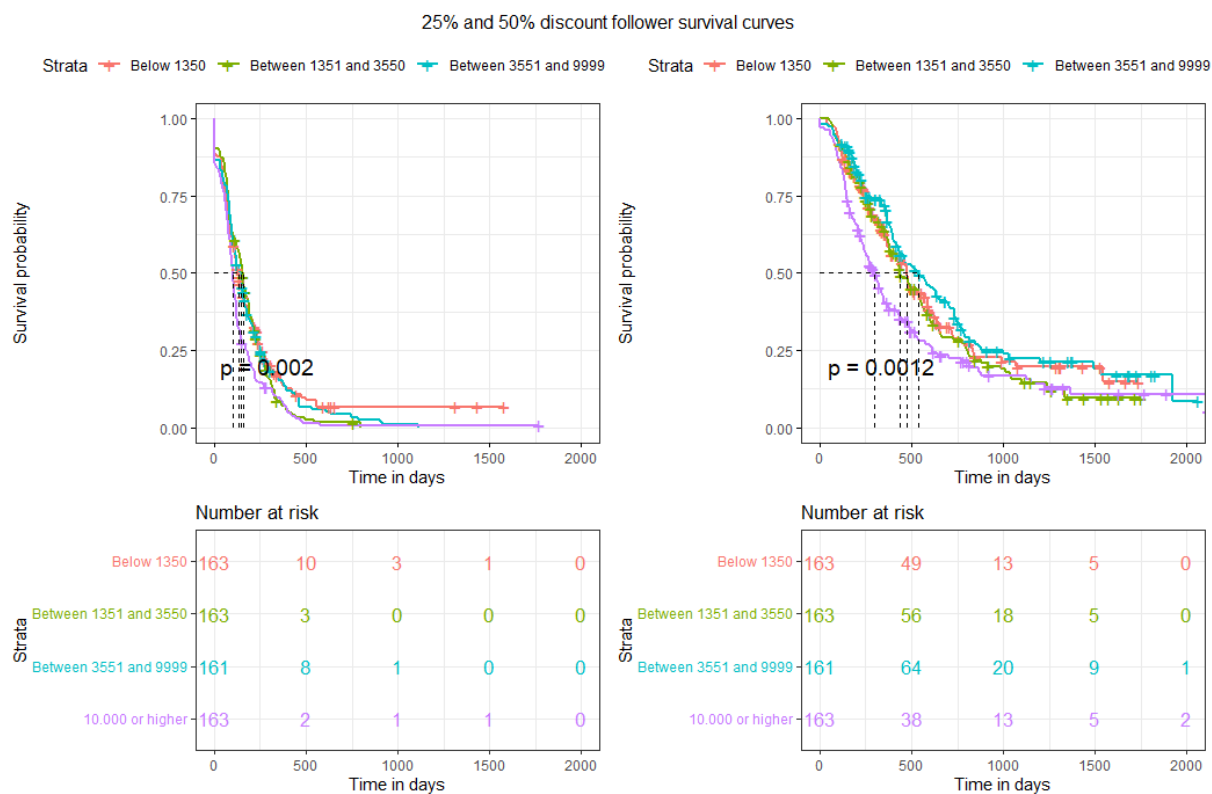


Figure 31: the 25% and 50% discount survival curves of the follower feature.

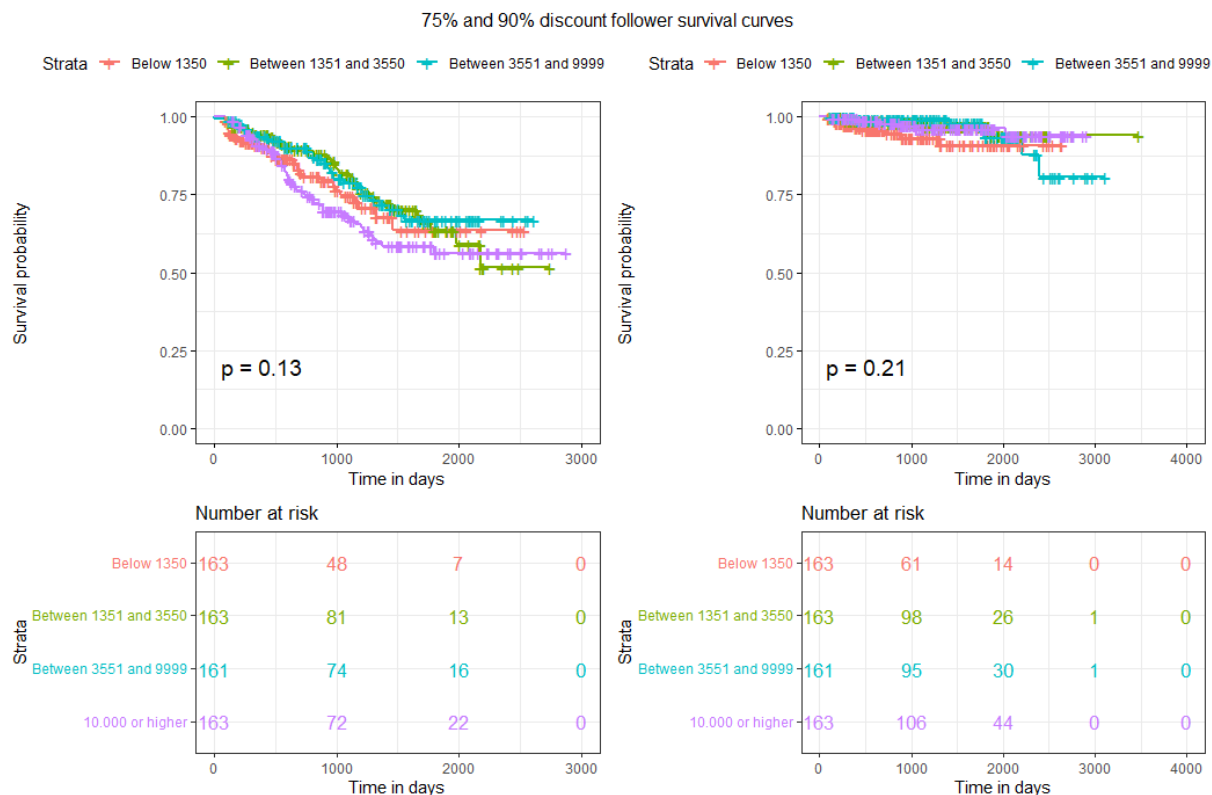


Figure 32: the 75% and 90% discount survival curves of the follower feature.

