Innovation Diffusion Analysis for the Project Gutenberg Open Audiobook Collection

Elina Ohanjanyan

2024-10-18

1. Choose an innovation from the list.

The innovation I chose from the list is the Project Gutenberg Open Audiobook Collection. It's an open Audiobook collection that has been created using text-to-speech. It is a bit different and difficult to compare to the rest of the innovations on the list and hard to analyze from the business perspective, as the collection is open and available to everyone. However, we can still try to predict its market potential.

2. Identify a similar innovation from the past.

A similar innovation that I chose from the past are the regular Audiobooks. The two innovations are different in their creation process: one is made through the work of real people reading and recording books, while the other is an automated process of using text and speech data in order to make more books available for listening. However, the underlying reason and the purpose for both of these innovations are the same - they were made to make books and other visual material more accessible to people who have problems with their vision or reading, or even just people who don't like to read but want to know the content of books.

Their functionalities are also practically the same - they are books in an audio format. Their market impact is a bit difficult to compare, as the books in the Project Gutenberg Open Audiobook Collection are pretty much an extension of the already existing audiobook market. So, the largest difference is in the underlying technology - one has been made and recorded by human why the other has been made using text-to-speech models.

3. Find historical data.

The data I chose to understand the audiobook market is this one: https://www.statista.com/statistics/305733/consumer-audiobook-download-sales-revenue-in-the-uk/ and it shows the audiobook download revenues in the UK from 2009 to 2023.

The problem with choosing this dataset is the fact that it shows revenue data, but not the data of how many units of audiobooks have been sold, which I would ideally need for the bass model. However, I had to settle for this data due to the limited audiobook data by units I could find online.

To remedy the problem of not havimg data on the number of audiobooks sold per year, I decided to find the average price of one audiobook in the UK and divide the revenues by that number, to get a rough estimate of how many units have been sold.

I was also unable to find any specific statistics of what the average audiobook price is in the UK, so I visisted amazon.co.uk, looked up the prices of the audiobooks (they ranged from around 12 pounds to 18), so I decided to divide the revenue by ~15 to get a rough estimate of the number of units sold. I understand that it's a bad way of doing things, and it doesn't take into consideration the possible inflation of audiobook prices, but the data was REALLY really limited.

```r
library(diffusion)
library(ggplot2)
library(reshape2)
library(ggpubr)
library(readxl)
library(gridExtra)
```

```r
data <- read_excel("../data/statistic_id305733_audiobook_revenue_in_the_united_kingdom_uk_2009.xlsx", sh
colnames(data) <- c("Year", "Sales")
data$Sales <- data$Sales * 1000000 / 15 #as the data is in millions and then divided by 15
data
```

```
## # A tibble: 15 x 2
##    Year      Sales
##    <chr>     <dbl>
##  1 2009    133333.
##  2 2010    200000
##  3 2011    333333.
##  4 2012    466667.
##  5 2013    800000
##  6 2014   1533333.
##  7 2015   2000000
##  8 2016   2600000
##  9 2017   3200000
## 10 2018   4600000
## 11 2019   6466667.
## 12 2020   8866667.
## 13 2021  10066667.
## 14 2022  10933333.
## 15 2023  13733333.
```

4. Estimate Bass Model parameters.

```r
t <- seq(1, length(data$Year)) #to get sequence for the years
diffusion_model <- diffusion(data$Sales)
diffusion_model$w
```

```
##           m           p           q
## 1.835094e+08 1.534590e-03 3.121426e-01
```

We can round up the values for m, p and q.

```r
m <- round(diffusion_model$w, 3)[1]
p <- round(diffusion_model$w, 3)[2]
q <- round(diffusion_model$w, 3)[3]
cat("Estimated Parameters:\n")
```

```
## Estimated Parameters:
```

```r
cat("Market potential:", m, "\n")
```

```
## Market potential: 183509423
```

```r
cat("Innovation coefficient:", p, "\n")
```

```
## Innovation coefficient: 0.002
```

```r
cat("Imitation coefficient:", q, "\n")
```

```
## Imitation coefficient: 0.312
```

