

# Analyzing Amazon CD Reviews with Bayesian Monitoring and Machine Learning Methods

by

Eric Su

Department of Statistical Science  
Duke University

Date: \_\_\_\_\_

Approved: \_\_\_\_\_

\_\_\_\_\_  
David Banks, Advisor

\_\_\_\_\_  
Sudipta Dasmohapatra

\_\_\_\_\_  
Merlise Clyde

A thesis submitted in partial fulfillment of the  
requirements for the degree of Master of Science  
in the Department of Statistical Science  
in the Graduate School of  
Duke University

2020

ABSTRACT

Analyzing Amazon CD Reviews with Bayesian Monitoring  
and Machine Learning Methods

by

Eric Su

Department of Statistical Science  
Duke University

Date: \_\_\_\_\_

Approved:

\_\_\_\_\_  
David Banks, Advisor

\_\_\_\_\_  
Sudipta Dasmohapatra

\_\_\_\_\_  
Merlise Clyde

An abstract of a thesis submitted in partial fulfillment of the  
requirements for the degree of Master of Science  
in the Department of Statistical Science  
in the Graduate School of  
Duke University

2020

Copyright © 2020 by Eric Su  
All rights reserved

## **Abstract**

This paper analyzes customer reviews of CDs sold on Amazon.com using various statistical and machine learning methods. We investigated the distribution properties through exploratory analyses and the Bayesian monitoring method was utilized to analyze life cycles of CDs. We proposed an adjustment to the classic Bayesian monitoring technique which allows it to deal with extreme changes in data. To predict how many reviews CDs get, we compared the performances of a range of machine learning models and identified important features affecting the number of reviews using permutation importance.

# Contents

<b>Abstract</b>	<b>iv</b>
<b>List of Figures</b>	<b>vii</b>
<b>List of Tables</b>	<b>ix</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 Exploratory Analysis</b>	<b>3</b>
2.1 Distributions of the Number of Reviews . . . . .	3
2.2 Popularity of Different Genres . . . . .	5
2.3 Purity of Dominant Genres . . . . .	6
2.4 Relationship Between Genres . . . . .	7
2.5 Popularity of Cross-Genre CDs . . . . .	9
2.6 Buyers' Interest in Genres Across Time . . . . .	10
2.7 Producers' Interest in Genres Across Time . . . . .	12
<b>3 CD Life Cycle Analysis using Bayesian Monitoring</b>	<b>15</b>
3.1 Motivation . . . . .	15
3.2 Life Cycles of CD Reviews . . . . .	15
3.3 Bayesian Monitoring . . . . .	17
3.3.1 Introduction . . . . .	17
3.3.2 Extending Bayesian Monitoring to Count Data . . . . .	18
3.3.3 Dealing with Extreme Model Failure . . . . .	19
3.3.4 Applying Bayesian Monitoring to CD Review Data . . . . .	26
3.4 Fitting Exponential Decay Function . . . . .	27

<b>4</b>	<b>Predicting the Number of Reviews</b>	<b>29</b>
4.1	Motivation . . . . .	29
4.2	Linear Regression Models . . . . .	30
4.2.1	Classical Linear Regression . . . . .	30
4.2.2	Bayesian Linear Regression . . . . .	31
4.3	Other Machine Learning Models . . . . .	35
4.3.1	K-Nearest Neighbors (KNN) . . . . .	35
4.3.2	Decision Tree . . . . .	37
4.3.3	Random Forest . . . . .	38
4.3.4	Gradient Boosting Tree . . . . .	40
4.3.5	Neural Network . . . . .	41
4.4	Model Comparison . . . . .	42
<b>5</b>	<b>Conclusion</b>	<b>45</b>
	<b>Bibliography</b>	<b>46</b>

# List of Figures

2.1	Frequency Distribution of Reviews . . . . .	4
2.2	Frequency Distribution with Fitted Power Law Line . . . . .	5
2.3	Average Rating for CDs in Different Genres . . . . .	5
2.4	Average Number of Reviews of CDs in Different Genres . . . . .	6
2.5	Genre Purity Score . . . . .	7
2.6	Correlation Between Music Genres . . . . .	8
2.7	Jaccard Similarity Between Music genres . . . . .	9
2.8	Average Rating and Number of Reviews for CDs with Different Number of Genres . . . . .	10
2.9	CD Review Number Over Time . . . . .	11
2.10	Number of reviews per month for different genres . . . . .	12
2.11	Number of CDs Released Over Time . . . . .	13
2.12	Number of CDs Released Over Time for Different Genres . . . . .	14
3.1	Life cycle patterns of CDs . . . . .	16
3.2	Bayes factors of Standard vs. Alternative Model . . . . .	20
3.3	Model Failure with Extreme Parameter Change . . . . .	21
3.4	Adjusted Bayes factors of Standard vs. Alternative Model . . . . .	24
3.5	Adjusted Bayesian Monitoring for Extreme Parameter Change . . . . .	25
3.6	Fitted Exponential Decay Lines to CD Life Cycles . . . . .	27

4.1	Posterior Inclusion Probabilities of Predictors . . . . .	34
4.2	Permutation Importances from KNN . . . . .	37
4.3	Permutation Importances from Decision Tree . . . . .	38
4.4	Permutation Importances from Random Forest . . . . .	39
4.5	Permutation Importances from Gradient Boosting Tree . . . . .	41
4.6	Permutation Importances from Neural Network . . . . .	42
4.7	Cross Validation MSE of Machine Learning Models . . . . .	44

# List of Tables

3.1	Distributions in Bayesian Monitoring for Positive Count Data . . . . .	19
4.1	Estimated Box-Cox Power Transformation . . . . .	31

# 1

## Introduction

The number of reviews a CD gets directly reflects how popular it is and thus serve as a great indicator for customer interests. Understanding various statistical properties of the number of reviews and what affects it is key to building better recommendation systems, advertising and providing suggestions for CD producers. Our goal in this paper is to identify and demonstrate useful statistical and machine learning methods for answering those questions, and evaluate their performances through analyses of real world data.

The dataset used in this paper is the Amazon product review dataset organized by Julian McAuley, Christopher Targett, Javen Shi and Anton van den Hengel [1]. Our focus will be on all customer reviews from the product category “CDs & Vinyl”. The dataset contains customer reviews of CDs spanning from 1996 to 2014. Each review consists of the user id, the product id, review text, review title, rating, helpfulness, and the time of which the review was written. In addition, product specific information including product name, product id and the category assigned are also available. The categories assigned to the CDs are the music genres associated with the albums on the Amazon website.<sup>1</sup>

Chapter 2 is focused on exploratory analyses on the properties of CD reviews. We showed that the frequencies of reviews generally follows power law distributions, implying that only a small portion of CDs receive high numbers of reviews. Next, we explored the effects of music genres on the ratings and reviews CDs get. At the end of the chapter, we investigated the trend of CD reviews and the number of CDs

---

<sup>1</sup>Categories are as follows:

Alternative Rock, Blues, Broadway & Vocalists, Children’s Music, Christian, Classic Rock, Classical, Comedy & Spoken Word, Country, Dance & Electronic, Folk, Gospel, Holiday & Wedding, Jazz, Karaoke, Latin Music, Metal, New Age, Opera & Classical Vocal, Pop, R&B, Rap & Hip-Hop, Reggae, Rock, Soundtracks, Special Interest, World Music

released over time.

In chapter 3 we analyzed the life cycles of CDs using the Bayesian monitoring method developed by Mike West (1986) [2]. We proposed an adjustment to the method which applies discounts on the Bayes factor when extreme model failure is detected. The adjusted Bayesian monitoring technique is better suited to deal with data that has extreme, rapid changes.

Finally, we compared different models' performance in predicting how many reviews CDs get in chapter 4. Models we investigated included linear regression, Bayesian linear regression, k-nearest neighbors, decision tree, random forest, boosting trees and neural network.

## 2

# Exploratory Analysis

For exploratory data analysis, we will focus on distributions of ratings and number of reviews across different music genres and try to identify aspects of the data that are suitable for further investigation. Firstly, we investigate properties of the frequency distribution of Amazon reviews and fit power law distributions to it. In the next two sections we investigate the ratings and number of reviews of different genres, then we look at how dominant each genre is for listeners who prefer it. Next we will use correlation to find out which genres are often reviewed by the same customers. In the fifth section we will look at whether albums with more than one genre labelled got better ratings or more reviews. Finally, the last two sections in this chapter are dedicated to finding how the popularity of music genres change over time.

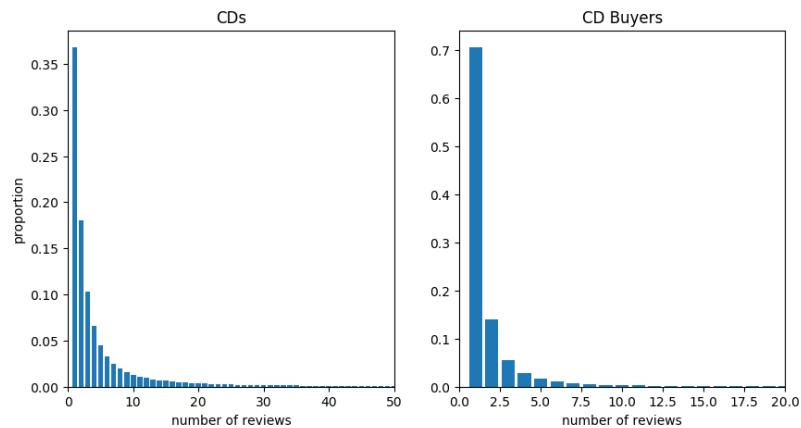
## 2.1 Distributions of the Number of Reviews

Understanding the distribution of the number of reviews gives us important information regarding how popular a CD is likely to be. We plotted the frequencies of reviews using barplots, as shown in Figure 2.1. The frequencies of reviews show a decreasing pattern as the number of reviews increases. This pattern is similar to power law distributions and thus we fitted power law distributions to the frequency of the reviews using the `powerlaw` package in Python [3], shown in Figure 2.2. A power law distribution with  $\alpha = 1.96$  fits decently well to the frequency of CD reviews, while a power law distribution does not seem like a good fit for the review frequency of buyers. Notice that the fitted power law distribution has a much larger  $\alpha$  parameter at the tails than the actual data. This is because the frequency of reviews decreases extremely fast at the beginning and quickly slows down at the tail. The likely reason of this is that a disproportionate number of buyers only reviewed one product when

using Amazon.

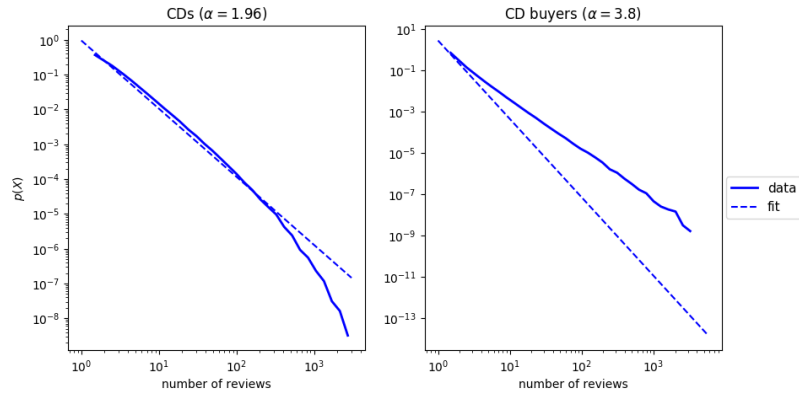
We suspected that the reason for a lack of fit on the buyer review frequency is that all buyers are fitted together regardless of the time joined. Since buyers who joined earlier are more likely to have written more reviews, a lack of fit may be caused by unequal number of users joining each year. We address this issue by fitting power law distributions to buyers who wrote their first review in the same year. This is done for the years 2001 to 2010 and the results indicate that the power law distribution is not a good fit for the frequency of review number regardless of the year and the estimated  $\alpha$  parameter is always larger than the data. On the other hand, power law distributions are decent fits for review frequencies of CDs released in any year. These results show that the distribution pattern is consistent for the entire dataset as well as data from individual years.

Our analysis showed that the probability of CDs getting more reviews decreases roughly proportional to the 2<sup>nd</sup> power of the number of reviews. This means that the probability of a CD getting one more review is always about 50% less. Since this is a critical feature of the data regardless of which year the CDs are released, it suggests that only a handful of CDs are going to get a lot of reviews.



**Figure 2.1: Frequency Distribution of Reviews**

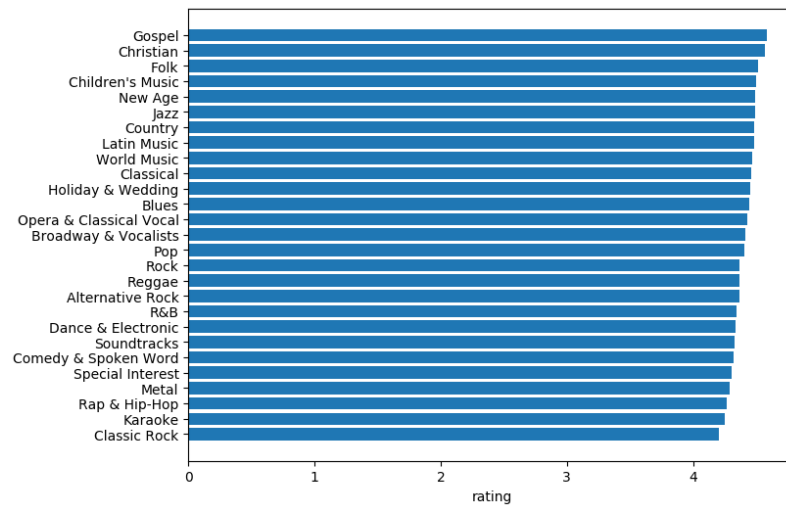
The plot on the left side shows how many reviews a CD got and the plot on the right displays the number of reviews a customer wrote. The number of reviews on the x-axis is truncated to a limited range to better show the decreasing pattern.