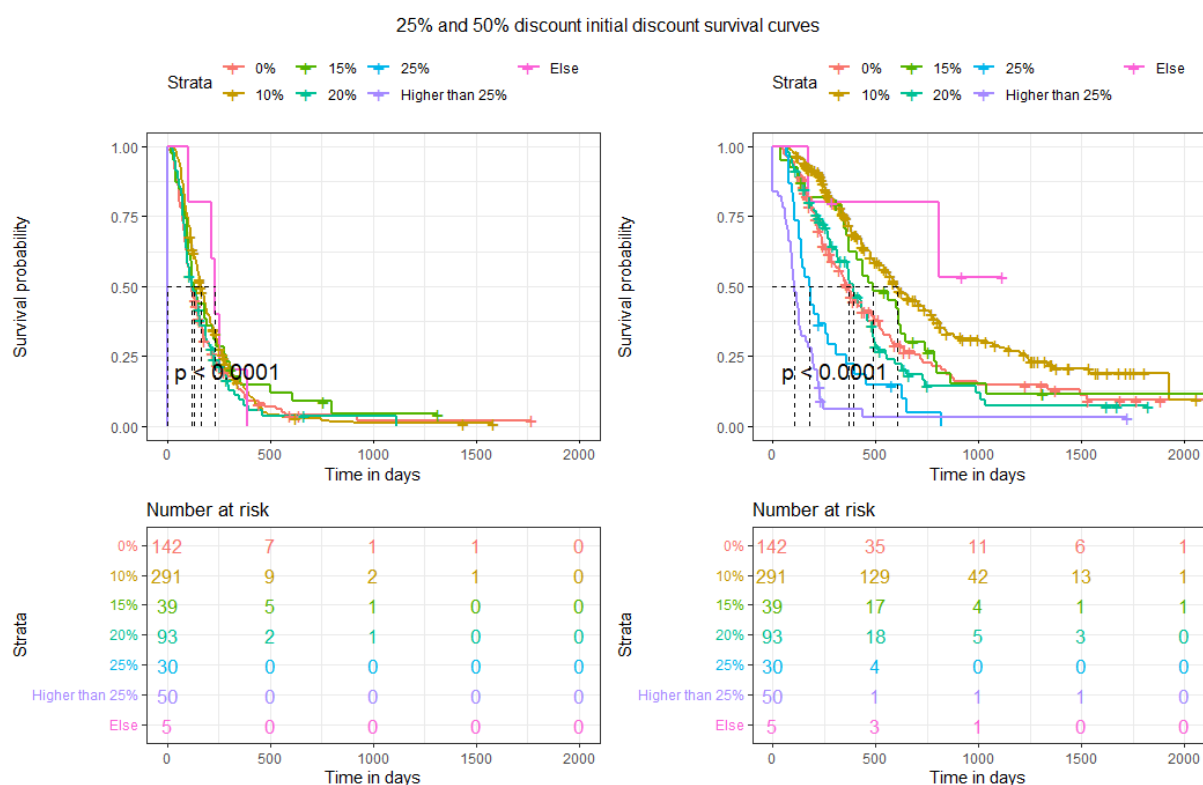


Figure 33: the 25% and 50% discount survival curves of the initial discount feature.

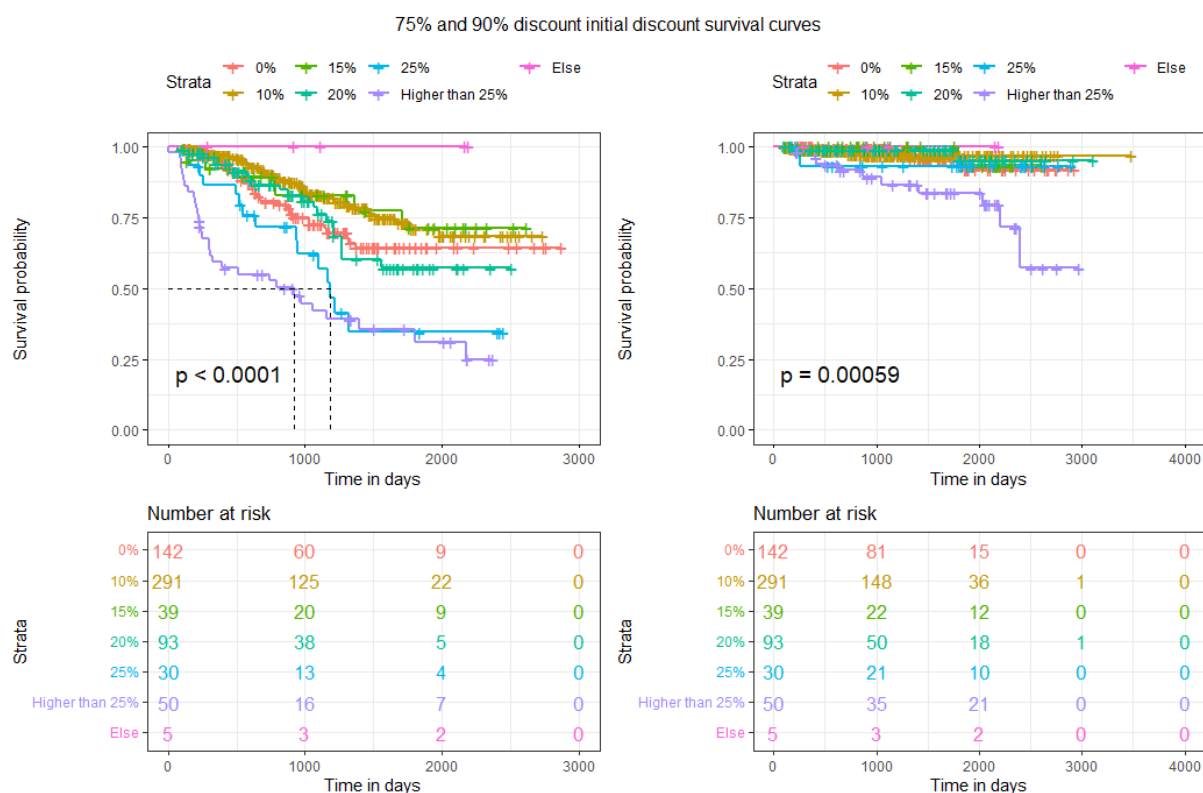


Figure 34: the 75% and 90% discount survival curves of the initial discount feature.

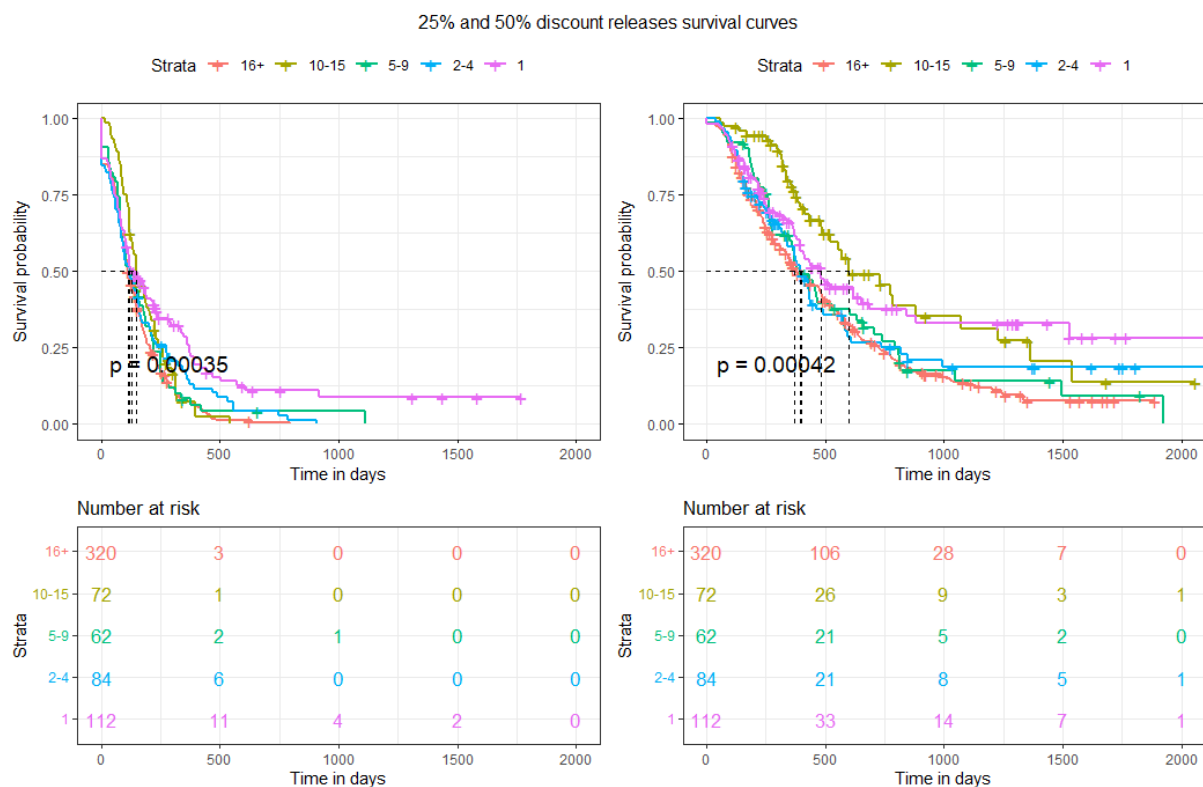


Figure 35: the 25% and 50% discount survival curves of the releases feature.

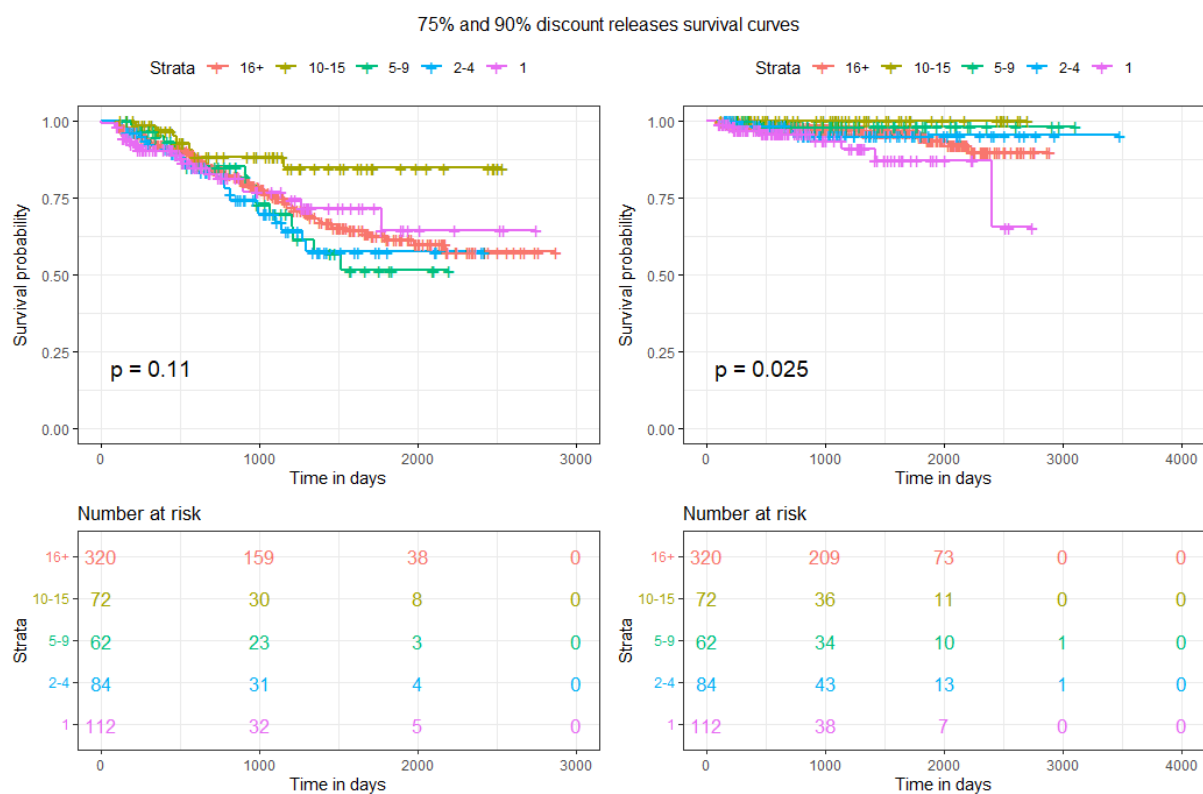


Figure 36: the 75% and 90% discount survival curves of the releases feature.