**Figure 2.2: Frequency Distribution with Fitted Power Law Line**  
 The number of reviews are logarithmically binned to better represent the power law line.  $\alpha$  represents the estimated  $\alpha$ /power parameter in the power law distribution.

## 2.2 Popularity of Different Genres

In this section we compare ratings and the number of reviews from different music genres. The average ratings CDs got across different genres is shown in Figure 2.3. Gospel or Christian music have the highest rating on average, while metal, rap & hip-hop, karaoke and classic Rock have lower ratings on average. In general, the ratings appears to be similar across different genres. Additionally, reviewers tend to give CDs a rating above 4, which is on the higher end of a 1-5 scale.



**Figure 2.3: Average Rating for CDs in Different Genres**

Next we looked at how many reviews CDs in each genre got, as shown in Figure 2.4. Alternative rock and metal albums receive the most reviews on average whereas classical, Latin music and karaoke got the least reviews. The number of reviews vary significantly across genres, ranging from 2.5 to 20.

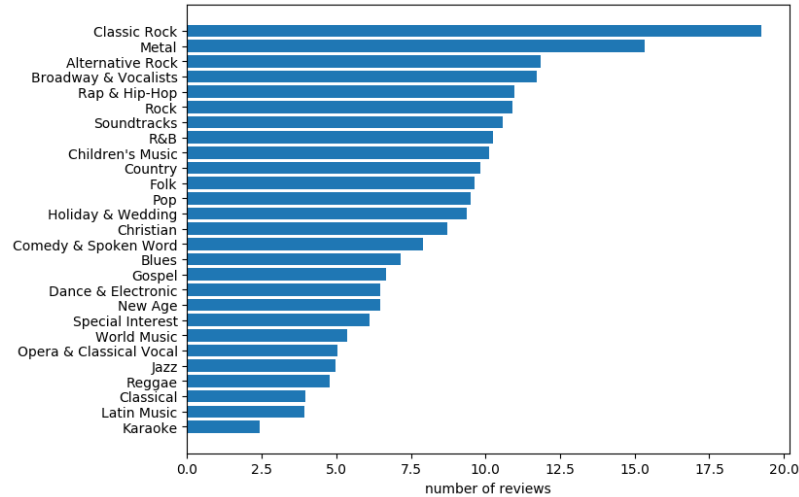


Figure 2.4: Average Number of Reviews of CDs in Different Genres

## 2.3 Purity of Dominant Genres

Some music listeners enjoy exploring other genres while others prefer to stick with their favourite. The goal in this section is to identify listeners of which music genres are more likely to review CDs from genres that they like most. This is done by calculating the “purity” of a buyer’s dominant genre. A dominant genre for a person is a genre that he/she has the most reviews on and the purity is defined as the number of reviews from the dominant genre divided by the total number of reviews that person has ever written. If a person has a high purity score, it indicates that he/she is most likely to review CDs that belongs to his/her favourite genre. Conversely, a low purity score suggests that the person is more diverse in the kinds of music he/she reviews.

Figure 2.5 shows the average purity score for each music genre. Metal, pop, rock and alternative rock are the genres with the highest purity scores. This means that

buyers who prefer these genres are more likely to review CDs of the same genre, over 90% of the time.

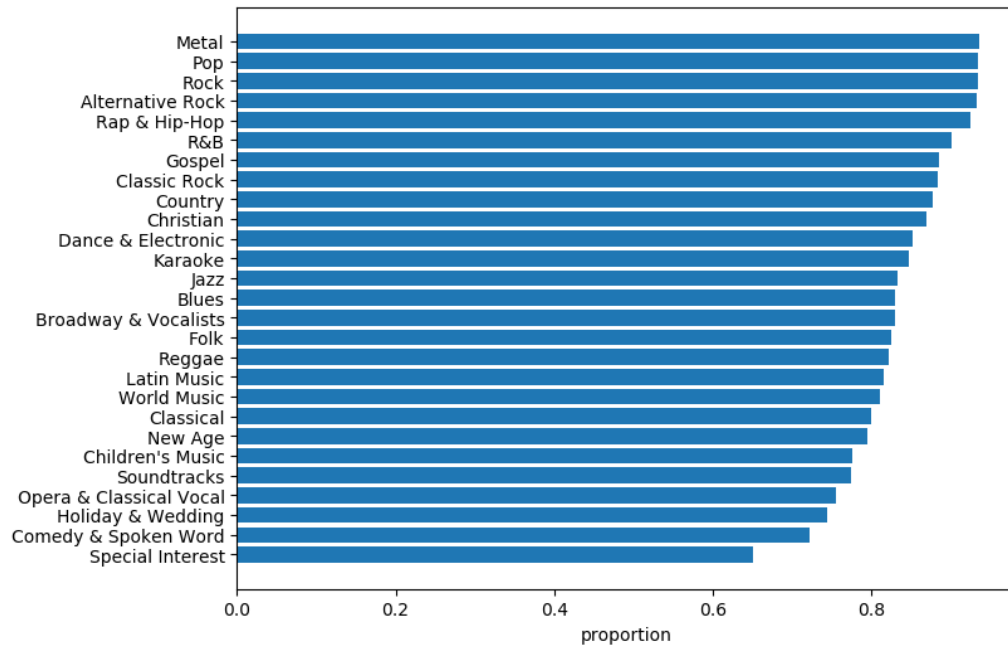
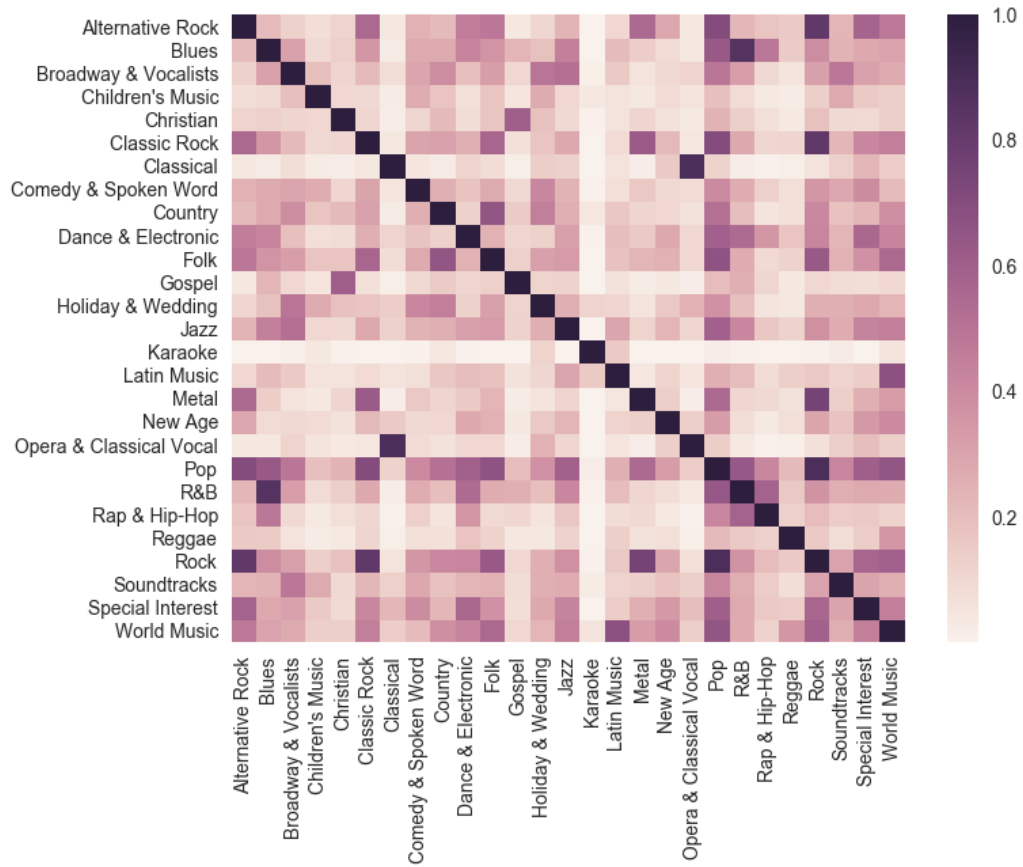


Figure 2.5: Genre Purity Score

## 2.4 Relationship Between Genres

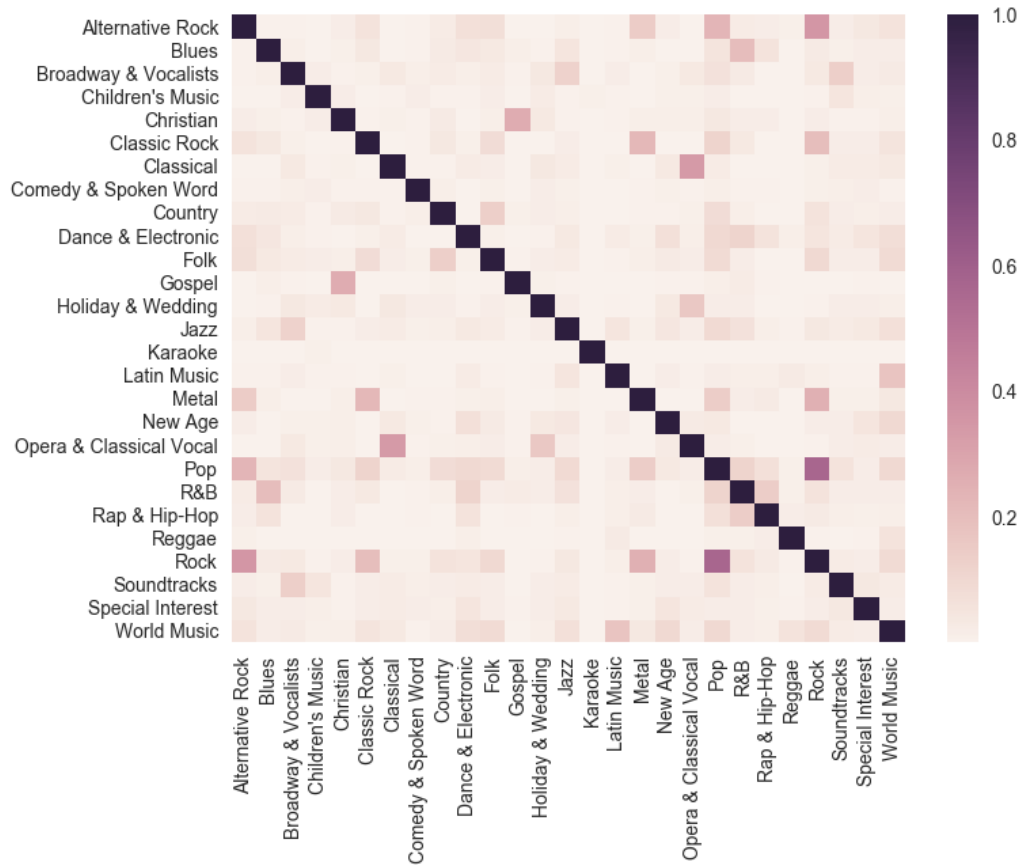
We want to know which genres are often enjoyed by the same people, as well as which genres often appear together in albums. To answer the first question, we summed up the number of albums a user reviewed for each genre and calculate the correlation coefficient between the number of reviews for each pair of genres. The results are shown in Figure 2.6. Rock, classic rock, alternative rock, metal and pop are all strongly correlated, indicating that they are often reviewed by the same people. In addition, classical and opera & classical vocal, blues and R & B as well as Latin music and world music are also highly correlated.



**Figure 2.6: Correlation Between Music Genres**

The correlation coefficient between the number of reviews each genre got with buyers being the observations. Buyers with only one review are removed since they do not provide information on what other genres they might be interested in.

Since the result above may be affected by the fact that a CD can be labelled as multiple genres, we also identified genres that often appear in the same albums. The Jaccard similarity [4] is used to evaluate how frequently each pair of genres appear together since the data is binary. The Jaccard similarity matrix is presented as a heat map in Figure 2.7. We found that the genres which often appear together are pop and rock, rock and alternative rock, classical and opera & classical vocal, gospel and Christian. These findings are consistent with the general impressions on music genres.



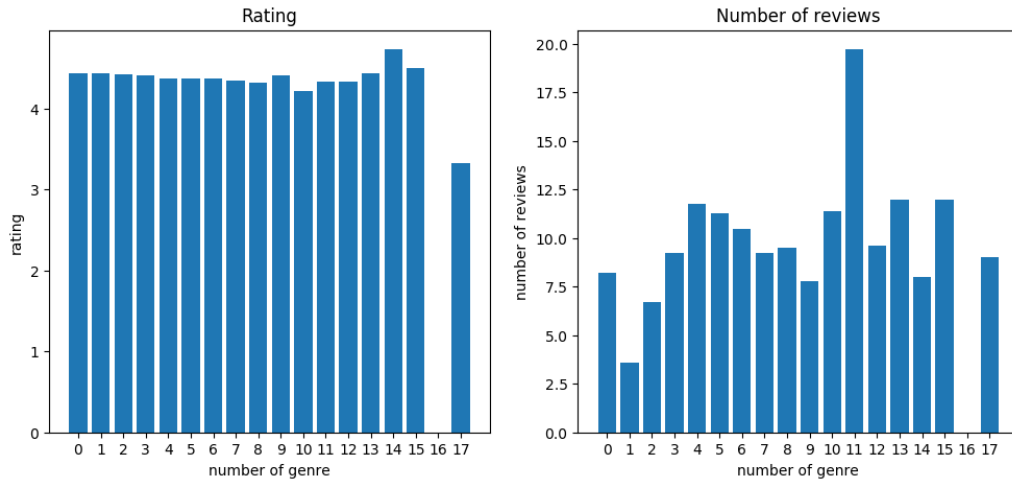
**Figure 2.7: Jaccard Similarity Between Music genres**  
 The Jaccard similarity between music genres with CD albums being the observations.