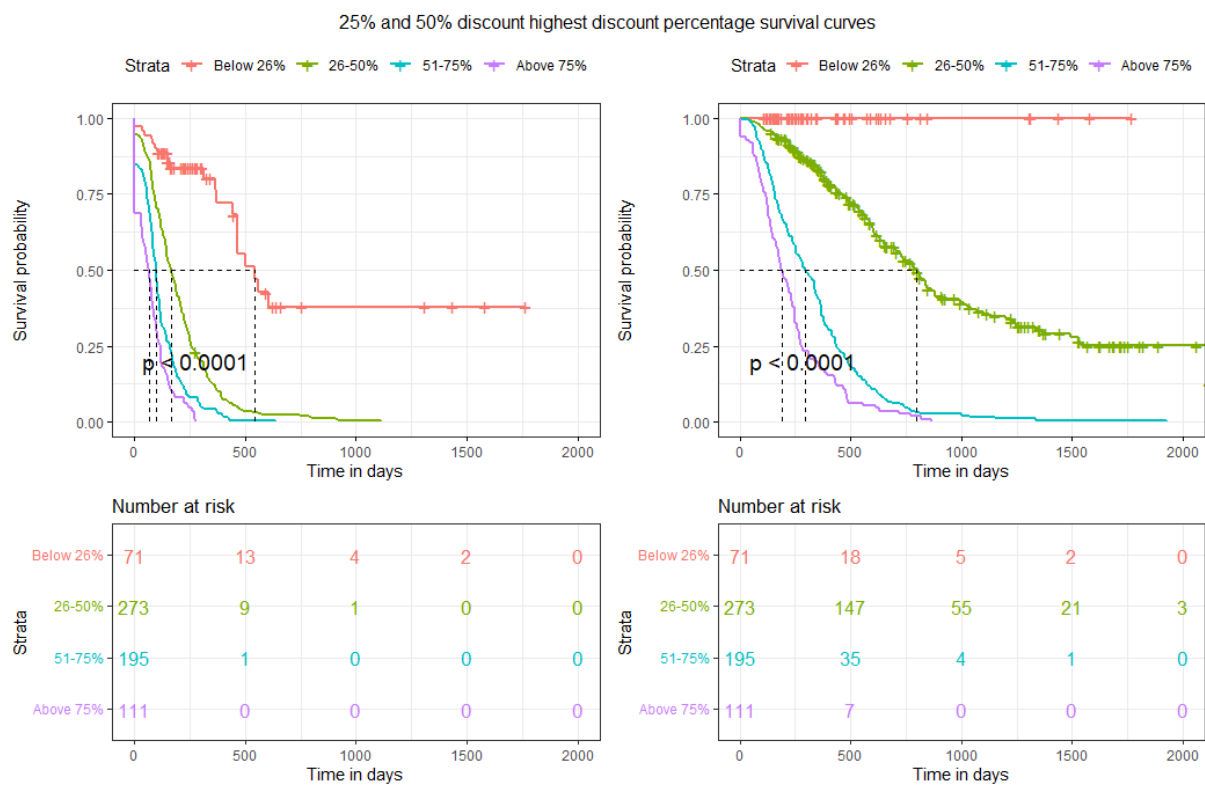


Figure 37: the 25% and 50% discount survival curves of the highest discount feature.

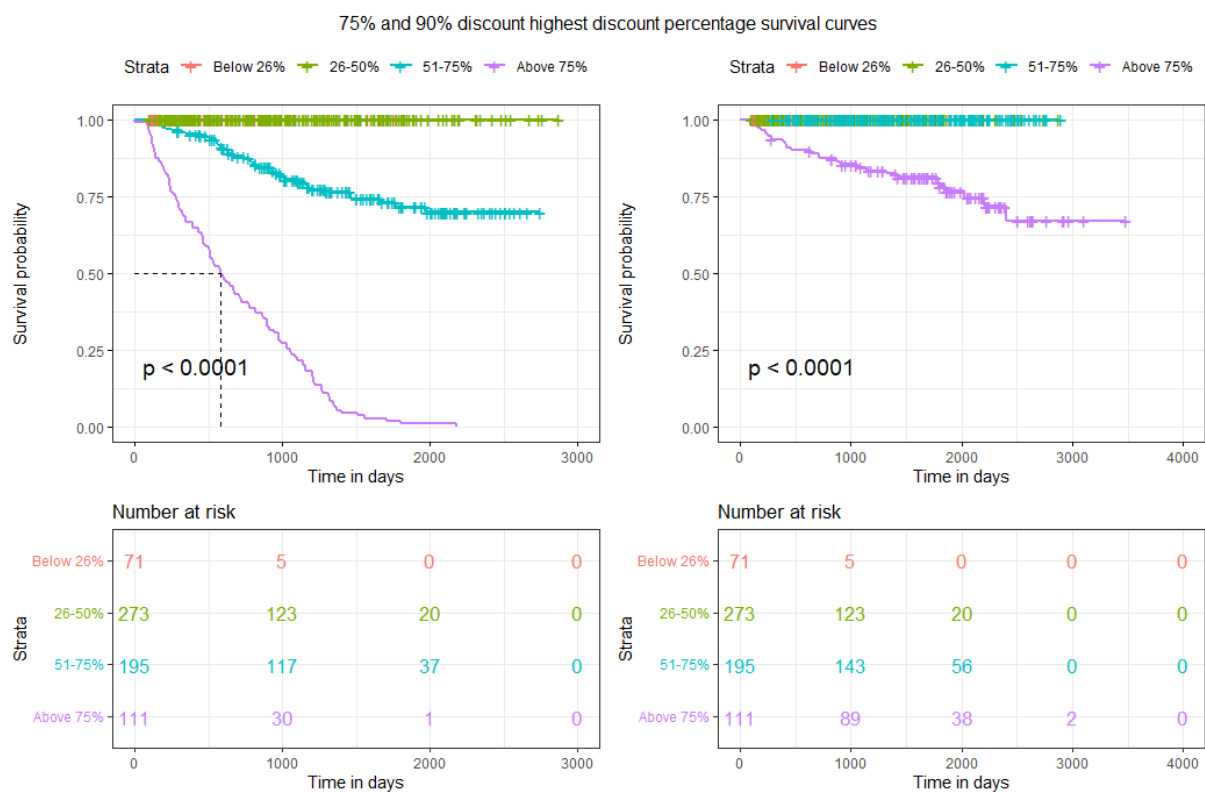


Figure 38: the 75% and 90% discount survival curves of the highest discount feature.

Appendix E: Cox Hazard Ratio Forests

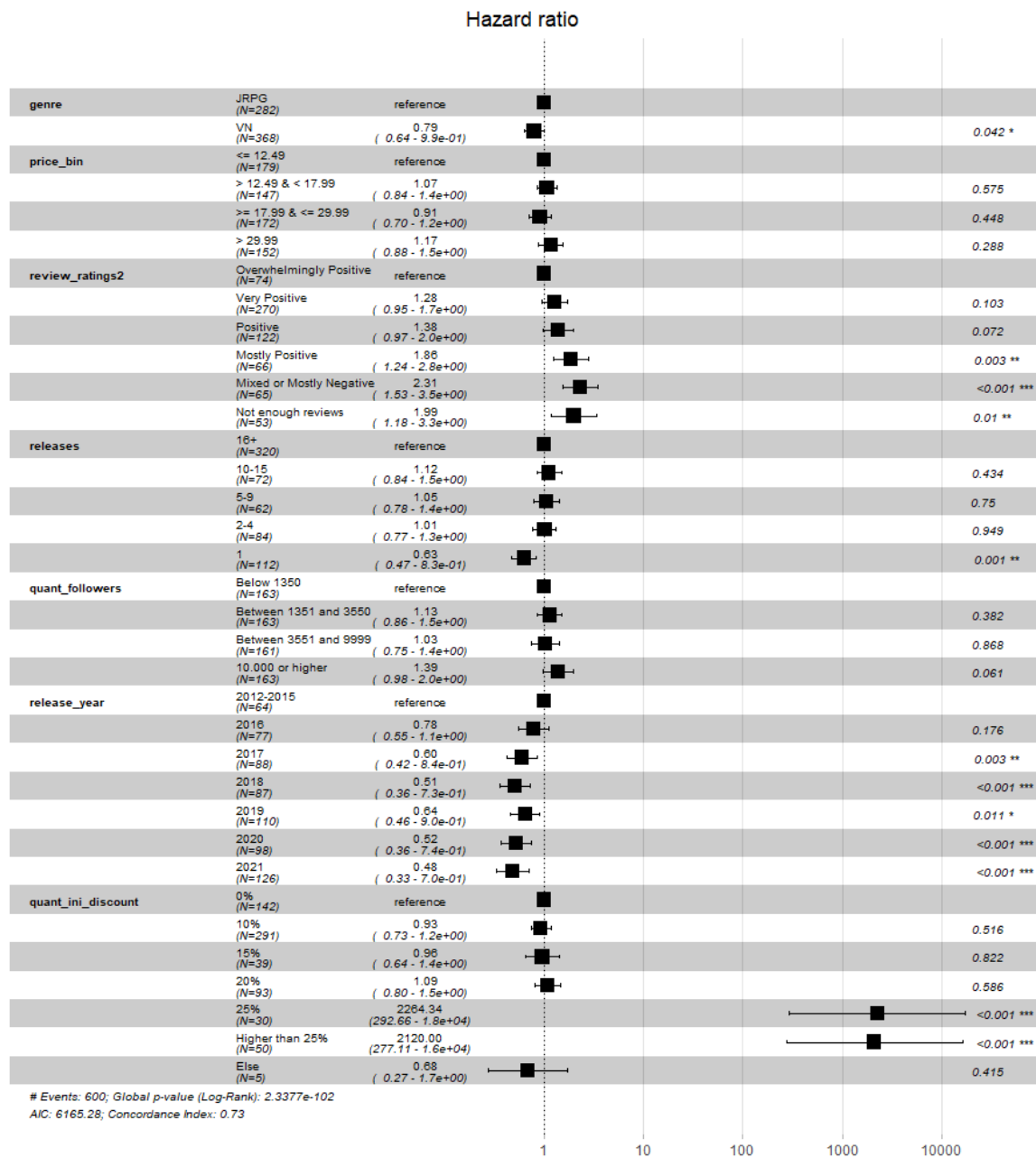


Figure 39: the Cox hazard ratio forest of the 25% discount threshold.

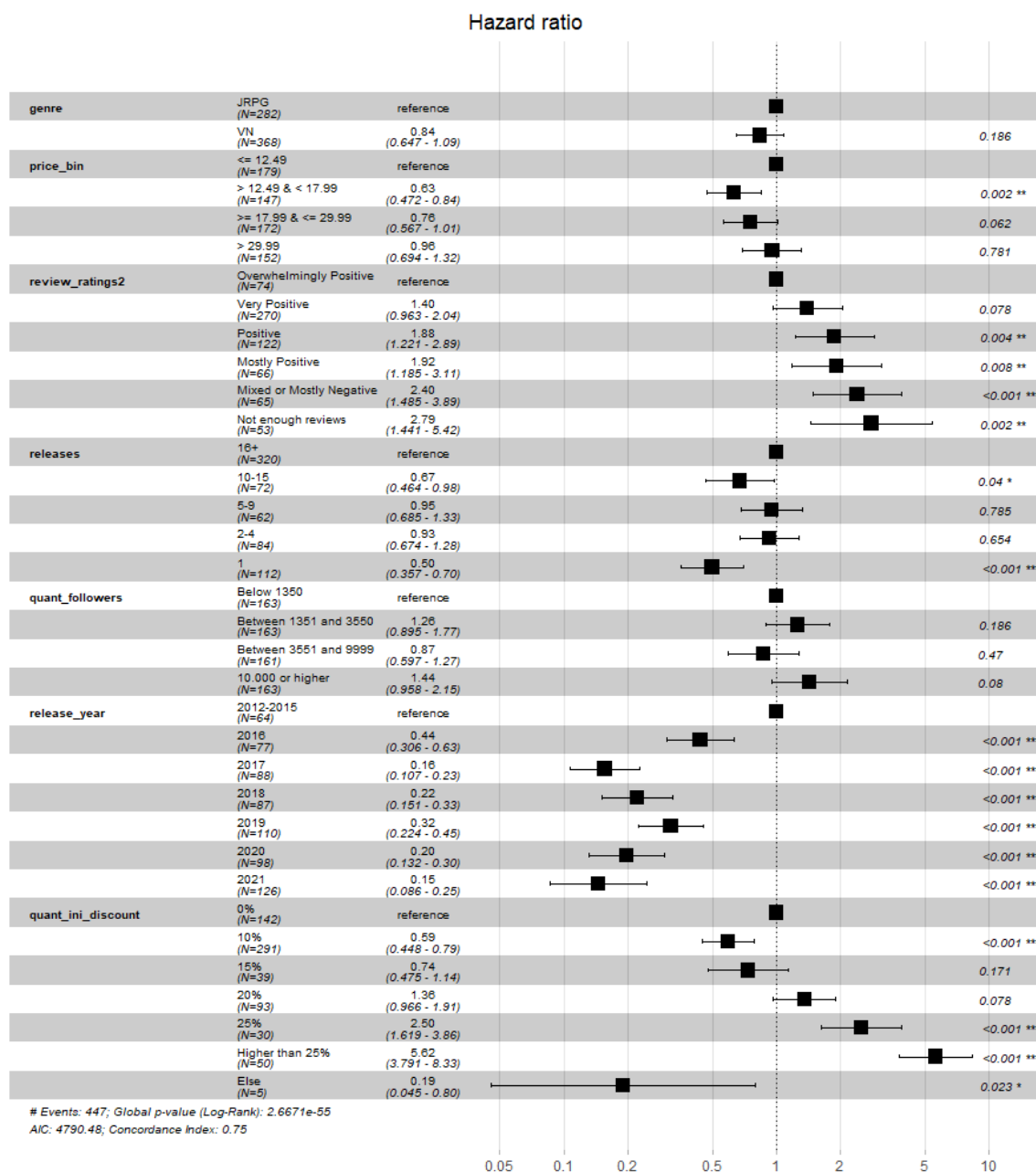


Figure 40: the Cox hazard ratio forest of the 50% discount threshold.