Relationships between genres can be used to design better recommendation systems. For example, a customer is likely to enjoy albums from a genre that is closely related to his/her favourite genres. As a result, even if that customer has never bought a CD from that genre, effective recommendations can be made based on the relationships between genres in our analysis.

## 2.5 Popularity of Cross-Genre CDs

As mentioned before, a CD can have multiple genres labelled to it and we would like to understand how the number of genres a CD belongs to might affect the rating and how many reviews it got. The left plot in Figure 2.8 shows the average rating of

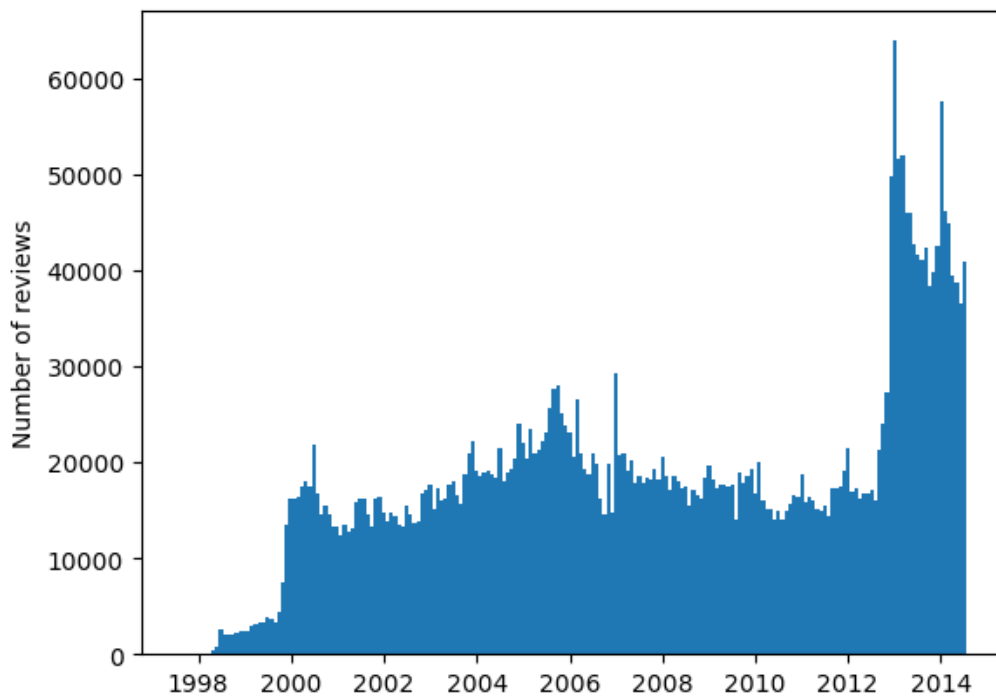
CDs with different number of genres. No clear pattern can be observed and it seems that the number of genres a CD has does not have a noticeable effect on the rating it got. Although no obvious trend exists, the number of reviews that CDs have is more varied across different number of genres. In addition, the sample size of CDs with a high number of genres are significantly smaller than other CDs. Therefore, no concrete conclusion can be made regarding the impact of the number of genres a CD has.



**Figure 2.8: Average Rating and Number of Reviews for CDs with Different Number of Genres**

## 2.6 Buyers’ Interest in Genres Across Time

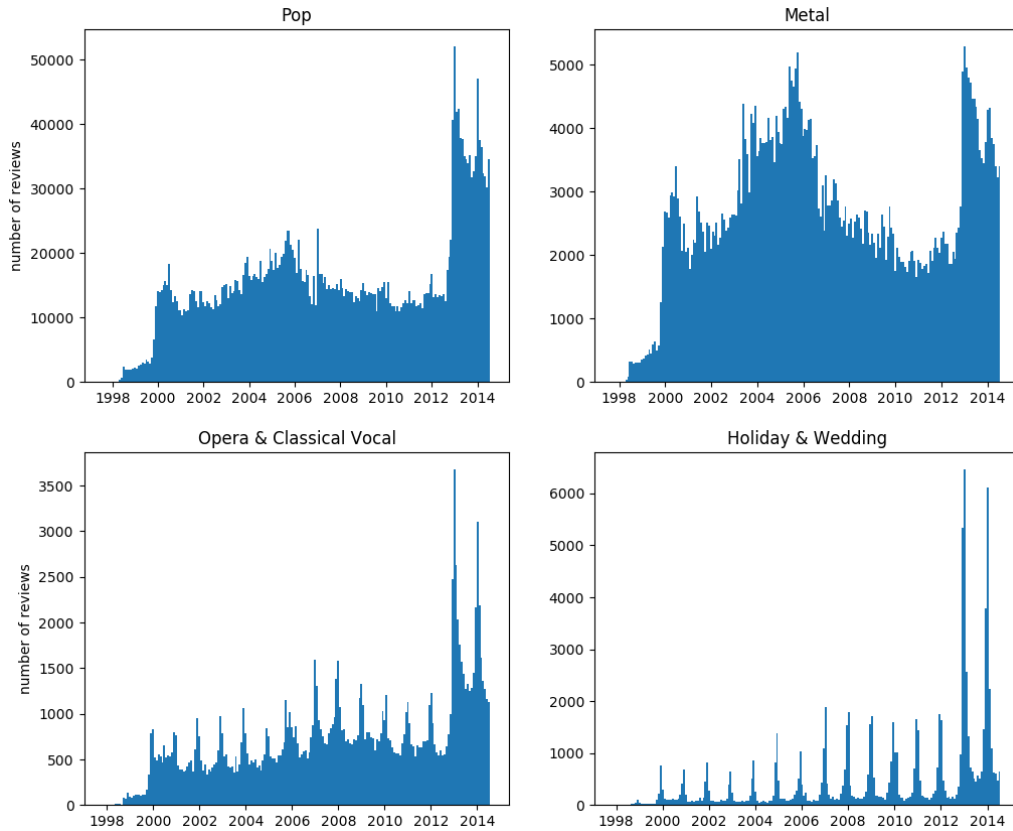
In this section and the next section, we will explore how the popularity of different music genres change over time. First we will look at popularity from the viewpoint of the customers. The number of reviews written is used as the indicator for popularity among buyers. Before looking at any specific genre, we first plotted out the number of reviews across time for the entire dataset to investigate the overall trend of the CD market. The result is shown in Figure 2.9. We see that there are roughly four “waves” of intense popularity for the CDs as a whole around 2001, 2006, 2013 and 2014 respectively.



**Figure 2.9: CD Review Number Over Time**

Each bar represents the number of reviews products in the CDs & Vinyl category got in a month.

Now we turn our attention to the number of reviews over time for some specific music genres. The top two plots in Figure 2.10 are the popularity trend for pop and metal music. One can see that the pattern for these two genres are similar to the general trend that we have seen above. Nevertheless, notice that Metal music albums have roughly equal popularity in the second and third wave. Most genres also have this same pattern but with varying strength of waves. On the other hand, some genres show a distinct cyclic pattern as shown in the bottom two plots in Figure 2.10. The cyclical behavior of opera & classical vocal might be caused by the opera house schedule and the cyclic pattern of the holiday & wedding category may be due to people buying albums from it during holiday seasons.



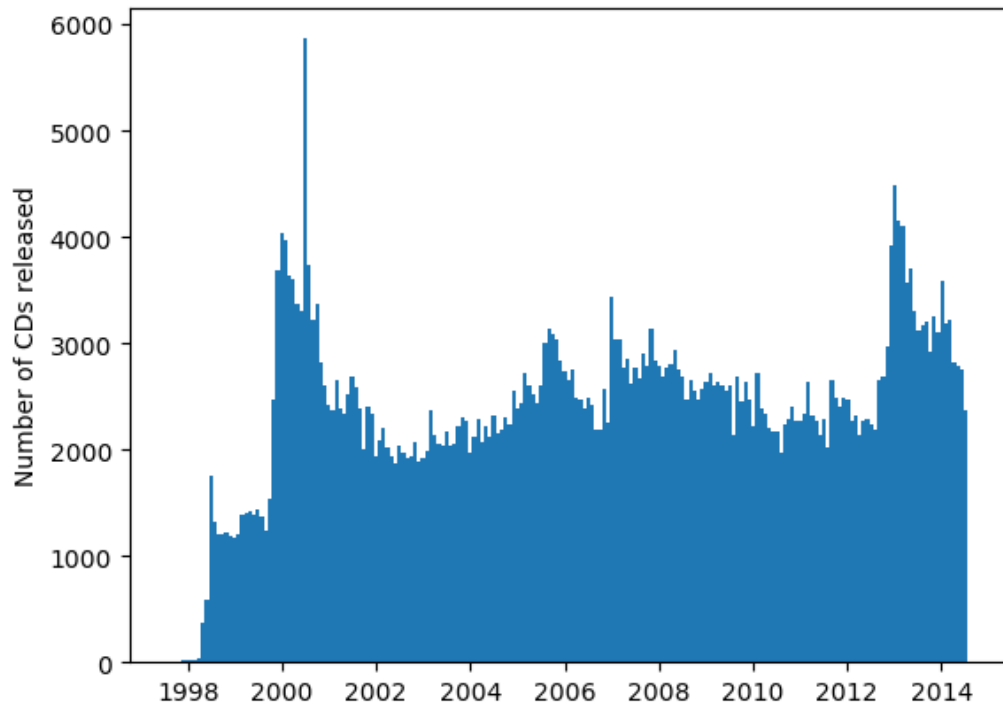
**Figure 2.10: Number of reviews per month for different genres**

From the analysis in this section, we conclude that CDs in most music genres follow the same overall popularity trend in the music industry, while CDs from some genres such as opera & classical vocal and holiday & wedding display cyclical patterns. This motivates different marketing strategies for products in these two genres.

## 2.7 Producers' Interest in Genres Across Time

We used the number of CDs released to determine the popularity of CDs from the producers' perspective. Since the release dates are not available, we used the time at which a product got its first review instead. Similar to the previous section, we first look at the general trend, presented in Figure 2.11. The popularity waves came at the same time as we have seen from the customers' perspective. However, the

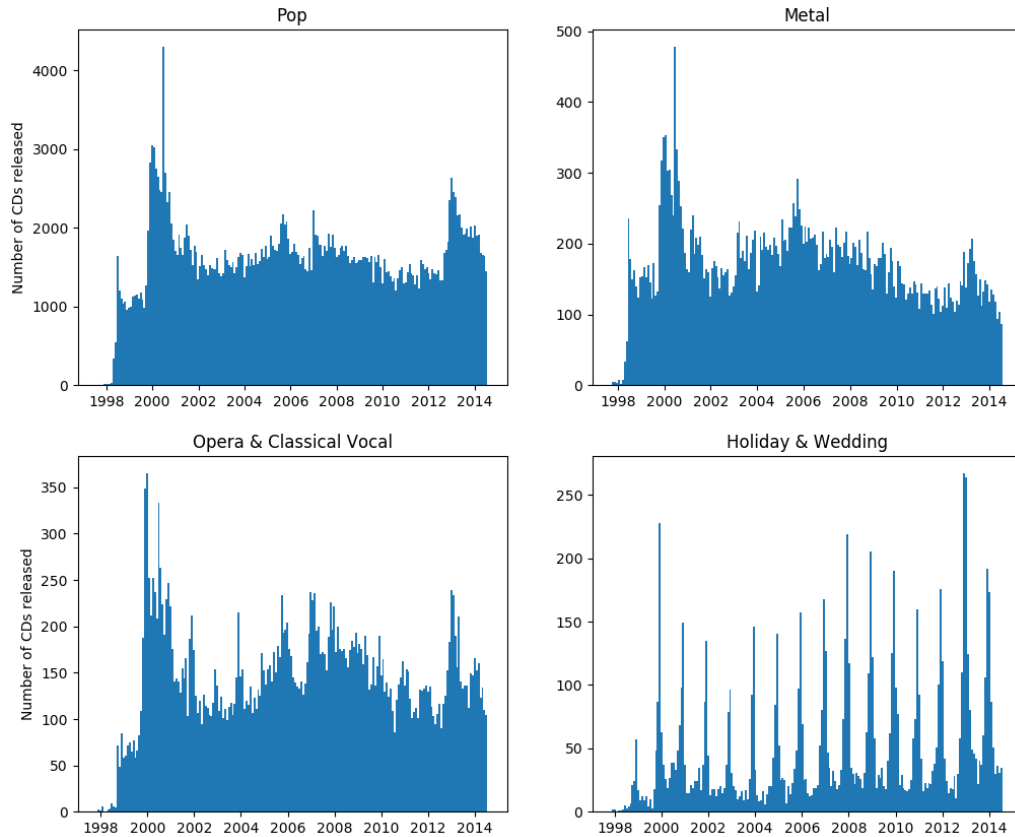
intensity of the waves are different, with the first wave having the most CDs released instead of the third.



**Figure 2.11: Number of CDs Released Over Time**

Each bar represents the number of products released in the CDs & Vinyl category within a month.

Similar to the previous section, we now turn our attention to CDs from individual genres. The number of CDs released for the same four genres in the previous section are shown in Figure 2.12. Pop and metal still have similar pattern as the overall trend, while opera & classical vocal and holiday & wedding still has cyclical behavior. Nevertheless, notice that the cyclical pattern in opera & classical vocal is less obvious and the pattern seems to be a mix of the overall trend and seasonality.



**Figure 2.12: Number of CDs Released Over Time for Different Genres**

Our analysis in this section showed that the general trend of CD released is slightly different than the CD popularity from the customers' perspective. This might suggest that companies should release more CDs to meet the interests of the market when disparity occurs, at least for genres that follows the general industry trend. On the other hand, products in the genres opera & classical vocal and holiday & wedding have their unique pattern, thus the marketing strategies for them should take the cyclical behavior into consideration.

# 3

## CD Life Cycle Analysis using Bayesian Monitoring

### 3.1 Motivation

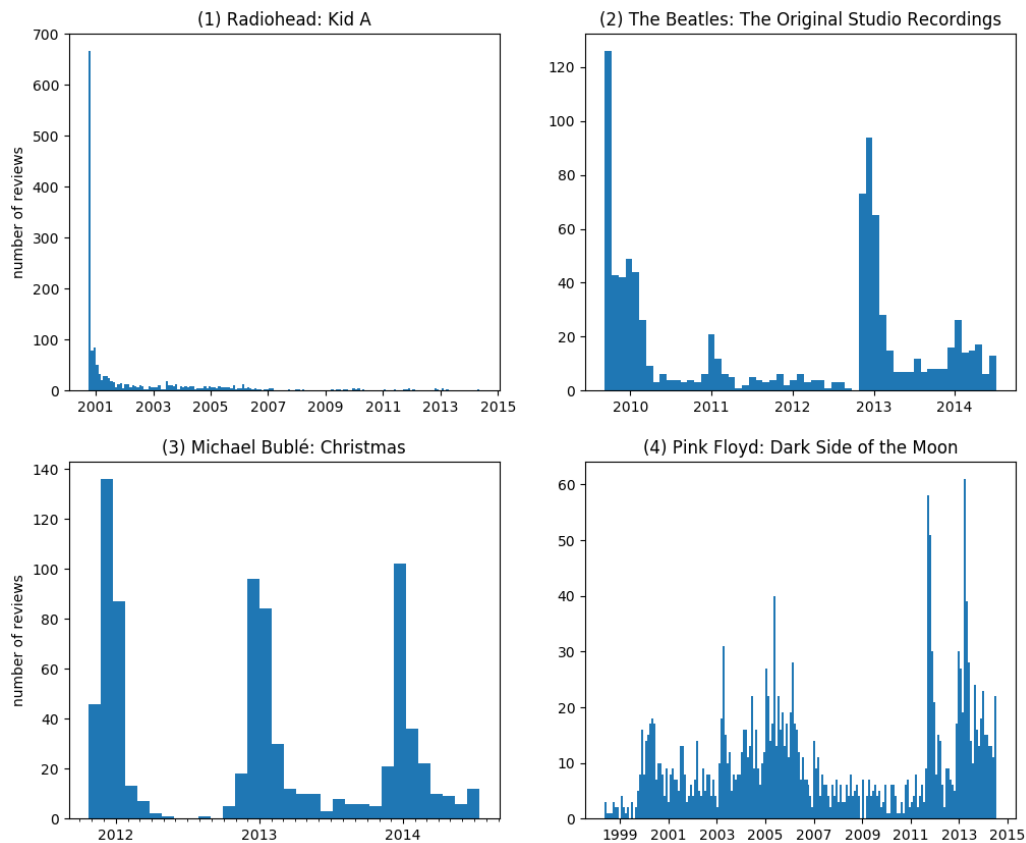
The life cycles of products have been one of the center focuses of the business world, especially in the marketing literature. Understanding how the sales/popularity of products evolve over time provides valuable information critical for formulating better business strategies. Traditionally, the general framework used in analyzing product life cycle is to divide the life cycle into four parts – Introduction, Growth, Maturity and Decline [5]. This suggests that a product grows slowly at the beginning, gradually reaching its peak, before quickly declining to the end of its life cycle.

We want to investigate whether this framework fits the CD review data as well as identify key features in life cycles of CDs. The first section of this chapter is dedicated to exploring different life cycle patterns of CD reviews. Our focus then moves toward the major pattern in the dataset, the single exponential decay life cycle. The Bayesian monitoring method is used to identify the initiation of popularity as well as quantify the magnitude of change in CDs' life cycles. We proposed an improved Bayesian monitoring method using prior predictive  $p$ -values which allows the method to deal with extreme changes. Finally, we will demonstrate how fitting exponential decay function to the single exponential decay life cycle helps differentiate CDs with a short-lived popularity versus CDs with a long-lasting interests from customers.

### 3.2 Life Cycles of CD Reviews

We discovered that there are four general patterns of life cycles for CD reviews. The majority of CDs have a life cycle with an exponential decay like shape, with a huge

spike in popularity early and the number of reviews in subsequent month quickly decreasing. An examples of this pattern is shown in plot (1) of Figure 3.1. The second type of life cycle is a "multiple-waves" pattern which contains several waves of intense popularity of the CD. Each wave is followed by a quick descent similar to the previous pattern. Additionally, some CDs exhibit periodic life cycles as shown in plot (3). Those CDs mostly features songs related to Christmas, suggesting that people tend to buy them during the holiday thus resulting in periodicity. The last life cycle pattern is exhibited by CDs released before 1998, the earliest time period from which the dataset have data. Since the data is incomplete in this case, no dominant pattern can be identified.



**Figure 3.1: Life cycle patterns of CDs**

The plots show the number of reviews a CD got per month since receiving its first review. The four distinct patterns are (1) Single exponential decay, (2) Multiple waves, (3) Periodic and (4) CDs released before 1998.

From the work above, we saw that the life cycles of CD reviews does not fit too well into the traditional life cycle analysis framework and requires a different approach for analyses. The vast majority of CD reviews display a single exponential decay pattern. This implies that customers are usually quickly interested in a CD when it is new and therefore its popularity reaches the peak almost instantly. For the next two sections of this chapter, we will focus on analyzing the single exponential decay pattern since it is the dominant life cycle of CDs. An improved Bayesian monitoring method will be used to detect the beginning of the popularity spike and the rate of popularity decline will be estimated by fitting an exponential decay line to the life cycles.

## 3.3 Bayesian Monitoring

### 3.3.1 Introduction

The Bayesian monitoring technique proposed by Mike West (1986) [2] revolves around the idea of constructing a standard model using Bayesian updating and checking whether it fails using an alternative model as new data come in. The degree of likelihood of the standard versus the alternative for new data can be evaluated using the Bayes factor of their respective prior predictive distributions, indicated as  $H_t$  in the paper. The standard model will be deemed unsatisfactory when the Bayes factor is lower than a pre-specified threshold and subsequently the alternative model will become the standard model from the next time point. Since changes can occur gradually over time, the Bayesian monitoring uses the cumulative Bayes factor of the most discrepant group of recent, consecutive observations to assess cumulative effects. This is defined as

$$V_t = H_t \min(1, V_{t-1})$$

The standard model will also be updated to the alternative model if  $V_t$  is lower than the threshold.

In general, the alternative model is a more diffused version of the standard model, having the same location parameter but a with higher variance. The paper suggested constructing the alternative using power discounting on the prior distribution for the standard model. As a result, the alternative will have prior distribution with the form

$$p_A(\eta_t|D_{t-1}) = p_S(\eta_t|D_{t-1})^\delta$$

where  $\eta$  is the natural parameter,  $D_{t-1}$  is all the data from time point 1 to  $t - 1$  and  $\delta$  is a discounting factor between  $[0, 1]$ .

The main strength of the Bayesian monitoring method is that it is not overly sensitive to outliers while still having the ability to adapt to changes in new data. In addition, it provides posterior distributions that can be used for prediction and constructing credible intervals.

### 3.3.2 Extending Bayesian Monitoring to Count Data

To apply the Bayesian monitoring method to our Amazon CD data, we must account for the fact that the data consists of non-negative integers. Each time point in our data represents how many reviews a CD got within a day. As a result, the Poisson distribution appears to be the appropriate choice for modelling this data. For convenience when updating, we used the Gamma distribution as the prior distribution for  $\lambda$  since it is the conjugate prior for Poisson. The resulting prior predictive distribution is then a negative binomial distribution, which is used to compute the Bayes factor. Distributions in our model are shown in Table 3.1.

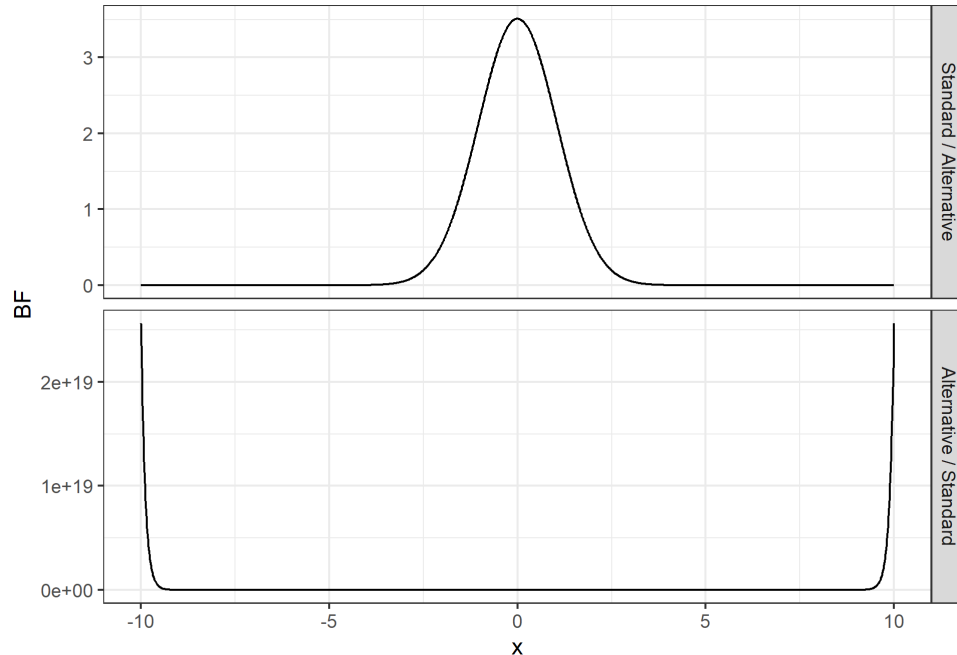
**Table 3.1: Distributions in Bayesian Monitoring for Positive Count Data**  
 $\lambda$  represents the parameter of the Poisson distribution.  $\alpha$  and  $\beta$  represents the shape and rate parameters of the Gamma distribution.  $\delta$  is the discount factor used to construct a more diffused alternative.  $y_t$  is the data observed at time  $t$ .

	Standard	Alternative
Sampling Distribution		Poisson( $\lambda$ )
Prior	Gamma( $\alpha, \beta$ )	Gamma( $\delta\alpha, \delta\beta$ )
Prior Predictive	NB( $\alpha, \frac{1}{\beta+1}$ )	NB( $\delta\alpha, \frac{1}{\delta\beta+1}$ )
Posterior	Gamma( $\alpha + y_t, \beta + 1$ )	Gamma( $\delta(\alpha + y_t), \delta(\beta + 1)$ )

### 3.3.3 Dealing with Extreme Model Failure

The Bayesian monitoring method uses the Bayes factor of the standard model against the alternative model in determining whether the standard model is satisfactory or not. Whenever the Bayes factor ( $H_t$ ) or cumulative Bayes factor ( $V_t$ ) is lower than some pre-specified threshold, the standard model is rejected and we use the alternative model instead in subsequent analysis.

However, the Bayes factor may exaggerate the evidence favoring the alternative when new data is unlikely to occur under both the standard and alternative models. This can be illustrated by plotting the Bayes factor of the alternative against the standard, as shown in Figure 3.2 below. We see that the Bayes factor of the alternative against the standard increases exponentially as the new data deviates further from our current model. This implies that the Bayes factor of the standard against the alternative decreases exponentially and this is due to the density of the exponential family distributions decreasing exponentially in the tails.