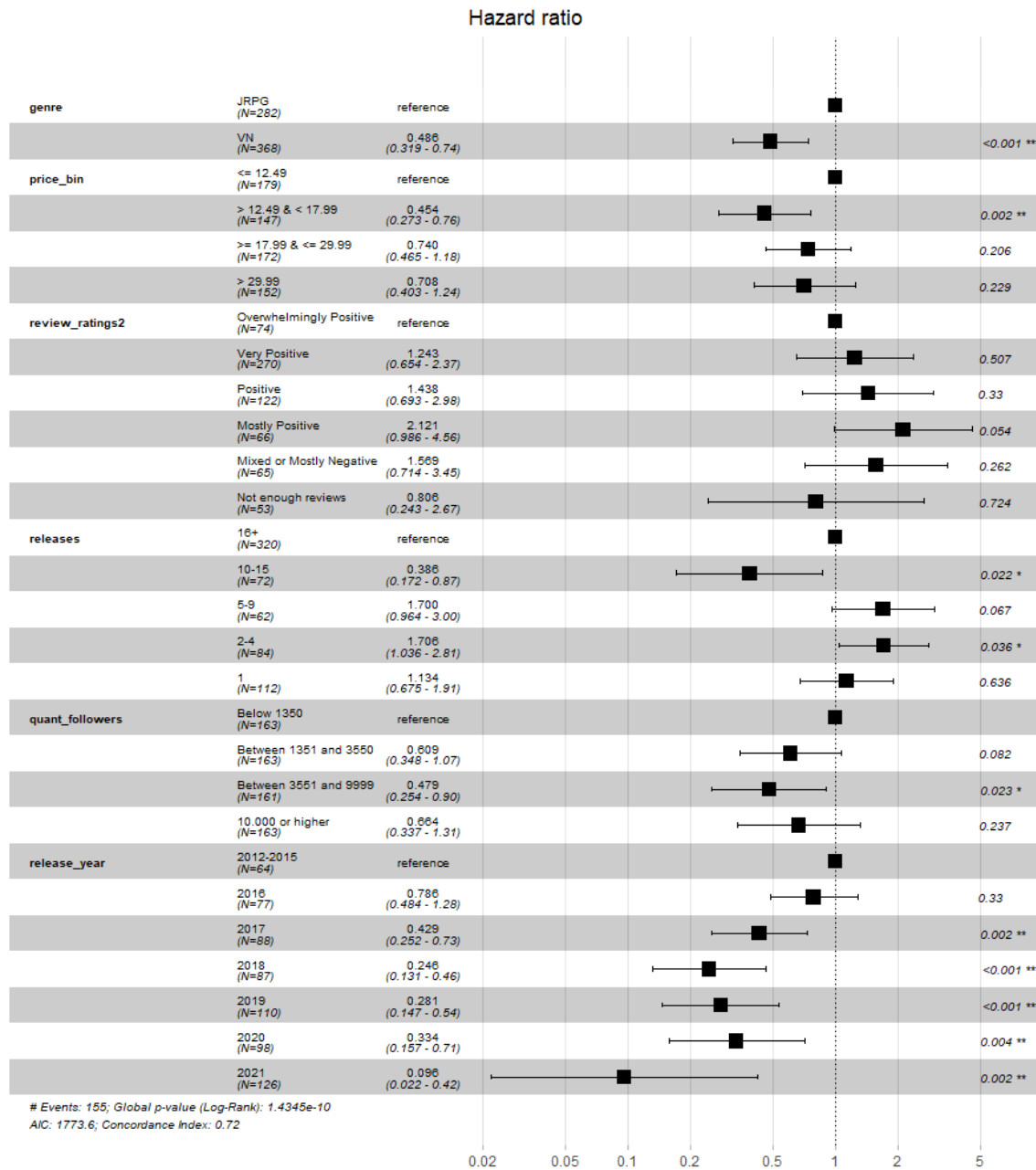


Figure 41: the Cox hazard ratio forest of the 75% discount threshold.

Appendix F: Aalen Additive Curves

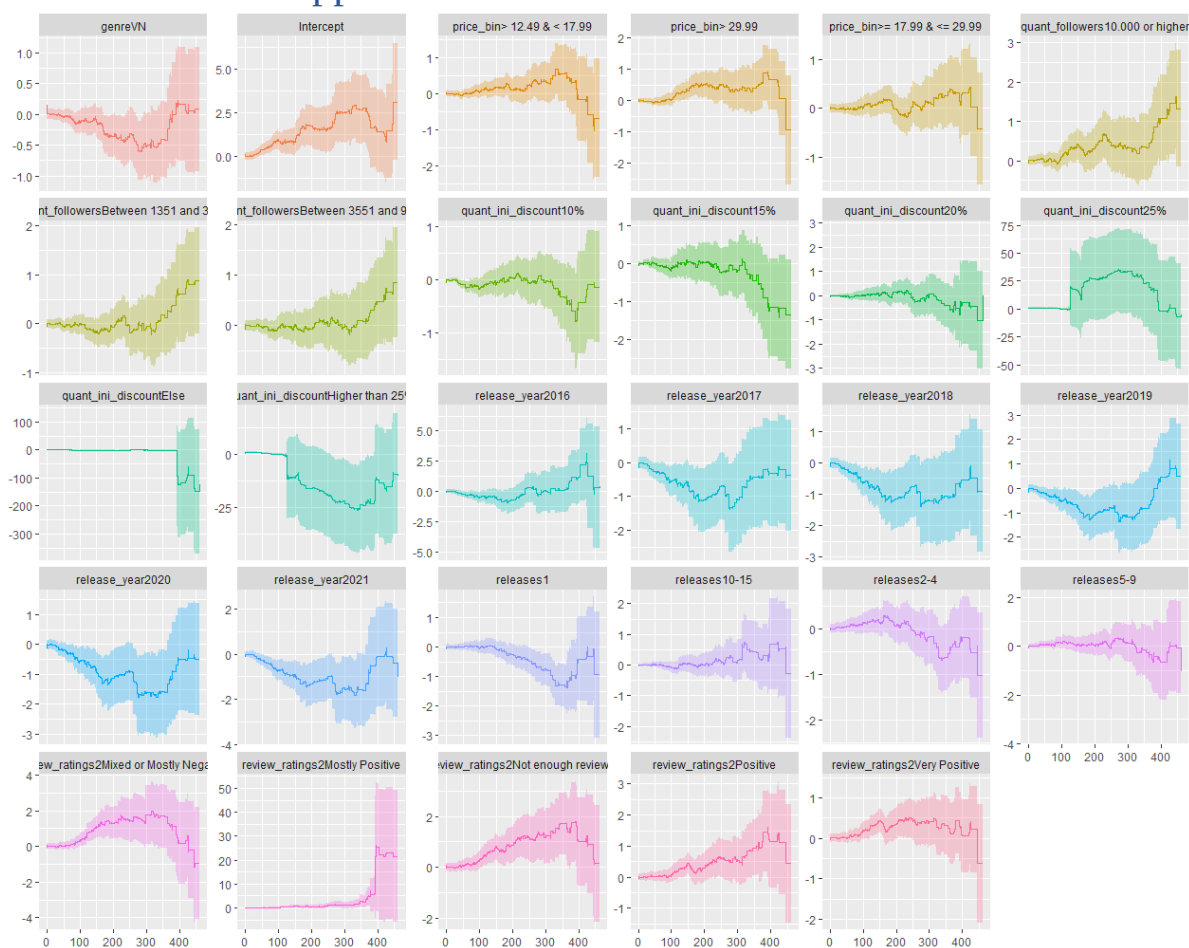


Figure 42: the Aalen additive hazard curves for the 25% discount threshold.