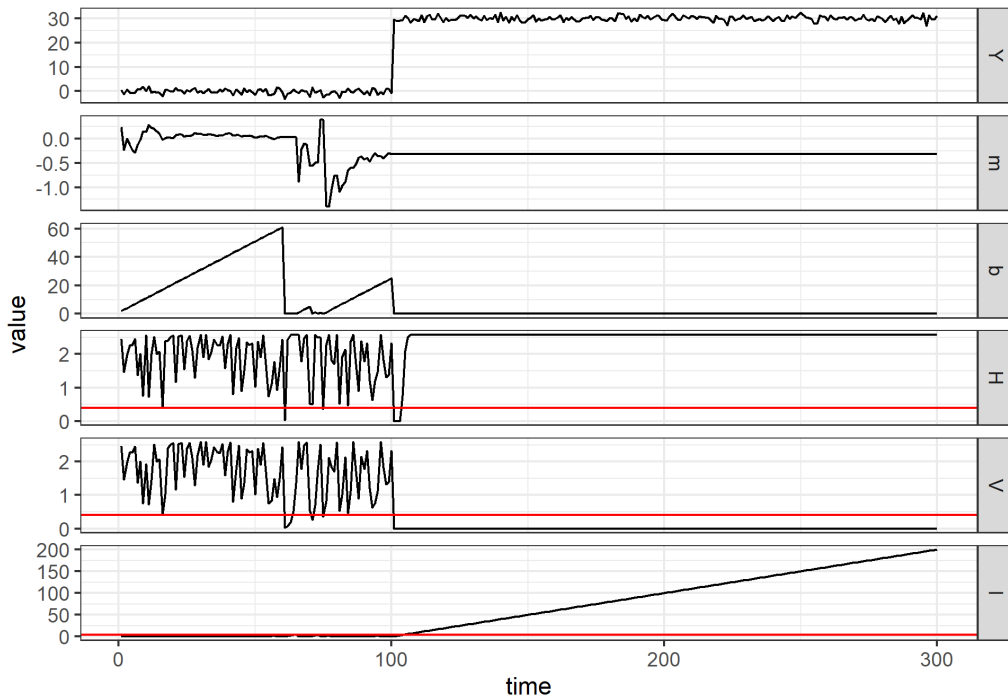
**Figure 3.2: Bayes factors of Standard vs. Alternative Model**

To illustrate the issue in a simple example, the standard model used in this plot is  $N(0, 1)$  and the alternative is  $N(0, 12)$ . The alternative is constructed by setting the variance of the sampling model ( $\phi$ ) as 1 and setting prior sample size ( $b$ ) as 10. This results in a discount factor ( $\delta$ ) around 0.08 and leads to the alternative having a variance around 12.

When this happens, the cumulative Bayes factor of the most discrepant group of recent observations ( $V$ ) becomes extremely low and could stay below the threshold ( $\tau$ ) for a very long time, preventing the model from updating to any new information. As a result, a single extreme outlier or an extreme parameter change has the potential of crippling the monitoring process and might require external intervention to resolve the issue.

In fact, this is precisely what happens when we apply the Bayesian monitoring method directly to our Amazon CD dataset. Since the number of reviews a CD got often increases drastically at some time point, the cumulative Bayes factor becomes too low to allow for any future updates. As we have seen in the last section, popularity spikes are an essential feature of CD life cycles, thus we will need to address this issue before we can apply Bayesian monitoring to detect the beginning of popularity spikes.

We will further illustrate how this problem may occur in action using a toy example of 100 data points generated from  $N(0, 1)$ , followed by 200 data points from  $N(100, 1)$ . Obviously, there is a parameter change at time 101, changing the mean parameter from 0 to 100 and we certainly expect the monitoring process to pick that up. The result of the analysis is shown in Figure 3.3. As mentioned previously, we see that  $V$  stays extremely low after the parameter change and the model stops updating the parameter  $m$  because  $V$  stays below the threshold  $\tau$  for all time points afterwards. This shows that the model is unable to adapt to new data (at least for a very long time) when one Bayes factor (at time 101) is extremely low.



**Figure 3.3: Model Failure with Extreme Parameter Change**

The sampling model is a Normal distribution with variance ( $\phi$ ) 1. The monitoring process is conducted using  $\rho = 0.15, \tau = 0.4$  as suggested in M. West (1996) and using prior parameters  $m = 0, b = 1$ . In the plot,  $Y$  is the data,  $m$  and  $b$  are the estimated (posterior) mean and precision parameters for  $\mu$ ,  $H$  and  $V$  are the Bayes factor and the cumulative Bayes factor, and  $l$  is the run length parameter. The red line indicates the threshold  $\tau$  and the threshold for  $l$ .

The cause of this issue is that the Bayes factor inappropriately favors the alter-

native when both the standard and alternative models fail to capture new data. Due to how the alternative is formulated, this almost always happens when new data lies in a small tail area of the alternative model.

To solve this issue, we need a value that measures how extreme new data is compared to the current models. An intuitive measure for this purpose is the proportion of the tail area (tail-area probability) more extreme than the new data under the prior predictive distribution. The idea is presented in detail by Andrew Gelman, Xiao-Li Meng and Hal Stern (1996) [6]. We will utilize a similar measure as the posterior predictive assessment mention in their paper but we will use the prior predictive distribution instead of the posterior predictive distribution since we want to assess the fitness of new data at time  $t$  compared to our current model (prior at time  $t$ ). This measure is sometimes referred to as the “prior predictive  $p$ -value” as it has a similar interpretation as the  $p$ -value often used in frequentist statistical analyses.

The tail-area probability for the alternative model is defined as

$$\min(\Pr_A(Y \geq y_t), \Pr_A(Y \leq y_t))$$

where  $y_t$  is the new data and  $p_A(Y_t|D_{t-1})$  is the prior predictive distribution under the alternative at time  $t$ . For a Normal sampling model, the prior predictive distribution is symmetric and we can also use the two-tailed tail-area probability, defined as

$$2 \times \min(\Pr_A(Y \geq y_t), \Pr_A(Y \leq y_t))$$

Note that the tail-area probability of the standard model is always smaller than that of the alternative. Consequently, the prior predictive tail-area probability is low when the new data is unlikely under both the standard and alternative models, suggesting model failure of both models.

We propose a power discounting method that decreases the impact of extreme outliers or changes. As mentioned previously, the model stops updating parameters when  $V$  stays low due to one extremely low Bayes factor. Therefore, the problem can be resolved if we lower the effect of Bayes factors of observations with a small

tail-area probability. Since the tail-area probability is between  $[0, 1]$ , taking the Bayes factor to the power of the tail-area probability always pulls it towards 1, effectively decreasing its effect on the cumulative Bayes factor  $V$ .

We define the “adjusted  $V$ ” as follows

$$H_t^{adj} = (H_t)^{p_t}$$

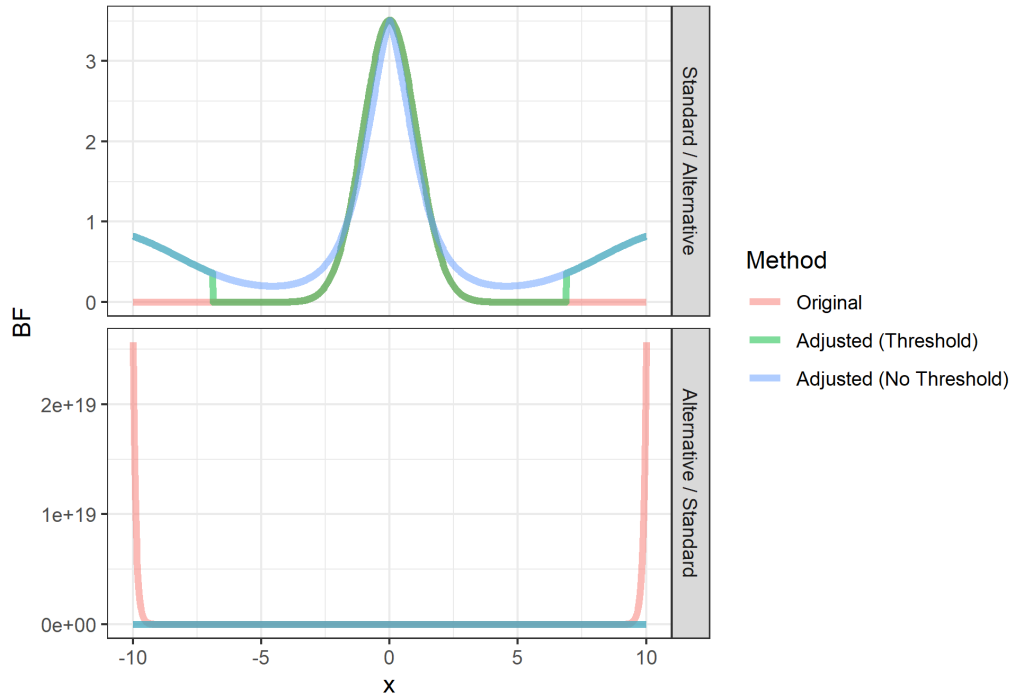
$$V_t^{adj} = H_t^{adj} \min(1, V_{t-1}^{adj})$$

where  $p_t$  is the tail-area probability.

This adjustment is only made after we checked for standard model failure using the original, unadjusted cumulative Bayes factor  $V_t$ . Since a low tail-area probability indicates model failure for both the standard and alternative, we still want to lower the estimate of the precision and thus should only adjust  $V$  after the usual adapting process. This way, this adjustment does not interfere with the original logic of Bayesian monitoring.

There are two ways to apply the tail-area probability adjustment: the threshold and non-threshold method. The threshold method only applies the discount when the tail-area probability is lower than some pre-specified threshold, similar to how model failure is decided by comparing the Bayes factor to a threshold. On the other hand, the non-threshold method adjusts  $V$  regardless of its value. The effects of both method to the Bayes factor is shown in Figure 3.4.

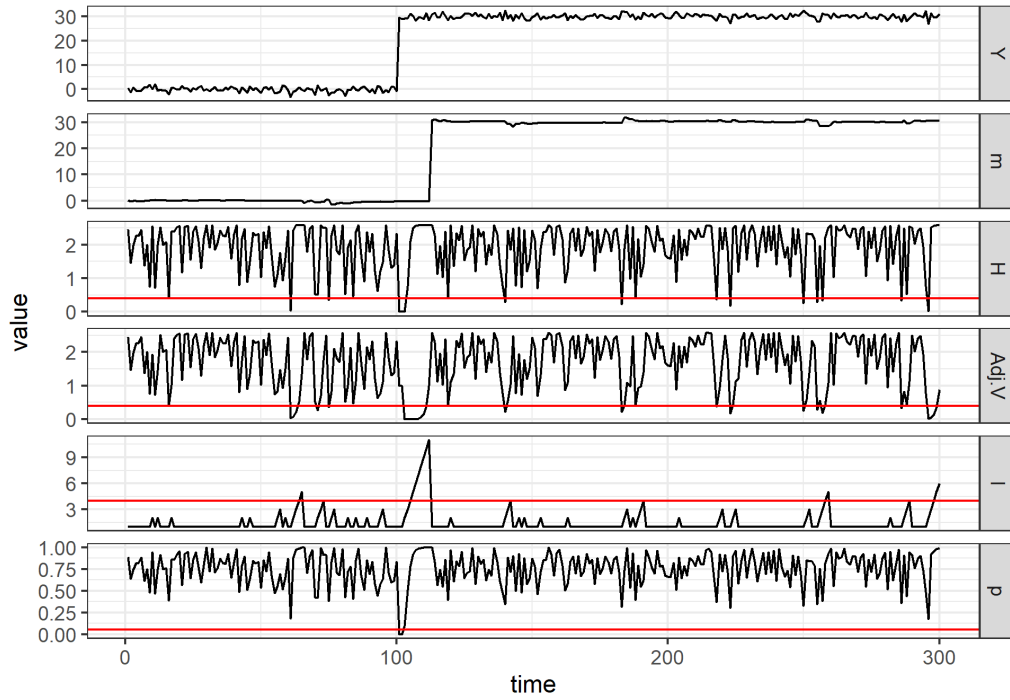
If the threshold method is used, the monitoring process remains the same when new data is not extreme enough under both the standard and alternative models. We believe this is desirable since this means it preserves all the merits of the original method while only guarding against extreme changes or outliers.



**Figure 3.4: Adjusted Bayes factors of Standard vs. Alternative Model**  
 The Bayes factors are calculated using the same set-up as in Figure 3.2.

Figure 3.5 shows the analysis result after we apply the discounting method on  $V$ . We see that the model can adapt to new data after around 10-15 observations following the parameter change. The Bayes factor ( $H$ ) stayed low for several time points but the adjusted cumulative Bayes factor (Adjusted  $V$ ) does not remain low and thus the model is able to update its parameters after a reasonable amount of new data has been collected.

The tail-area probability adjustment provides an easy fix to the problem of dealing with extreme outliers or parameter changes in the original Bayesian monitoring technique. It does not change the original analysis process and only protects the model against data with low tail-area probability. The tail-area probability also serves as an additional check for model failure. We believe the method will be useful for applications with potential large changes and would be critical for our analysis in the next section.



**Figure 3.5: Adjusted Bayesian Monitoring for Extreme Parameter Change**  
 $p$  shows the tail-area probability for each observation. The threshold method is used here with a threshold of 0.05 (the red line in  $p$ ), thus  $V$  is only adjusted when the tail-area probability is lower than 0.05.

### 3.3.4 Applying Bayesian Monitoring to CD Review Data

With the adjustments in the two previous sections, the improved Bayesian monitoring method has the following algorithm.