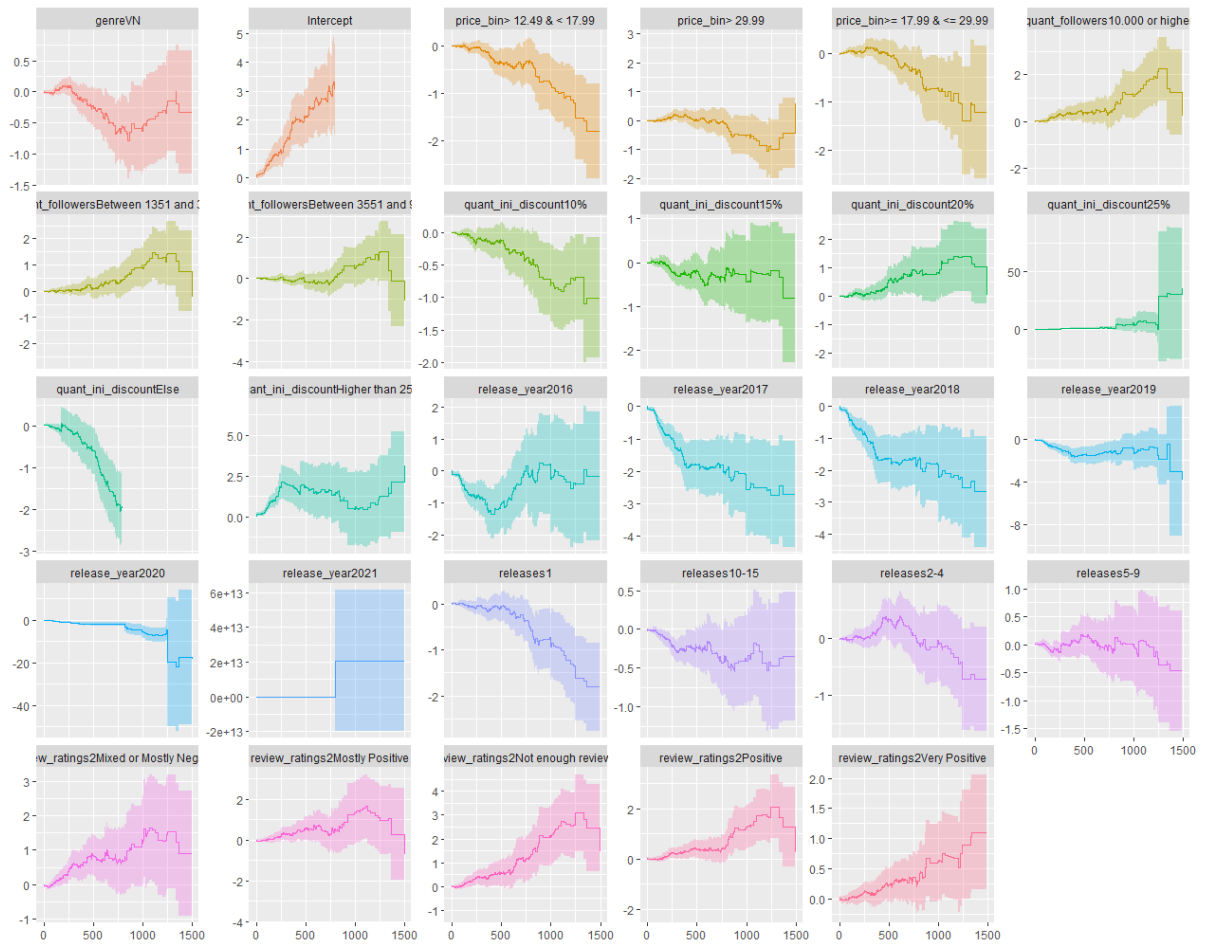


Figure 43: the Aalen additive hazard curves for the 50% discount threshold.

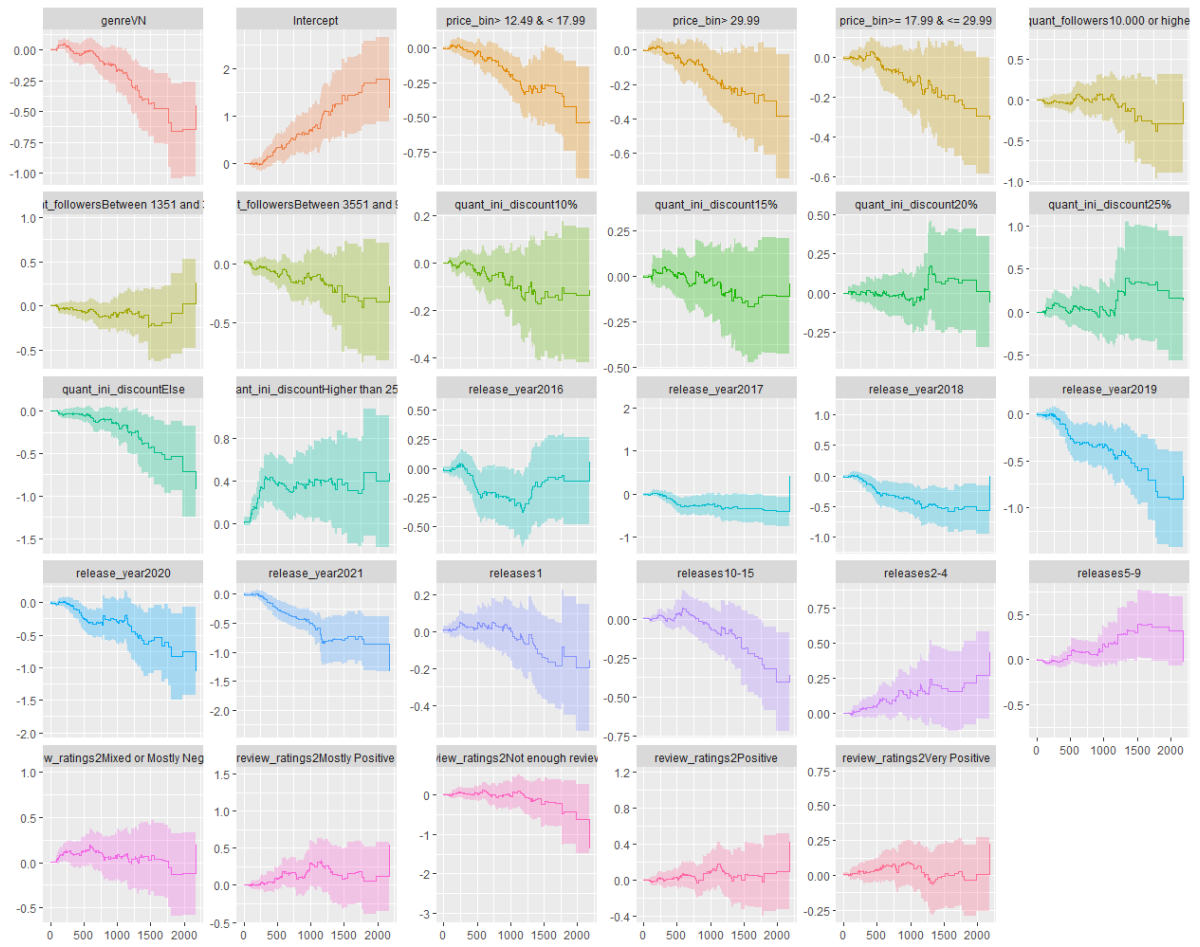


Figure 44: the Aalen additive hazard curves for the 75% discount threshold.