---

**Algorithm 1:** Adjusted Bayesian Monitoring Method for Count Data

---

**Inputs:**

$y_1, \dots, y_n$

**Initialize:**

$\alpha_0 = 1$  (or value of user's choice)

$\beta_0 = 1$  (or value of user's choice)

**for**  $t = 1$  *to*  $n$  **do**

$\delta_t = \rho/[1 + \beta_{t-1}(1 - \rho)]$

$H_t = \text{NB}(y_t, \alpha_{t-1}, \frac{1}{\beta_{t-1}+1})/\text{NB}(y_t, \delta_t\alpha_{t-1}, \frac{1}{\delta_t\beta_{t-1}+1})$

$V_t = H_t \min(1, V_{t-1})$

**if**  $V_{t-1} > 1$  **then**

$l_t = 1$

**else**

$l_t = l_{t-1} + 1$

**end**

$p_t = 2 \min[\Pr(Y > y_t), \Pr(Y \leq y_t)]$  where  $Y \sim \text{NB}(\delta_t\alpha_t, \frac{1}{\delta_t\beta_t+1})$

**if**  $V_t > \tau$  **or**  $l_t > k$  **then**

$\alpha_t = \delta_t\alpha_{t-1}$

$\beta_t = \delta_t\beta_{t-1}$

**else**

$\alpha_t = \alpha_{t-1} + y_t$

$\beta_t = \beta_{t-1} + 1$

**end**

**if**  $p_t < p_{\text{threshold}}$  **then**

$V_t = H_t^{p_t} \min(1, V_{t-1})$

**end**

**end**

---

The fix parameter used in this analysis are  $\rho = 0.15$ ,  $k = 4$ ,  $\tau = 0.1$  and  $p_{\text{threshold}} = 0.05$ . All notations are consistent with Mike West (1986) [2].

We applied the adjusted Bayesian monitoring method to the top 500 most reviewed CDs with a single exponential decay life cycle. 88.8% (444 out of 500) of CDs have their major popularity spike at the beginning (day 1) with the other CDs having the major review number spike occurring from day 2 to an extreme outlier of day 4722 (*The Best of Simon & Garfunkel*). Recall that since we are only observing

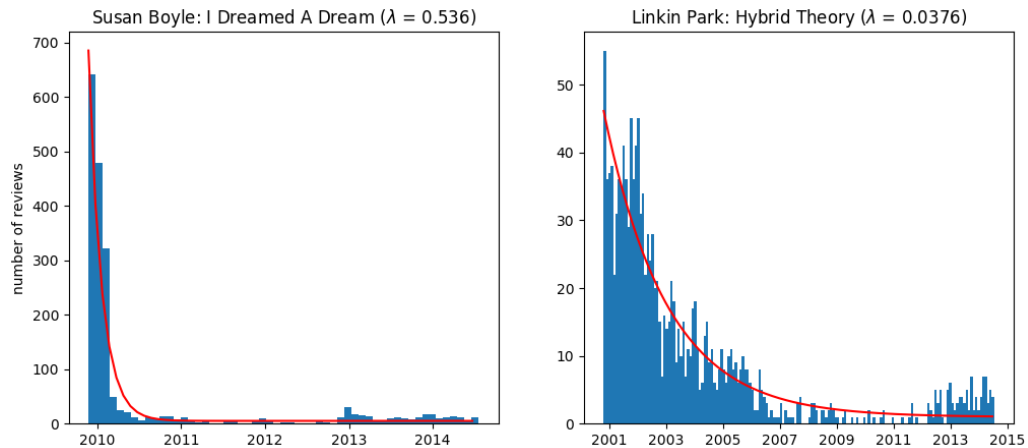
the review data, day 1 is not the release day but the day at which the CD received its first review. Nevertheless, our analysis showed that CDs usually attract customer interests quickly and often enter the maturity stage of their life cycles immediately. Our result suggests that companies should plan their business strategies for CD products before the release since the popularity of CDs is almost always at its highest immediately after release.

### 3.4 Fitting Exponential Decay Function

After identifying the start of of a CD's popularity, one may also want to estimate how fast its popularity decreases across time. We fitted exponential decay to CDs with the first life cycle pattern identified above. The function we fitted has the form

$$N(t) = N_0e^{-\lambda t} + C$$

where  $t$  represents time,  $N_0$  is the initial value,  $\lambda$  is the decay rate and  $C$  is a intercept constant. The parameter we are most interested in is the decay rate, which tells us how fast the popularity of CDs decrease over time. Two examples are shown in Figure 3.6.



**Figure 3.6: Fitted Exponential Decay Lines to CD Life Cycles**

The estimated decay rate  $\lambda$  tells us how quickly a CD's popularity decreases over

time. Using the CDs in Figure 3.6 as examples, we see a distinct difference in their decay rates. The *I Dreamed a Dream* album by Susan Boyle enjoyed only a short-lived interest from customer, as indicated by a large  $\lambda$ . On the other hand, *Hybrid Theory* by Linkin Park has a much smaller  $\lambda$  value and a much smoother decay in popularity.

Identifying the rate of decay for CDs and other products helps companies decide what products to focus its production and advertisements on. Differences in the speed of decay suggests the need for different business strategies. By utilizing the methods outlined in this chapter, companies will be able to distinguish between products with different life cycles and adapt their business strategies accordingly.

## 4

# Predicting the Number of Reviews

## 4.1 Motivation

Since the number of reviews CD got is a direct indicator of popularity, accurate prediction of how many reviews CDs got will be invaluable for informing business strategies. With the ability to predict the popularity of a CD, the marketing team can target products that has the most potential to be the next big hit. In addition, the producers can use this information to design albums that has a better chance to be popular.

We built various different predictive models for predicting the number of reviews and evaluate their performances as well as analyze which variables have the highest impact. In the first section, we considered the classical linear regression and the Bayesian linear regression with model selection. The section that follows will be dedicated to other popular machine learning models including k-nearest neighbors, decision tree, random forest, boosting trees and neural network.

The predictors used in our analysis are the music genres of CDs, rating, time elapsed after released, how many reviews related CDs got on average, the mean rating of related CDs and the number of related CDs. The only CD with more than 3,000 reviews is removed to prevent the models from focusing too much on this one observation. We split the dataset into a training set with 80% of all the samples and a test set of 20% of samples. This results in a training set of around 250,000 CDs and a test set of approximately 60,000 samples.

## 4.2 Linear Regression Models

### 4.2.1 Classical Linear Regression

Linear regression models the response variable as a linear combination of the predictors. This is represented by the formula  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ , where  $\mathbf{y}$  is the vector of response,  $\mathbf{X}$  is the matrix of predictors,  $\boldsymbol{\beta}$  is the vector of estimated regression parameters, and  $\boldsymbol{\epsilon}$  is the vector of random errors. Using the sum of squared errors as the loss function, the optimal parameter vector is given by the formula  $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$ .

Since linear regression assumes linearity and normality, we first investigated what transformations are appropriate for our data. The Box-Cox power transformation for multinormality described in Box and Cox (1964) [7] is applied to the response and all predictors except the music genres. The results are shown in Table 4.1. The suggested power transform for the number of related CDs and the mean number of reviews related CDs got are the log transformation. The recommended power transformation for the number of reviews is also close to the log transformation, thus we applied the log transformation for better interpretability. Next, we chose to use the square root transformation to time elapsed since the estimated power 0.66 is close to 0.5 which corresponded to the square root. Lastly, no transformations were applied to rating and mean rating of related CDs since it will become difficult to interpret their coefficients if the power transforms suggested are used.

**Table 4.1: Estimated Box-Cox Power Transformation**

Related Mean NumReview is the mean number of reviews related CDs got and Related Mean Rating is the mean rating of related CDs.

	Estimated Power	Lower Bound	Upper Bound
Number of Reviews	-0.26	-0.2618	-0.2567
Rating	4.74	4.7101	4.7614
Number of Related CDs	0.03	0.0278	0.0329
Related Mean NumReview	-0.06	-0.0601	-0.0555
Related Mean Rating	6.03	5.9914	6.0692
Time Elapsed	0.66	0.6588	0.6668

We started with a simple linear regression model with all the predictors and no interactions. The mean squared error of the test set for this model is 1173.7 and the 10 fold cross validation MSE of the training set is 1134.4. To improve this model, we added all possible two-way interactions and fitted another linear regression model. The model with all two-way interactions has a test set MSE of 952.6 and the cross validation MSE is 918.7. Clearly, including the interactions improved the predictive accuracy of the model significantly. We did not add any higher order interactions due to the computation time required and including more complex interactions also makes the model harder to interpret. In the next section, we will investigate how the Bayesian linear regression can help with selecting the important predictors.

### 4.2.2 Bayesian Linear Regression

Bayesian linear regression [8] is the classic linear regression model done in the Bayesian inference context. The structure of the model is exactly the same, but now prior dis-

tributions are assigned to all the parameters that needs to be estimated. Bayesian inference utilizes the Bayes theorem to update our model parameters from the prior distribution to the posterior distribution by learning from the data we collect, represented using the likelihood/sampling distribution. The Bayes theorem tells us that the posterior is proportional to the prior multiplied by the likelihood and the posterior contains all the information we need regarding the parameter.

In Bayesian linear regression, a multivariate Normal distribution is used as the prior for the mean of regression coefficients  $\beta$ , an inverse Wishart distribution is assigned as the prior for the covariance matrix of the regression coefficients, and an inverse gamma distribution is used as the prior for the variance of random errors. These prior distribution are chosen due to their conjugacy to the likelihood/sampling distribution of the parameters and makes the posterior updating calculations much easier.

As mentioned in the previous section, we ran into the problem of having too many possible interactions and thus variable selection is needed in linear regression. The common Bayesian approach to this problem is to quantify how good each model is through the posterior model probability, and choosing the predictors that appears in models with high posterior probabilities. This method is called “Bayesian model averaging” and it provides a general framework to selecting models in Bayesian statistics, see Clyde (1999)[9] for more details.

Suppose we are considering  $k$  models, indexed by  $m = 1, \dots, k$ , the posterior model probability of the  $m^{\text{th}}$  model is defined through Bayes theorem as

$$Pr(\mathcal{M}_m | Y) = \frac{Pr(\mathcal{M}_m)p(Y | \mathcal{M}_m)}{\sum_{m=1}^k Pr(\mathcal{M}_m)p(Y | \mathcal{M}_m)}$$

where  $\mathcal{M}_m$  represents the  $m^{\text{th}}$  model and  $Y$  is the data. This provides a quantitative measure of how likely each model is after accounting for the data. The “Bayesian Model Averaging” (BMA) model is the model with regression coefficients weighted by

the posterior model probabilities, with coefficients defined as

$$\beta_{\text{BMA}} = \sum_{m=1}^k \beta_m Pr(\mathcal{M}_m | Y)$$

with  $\beta_m$  being the coefficients in model  $m$ . This approach solves the problem of model selection by averaging over all the models using their respective posterior model probability as weights.

After calculating the posterior model probabilities, we would also want to know which variables are often included in models with high posterior probabilities. This is often referred to as the posterior inclusion probability and is defined as

$$\sum_{m=1}^k \mathbf{1}(\mathcal{M}_m \text{ includes predictor } i) Pr(\mathcal{M}_m | Y)$$

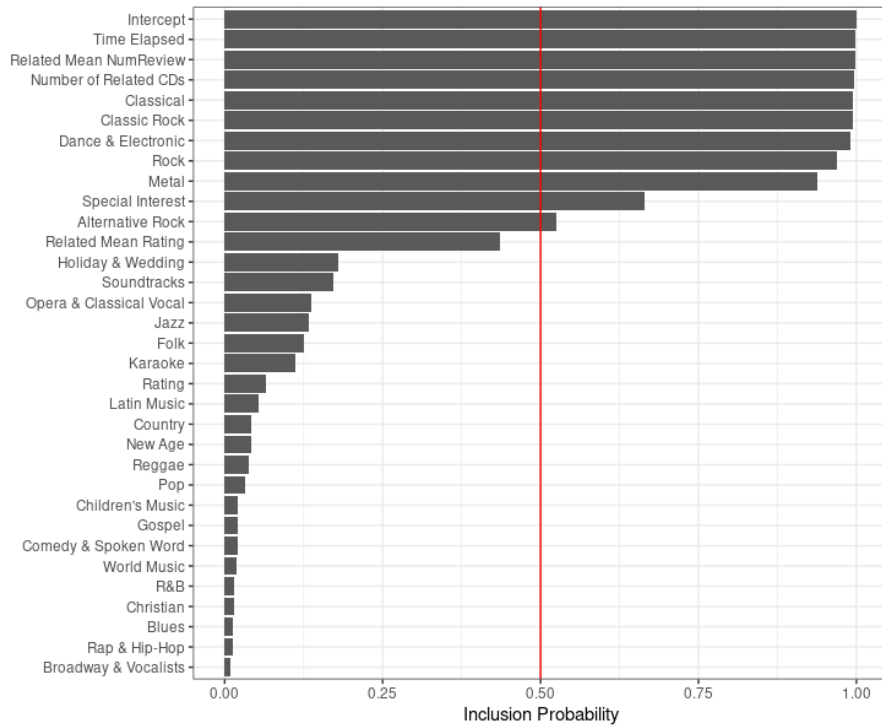
For a predictor, this is calculated by simply finding all the models that predictor  $i$  is included and summing up their posterior model probabilities. The posterior inclusion probability provides us a measurement of how often predictors provides important information to the all the models and is thus very useful in variable selection.

Since the total number of possible models is too big ( $2^p = 2^{529}$  in our case) and enumerating through all of them would be impossible in a reasonable amount of time, we need a clever strategy to explore models with high posterior probabilities without going through the entire model space. We used the BAS package written by Clyde et al. [10], which implements an adaptive sampling technique to sample from the posterior distributions of parameters that is effective in finding models with high posterior probabilities.

We fitted a Bayesian linear regression model with the same set-up as in section 4.2.1. We used 500,000 MCMC samples with a Jeffreys-Zellner-Siow prior [11] with  $\alpha = 1$  for regression coefficients. The BMA model has a test set MSE of 956.4 and the cross validation MSE is 919.8 on average. This performance is almost the same as the classic linear regression model with all the two-way interactions.

The posterior inclusion probabilities of the main predictors (no interactions) are shown in Figure 4.1. The intercept, time elapsed, mean number of review of related

CDs, the number of related CDs and the music genres classical, classic rock, dance & electronic, rock, metal, special interests and alternative rock have posterior inclusion probabilities over 50%. In addition, 114 two-way interactions have posterior inclusion probabilities over 50%. Our goal is to select important predictors and we do so by using the model that includes variables with posterior inclusion probabilities over 50%. This is referred to as the “Median Probability Model” (MPM) and the test set MSE for it is 954.9. It has a cross validation MSE of 919.8, which is also very similar to that of the classic linear regression model. Nevertheless, the MPM model uses only 125 predictors compared to 529 predictors used by the classic linear regression model. As a result, the MPM model in Bayesian linear regression is preferable over all the other linear regression models that we have investigated, since it has one of the best performance with the least number of parameters.



**Figure 4.1: Posterior Inclusion Probabilities of Predictors**

The red line indicates a posterior inclusion probability over 50%. The inclusion probabilities of the interactions are not plotted due to the size of the plot.

## 4.3 Other Machine Learning Models

Besides linear regression, we wanted to investigate the performance of other popular machine learning models on our problem and see if they can predict how many reviews CDs have more accurately. The metric used to evaluate performance will also be the mean squared error of all the predictions. Models investigated include k-nearest neighbors, decision tree, random forest, boosting trees and neural network. None of these models assumes linearity or normality, thus no transformations are applied to the variables. All the models mentioned in this section except neural network were implemented using the Scikit-learn package [12] in Python and the neural network model is implemented through the Python package Keras [13].

Since some of these models do not have an internal measure of variable importance, we evaluated the contribution of each variable using the permutation importance [14]. The permutation importance measures variable importance by calculating the impact on the loss function when permuting a predictor. Intuitively, permuting a predictor breaks its relationship with the response variable and gives us an estimation of how the model would have performed if that predictor's impact is removed. To calculate the permutation importance of a predictor, we permute it for a number of times and compute the average change in the loss of the model between the permuted and unpermuted dataset. In our analyses, the importance of a predictor is thus defined as the increase in mean squared error when permuting that predictor.

### 4.3.1 K-Nearest Neighbors (KNN)

The k-nearest neighbors model finds the k most similar training samples to the target and its prediction is the average/weighted average response of those samples. To fit a KNN model, one must first define a way to measure how similar/dissimilar each samples are to each other. The most common choice is the Euclidean distance, which is defined as  $\sqrt{\sum_{i=1}^p (x_i - y_i)^2}$  with  $x_i, y_i$  representing the predictors of two samples and  $p$  is the number of predictors. The Euclidean distance measures the dis-

tance/dissimilarity between two samples and the “k-nearest neighbors” for a sample would be the k training samples with the smallest Euclidean distance to it. After finding the closest samples, the prediction would simply be the average/weighted average response of those samples.

We fitted a k-nearest neighbors model to the training set with k as 11 and the samples weighted by how close they are to the prediction target. The value of k was chosen by performing a grid search over a range of possible values and  $k = 11$  yields the lowest MSE. The resulting model has a test set MSE of 825.6 and the cross validation MSE is 841.8. The performance of the KNN model is noticeably better than the linear regression models, which have MSEs around 900.

As mentioned at the beginning of this section, we evaluated the importance of predictors using permutation importance, as shown in Figure 4.2. The most important variable in this KNN model is the number of related CDs, reducing around 500 MSE on average. Additionally, the mean number of reviews of related CDs, being a CD featuring rock music and the time elapsed after release also greatly affect how many reviews a CD got. The other predictors are of lesser importance, but most still help reduce the MSE by some amount.

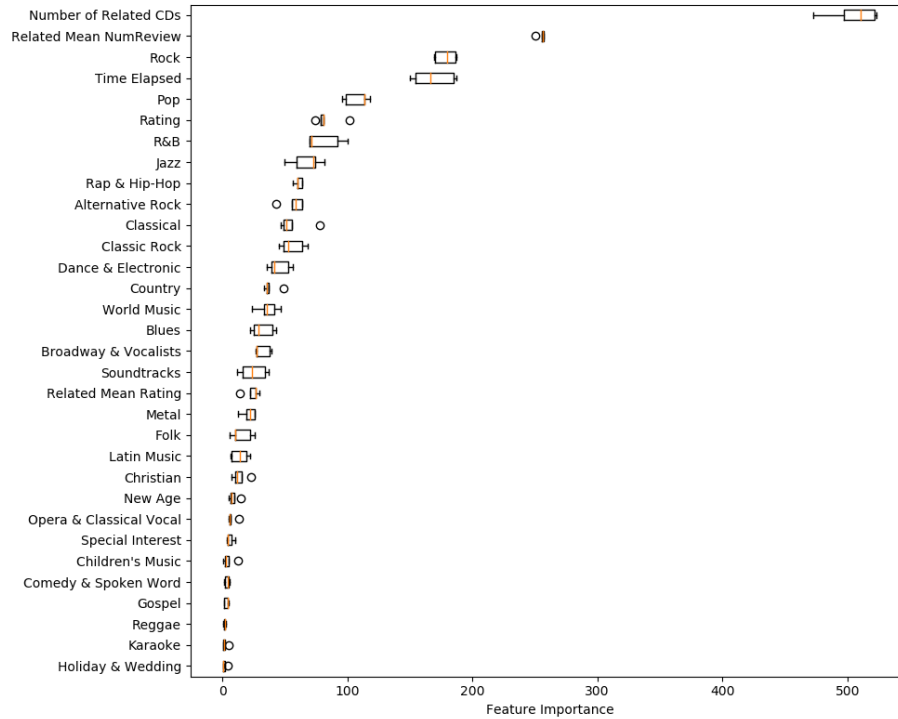


Figure 4.2: Permutation Importances from KNN

### 4.3.2 Decision Tree

A decision tree model has a tree/flowchart-like structure, with each node splitting the data into two groups until there is not enough samples in all the nodes. Each node splits the data based on the value of one predictor, and the goal at each node is to maximize the difference in the response variables of the two groups.

Using a grid search, the optimal hyperparameters for the decision tree model is a max tree depth of 6 and 1 as the minimum samples required in each leaf. The test set MSE of the decision tree model we fitted is 774.2 and the cross validation MSE is 742.2. This is further improvement over the KNN model and it also suggests that there are important non-linear relationships between the predictors when predicting the number of reviews.

The permutation importances of predictors in the decision tree is presented in Figure 4.3. Unlike the KNN model, the mean number of reviews of related CDs is the

most important variable in this model, being able to reduce the MSE by nearly 1000. The number of related CDs, which had the biggest impact in the KNN model, now decreases MSE by 600 to 800. Furthermore, the time elapsed and the mean ratings of related CDs improve MSE mildly. Lastly, all the other variables have almost no impact on the prediction error. This is related to the nature of the decision tree model not using all the predictors available.

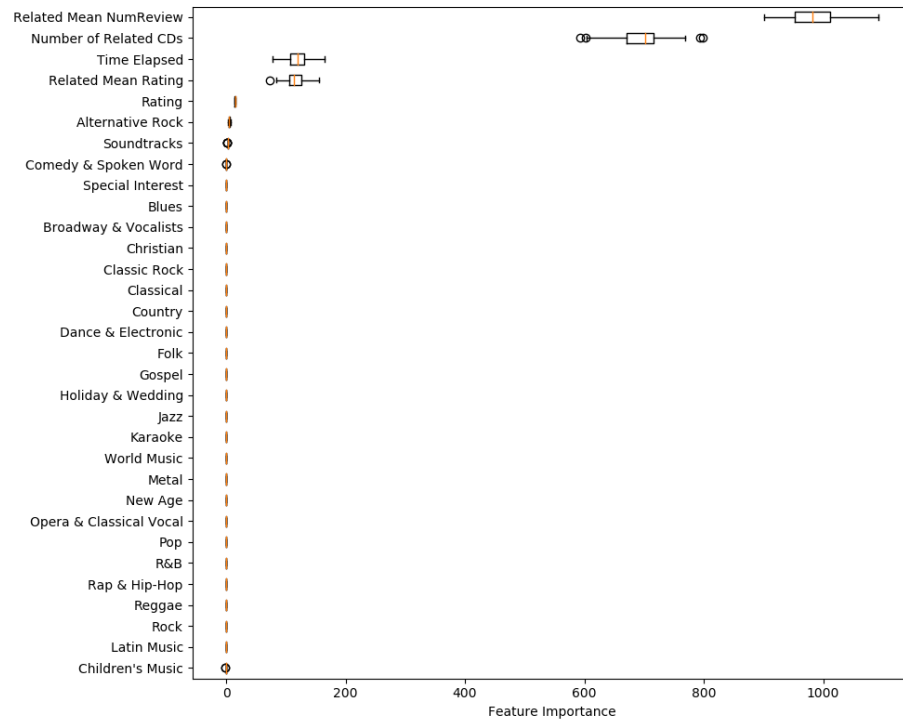


Figure 4.3: Permutation Importances from Decision Tree

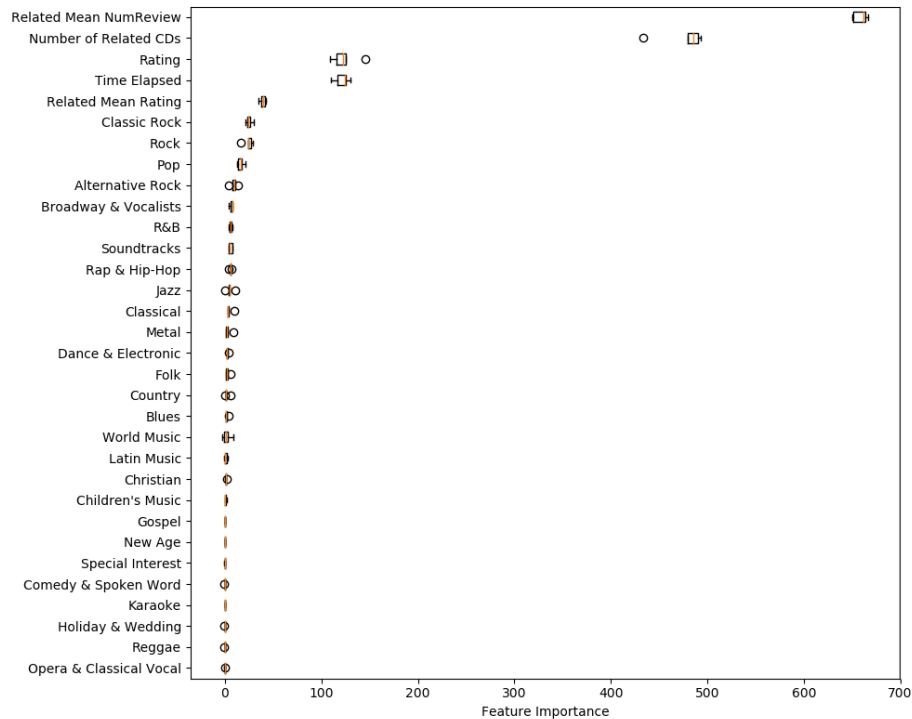
### 4.3.3 Random Forest

Random forest [15] is an ensemble method that aggregates the results of a collection of decision trees. The main motivation for the random forest model is to overcome the issue of overfitting the training set, which individual decision trees often do. This is done by fitting a multitude of decision trees, with each having access to only a portion of samples and predictors. The prediction of a random forest model is the mean of the predictions of all trees. Doing so increases the bias of individual decision

trees (because they have access to less data) but reduces the variance of the overall prediction, leading to a decrease in MSE.

The random forest model fitted to our dataset has 500 trees, each using  $\sqrt{32} \approx 6$  predictors and the minimum sample required at a leaf is 1. Its MSE on the test set is 637.9 and has an average cross validation error of 620 MSE. Clearly, the random forest model is able to make even more accurate predictions than a single decision tree, and it also beats all the models we have investigated so far.

The permutation importances of predictors in the random forest model are plotted in Figure 4.4. Similar to the results in decision tree, the mean number of reviews of related CDs and the number of related CDs are still the top two most important variables. In addition, time elapsed has similar impact to the random forest model as in decision tree, while rating now plays an important role in reducing MSE for the random forest.



**Figure 4.4: Permutation Importances from Random Forest**

### 4.3.4 Gradient Boosting Tree

Gradient boosting tree [16] is also a machine learning method that constructs an ensemble of decision trees. In contrary to random forest, gradient boosting tree utilizes the boosting algorithm that builds trees iteratively such that trees grown later focuses on samples which are predicted poorly in previous trees. The model is constructed by adding decision trees, each regressing on the residuals of previous trees. Since the residuals represent aspects that older trees failed to predict, the newer tree is able to improve the overall model by concentrating on predicting the samples with large residuals (poor prediction) previously. This feature of boosting tree models allows it to target samples that are difficult to predict as it includes more trees, resulting in higher predictive power.

Using grid search with cross validation, the optimal hyperparameters found for our boosting tree model are 600 trees, max tree depth of 10, minimum samples required at each split being 20, and a learning rate of 0.01. The test set MSE of this model is 640 and it has a cross validation MSE of 615.3. The predictive performance of the gradient boosting tree is almost the same as the random forest model and is also one of the best models we have.

The variable importance in the gradient boosting tree model is very similar to those in random forest. From Figure 4.5, we see that the top five most important variables are exactly the same and only the ordering is slightly different from the result of random forest. This should not come as a surprise since the two model are both ensemble methods based on decision trees.