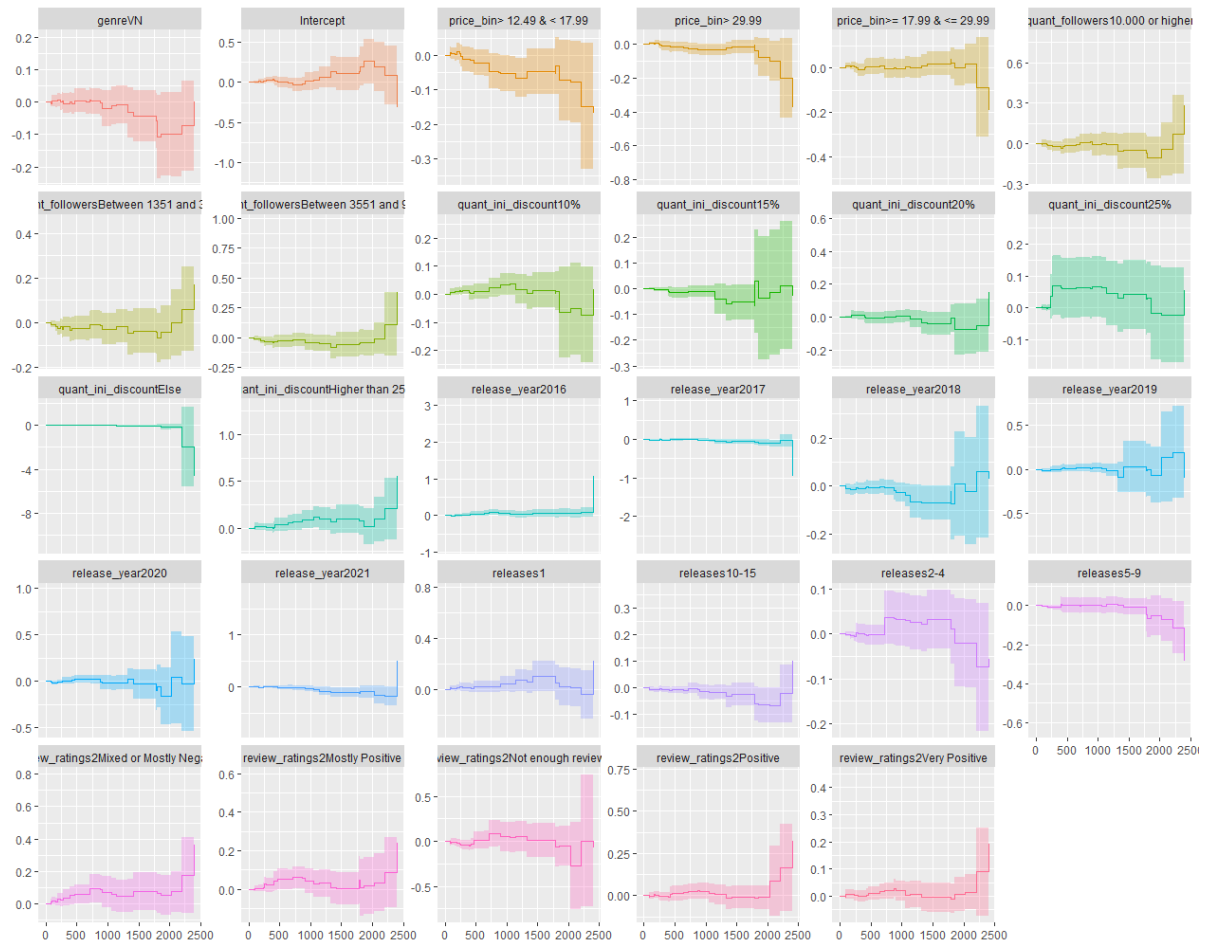


Figure 45: the Aalen additive hazard curves for the 90% discount threshold.

Appendix G: Feature Selection

Table 8

Feature selection Cox on evaluation dataset.

Label & model	Performance metrics			
	C-index R	C-index	Brier	IBS
25%				
Release year	0.604	0.634	0.213	0.130
Review ratings	0.649	0.661	0.202	0.129
Followers	0.657	0.687	0.190	0.125
50%				
Release year	0.649	0.648	0.208	0.358
Review ratings	0.687	0.707	0.194	0.344
Initial discount	0.735	0.721	0.231	0.272
Price bin	0.742	0.725	0.226	0.239
Genre	0.743	0.726	0.223	0.236
75%				
Review ratings	0.594	0.648	0.137	0.353
Initial discount	0.697	0.685	1.370	0.381
Followers	0.711	0.719	1.290	0.372
90%				
Initial discount	0.661	0.737	0.021	0.025
Price bin	0.761	0.943	0.021	0.025
Genre	0.783	0.962	0.020	0.024
Release year	0.855	0.972	0.019	0.021

Note. The C-index R is a reference performance based on the whole dataset. For the C-index a higher score is better. For the Brier and IBS score lower is better.

Table 9
Feature selection RSF on evaluation dataset.

Label & model	Performance metrics		
	C-index	Brier	IBS
25%			
Release year	0.623	0.214	0.130
Price bin	0.667	0.191	0.124
Genre	0.698	0.174	0.124
50%			
Review ratings	0.636	0.219	0.358
Initial discount	0.658	0.280	0.278
Release year	0.670	0.295	0.341
Genre	0.711	0.254	0.312
Price bin	0.696	0.282	0.273
75%			
Genre	0.624	0.141	0.357
Review ratings	0.648	0.132	0.353
Release year	0.683	0.138	0.417
Initial discount	0.737	2.421	0.556
Followers	0.743	2.179	0.479
Price bin	0.736	2.114	0.448
90%			
Initial discount	0.935	0.021	0.026
Price bin	0.960	0.018	0.025
Release year	0.957	0.017	0.024

Note. For the C-index a higher score is better. For the Brier and IBS score lower is better.

Appendix H: Full Featured Model Comparison

Table 10

Model performance comparison on evaluation data.

Metric	Models		
	Kaplan	Cox	RSF
25%			
C-index		0.689	0.657
Brier	2.733	1.936	2.357
IBS	0.15	0.125	0.15
50%			
C-index		0.728	0.693
Brier	0.294	0.230	0.279
IBS	0.296	0.245	0.275
75%			
C-index		0.701	0.736
Brier	0.899	1.877	2.114
IBS	0.333	0.458	0.448
90%			
C-index		0.927	0.943
Brier	0.022	0.021	0.020
IBS	0.027	0.026	0.031

Note. Kaplan is the baseline performance indicator. The models include all six features.

Appendix I: Survival Probabilities on unseen data

Table 11

Predicted survival probabilities compared to actual discount values.

Game title & model	Days since release for breach & survival probability			
	25%	50%	75%	90%
Highway Blossoms (VN)	159	286	552	608
Cox	0.466 [6]	0.751 [18]	0.929	0.935
RSF	0.421 [5]	0.590 [14]	0.740 [40]	0.954
Crystalline (VN)	0	240	537	X (1400)
Cox	0.936 [8]	0.557 [10]	0.768 [40]	1
RSF	0.886 [9]	0.081 [5]	0.620	0.977
Neptunia x SENRAN... (JRPG)	X (42)	X	X	X
Cox	0.889 [7]	[19]	- {0.721}	- {1}
RSF	0.859 [5]	[14]	- {0.885}	- {1}
Isekai Succubus ~ My... (JRPG)	48	X (158)	X	X
Cox	0.898 [8]	0.938 [23]	- {0.670}	- {0.934}
RSF	0.801 [7]	0.871 [18]	- {0.733}	- {0.988}
NEKO-MIMI SWEET... (VN)	X (75)	X	X	X
Cox	0.844 [8]	- {0.629}	- {0.747}	- {0.968}
RSF	0.917 [11]	- {0.831}	- {0.976}	- {0.984}

Note. The values indicate the survival probability assigned to the date on which a game went on sale. The brackets behind these indicate the amount the amount of months it took for the survival probability to become 50%. When this condition was not met, the lowest survival probability was calculated and given in curly brackets.