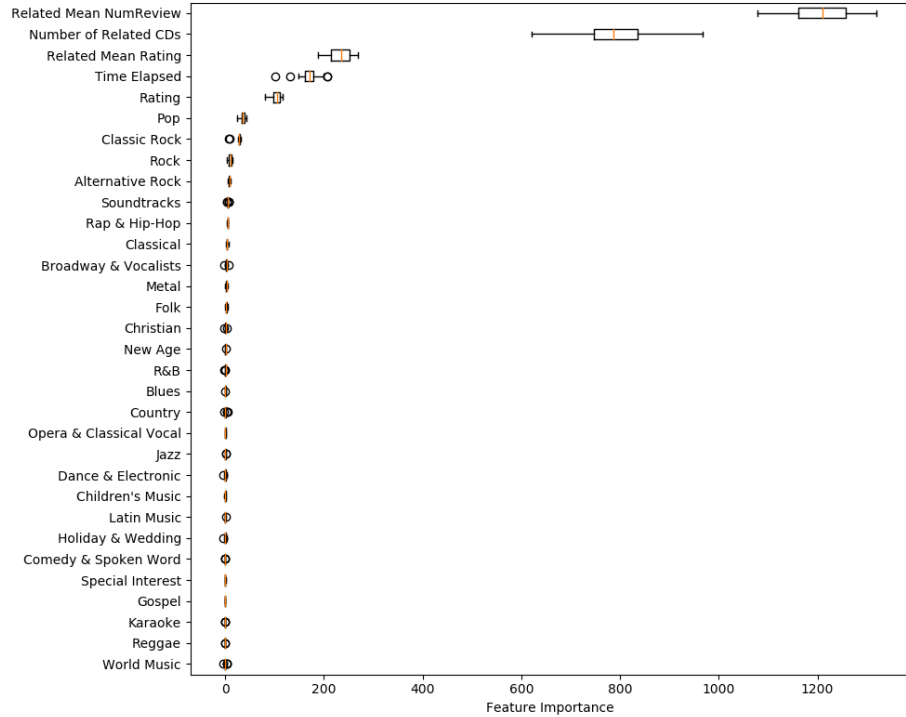


Figure 4.5: Permutation Importances from Gradient Boosting Tree

### 4.3.5 Neural Network

A neural network model is a collection of connected nodes, inspired by the biological neurons in brains of animals. Each node summarizes the information of inputs it receives and outputs to other nodes. Neural networks often consists of three parts—the input layer, the hidden layers, and the output layer. The input and output layers contains nodes with the predictors and response variables, respectively. Hidden layers can contain any number of nodes the user desires and using multiple hidden layers gives neural network models the ability to approximate complex non-linear functions.

We fitted a fully connected neural network with 3 hidden layers, each having 256 nodes. To avoid overfitting, all hidden layers uses a dropout [17] rate of 20%. Additionally, we splitted 20% of the training set into a validation set and the model stops updating its parameters if the error on the validation set stops improving for 30 epochs. Using the Adam optimizer [18] with a batch size of 128, our neural network

model achieved a test set MSE of 703.6 and the average cross validation MSE is 687.4. The performance of the neural network is slightly worse than random forest and gradient boosting tree but is still better than all other models.

Figure 4.6 displays the feature importances in the neural network model. Once again, the mean number of reviews of related CDs and the number of related CDs take the top two spots, being able to reduce the MSE by 700–800. Nevertheless, the rating of the CD and the mean rating of related CDs have very little impact on the prediction error in the neural network model.

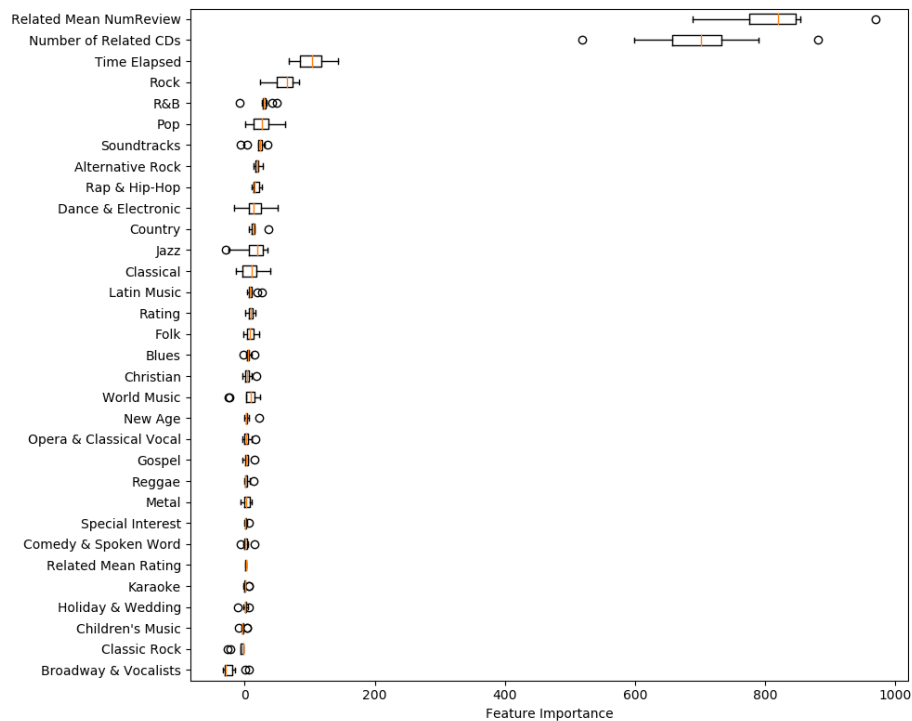


Figure 4.6: Permutation Importances from Neural Network

## 4.4 Model Comparison

We have fitted eight machine learning models to predict how many reviews CDs got in the Amazon dataset. Using permutation importance to measure how much MSE each predictor helps reduce, we found that the mean number of reviews of related CDs is consistently the most prominent feature in the models with good performances.

This shows how people often review/purchase CDs that are related or suggested to them on the Amazon website. Moreover, the rating of a CD, the number of related CDs, the mean rating of related CDs and the time elapsed after release are also important features in most models. Lastly, the music genres Rock, Classic Rock and Pop sometimes have minor impact on the average number of reviews CDs got, while all other genres appear to have little to no impact on the predictive accuracy.

Based on the analyses in this chapter, we discovered that the most important features for predicting the number of CD reviews are various summary statistics of related products as well as how old the CD is (time elapsed). Surprisingly, the rating of a CD or the music genres it belongs to do not have major impacts on how many reviews CDs got. This could be due to the power law nature observed in section 2.1. Since the vast majority of CDs will have very low number of reviews, the predictive accuracies of the models depend heavily on the small portion of CDs with a lot of reviews. Those CDs are rare outliers and general features such as their ratings or music genres may not be very useful in predicting them. This is because too many CDs with a low number of reviews have good ratings and belongs to popular music genres, resulting in a overall low average number of reviews for those categories. Consequently, the only predictors that help predict these CDs with a lot of reviews are the features of related CDs (likely popular as well) and the time elapsed, which are different for every CD.

The mean squared error is used to evaluate prediction accuracies of our models. Figure 4.7 shows the cross validation MSEs of all eight models. The best models in predicting the number of reviews are the random forest and gradient boosting tree model, both having an average MSE near 600. The neural network model is a close third, with a mean MSE around 700. Despite being a very simple model, the decision tree performed relatively well compared to the first five models. The KNN model has the worst performance out of all the models in section 4.3, however it still managed to beat linear regression models. The linear regression models do not perform well

on predicting the number of reviews. We suspected that this is because important non-linear relationships exists between the predictors and the response and the linear models are unable to capture them. Overall, we recommend using the two tree based ensemble model for this task and use them as benchmarks when exploring better models.

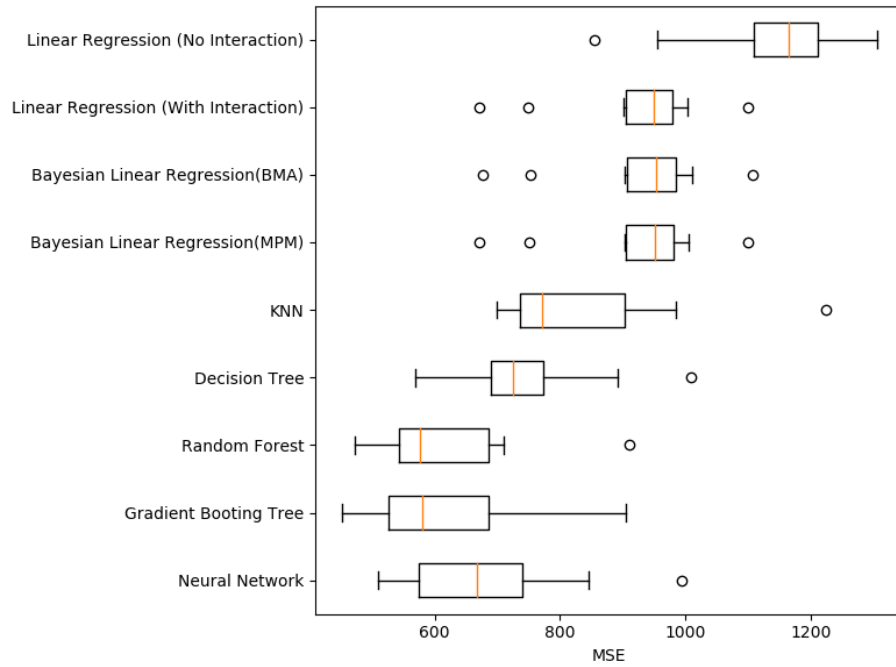


Figure 4.7: Cross Validation MSE of Machine Learning Models

# 5

## Conclusion

We analyzed the Amazon CD review dataset using a range of statistical and machine learning methods. Based on exploratory analyses, we discovered that the number of reviews follows a power law distribution regardless. In addition, we also investigated the relationships between music genres and other features. Furthermore, we identified general patterns of CD popularity from both the viewpoint of the customers and the producers. When analyzing CD life cycles, four different patterns were found and the most common pattern is the single exponential decay life cycle. For CDs with this type of life cycle, the Bayesian monitoring technique was used to pinpoint large popularity spikes in the data. To deal with extreme changes present in CD life cycles, we proposed an adjusted Bayesian monitoring method through the use of power discounting. Lastly, we compared several popular models for predicting how many reviews CDs get and the best performing models are the random forest and boosting tree. The work presented in this paper would help formulate better business strategies as well as provide guidance in analyzing data with similar features.

## Bibliography

- [1] Julian McAuley et al. “Image-based recommendations on styles and substitutes”. In: *SIGIR* (2015).
- [2] Mike West. “Bayesian Model Monitoring”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 48, No. 1 (1986), pp. 70–78.
- [3] Jeff Alstott, Ed Bullmore, and Dietmar Plenz. “powerlaw: a Python package for analysis of heavy-tailed distributions”. In: *PLoS ONE* 9(1): e85777 (2014).
- [4] Paul Jaccard. “Étude comparative de la distribution florale dans une portion des Alpes et des Jura”. In: *Bulletin de la Société vaudoise des sciences naturelles* 37 (1901), pp. 547–579.
- [5] G. Day. “The product life cycle: Analysis and applications issues”. In: *Journal of Marketing* 45 (1981), pp. 60–67.
- [6] Andrew Gelman, Xiao-Li Meng, and Hal Stern. “Posterior Predictive Assessment of Model Fitness Via Realized Discrepancies”. In: *Statistica Sinica* 6, No. 4 (1996), pp. 733–760.
- [7] G. E. P. Box and D. R. Cox. “An analysis of transformations”. In: *Journal of the Royal Statistical Society, Series B.* 26 (1964), pp. 211–246.
- [8] Andrew Gelman et al. *Bayesian Data Analysis, Second Edition*. Chapman and Hall/CRC, 2003.
- [9] Merlise Clyde. “Bayesian Model Averaging and Model Search Strategies (with discussion)”. In: *Bayesian Statistics*. Ed. by J.M. Bernardo et al. Vol. 6. Oxford University Press, 1999, pp. 157–185.
- [10] Merlise Clyde. *BAS: Bayesian Variable Selection and Model Averaging using Bayesian Adaptive Sampling*. R package version 1.5.5. 2020.
- [11] Feng Liang et al. “Mixtures of g Priors for Bayesian Variable Selection”. In: *Journal of the American Statistical Association* 103.481 (2008), pp. 410–423. DOI: 10.1198/016214507000001337. eprint: <https://doi.org/10.1198/016214507000001337>. URL: <https://doi.org/10.1198/016214507000001337>.
- [12] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

- [13] François Chollet et al. *Keras*. <https://keras.io>. 2015.
- [14] André Altmann et al. “Permutation importance: a corrected feature importance measure”. In: *Bioinformatics* 10 (2010), pp. 1340–1347.
- [15] Breiman. “Random Forests”. In: *Machine Learning* 45(1) (2001), pp. 5–32.
- [16] J. Friedman. “Greedy Function Approximation: A Gradient Boosting Machine”. In: *The Annals of Statistics* 29, No.5 (2001).
- [17] Nitish Srivastava et al. “Dropout: A Simple Way to Prevent Neural Networks from Overfitting”. In: *Journal of Machine Learning Research* 15 (2014), pp. 1929–1958. URL: <http://jmlr.org/papers/v15/srivastava14a.html>.
- [18] Jimmy Ba Diederik P. Kingma. “Adam: A Method for Stochastic Optimization”. In: 3rd International Conference for Learning Representations (San Diego). 2014.