5. Predict the diffusion of the innovation selected in step 1.

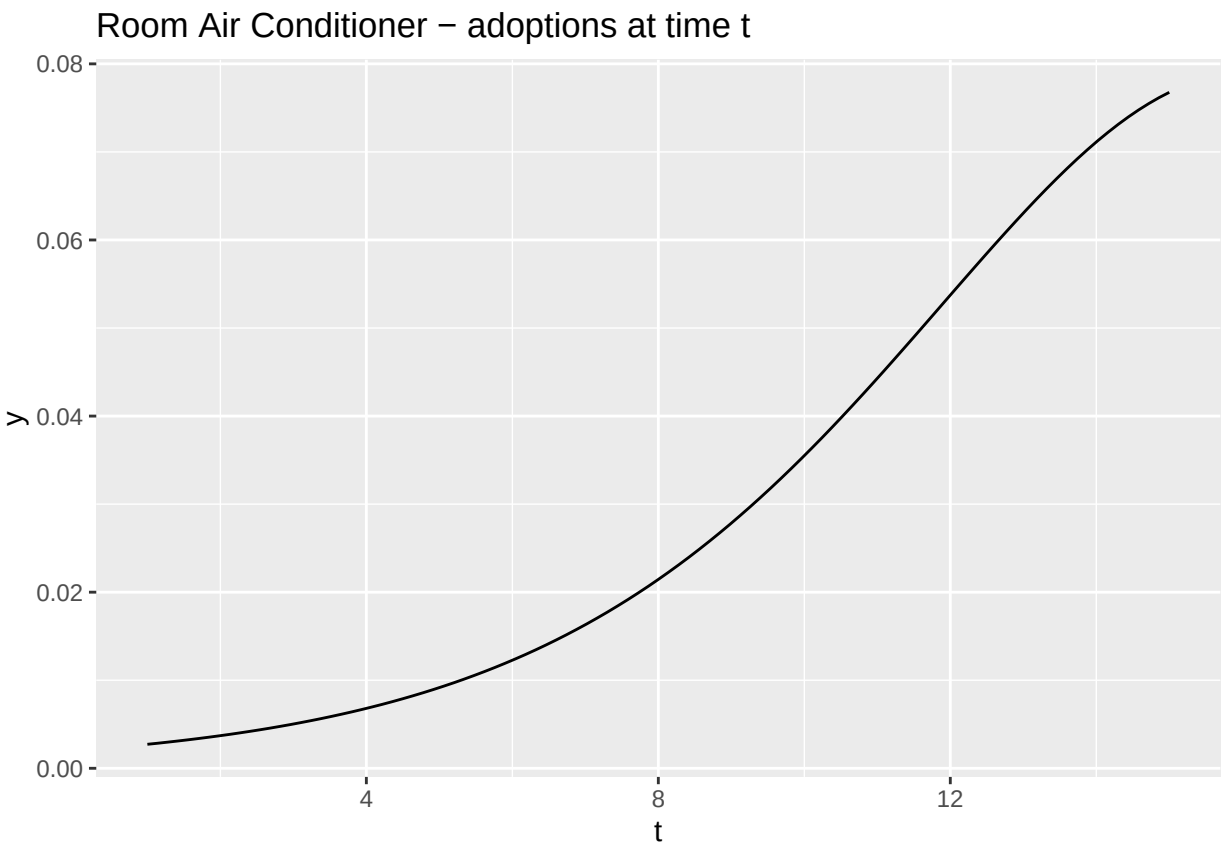
The formula for f(t) from the slides:

```r
bass.f <- function(t,p,q){
((p+q)^2/p)*exp(-(p+q)*t)/
(1+(q/p)*exp(-(p+q)*t))^2
}
```

The formula for F(t) from the slides:

```r
bass.F <- function(t,p,q){
(1-exp(-(p+q)*t))/
(1+(q/p)*exp(-(p+q)*t))
}
```

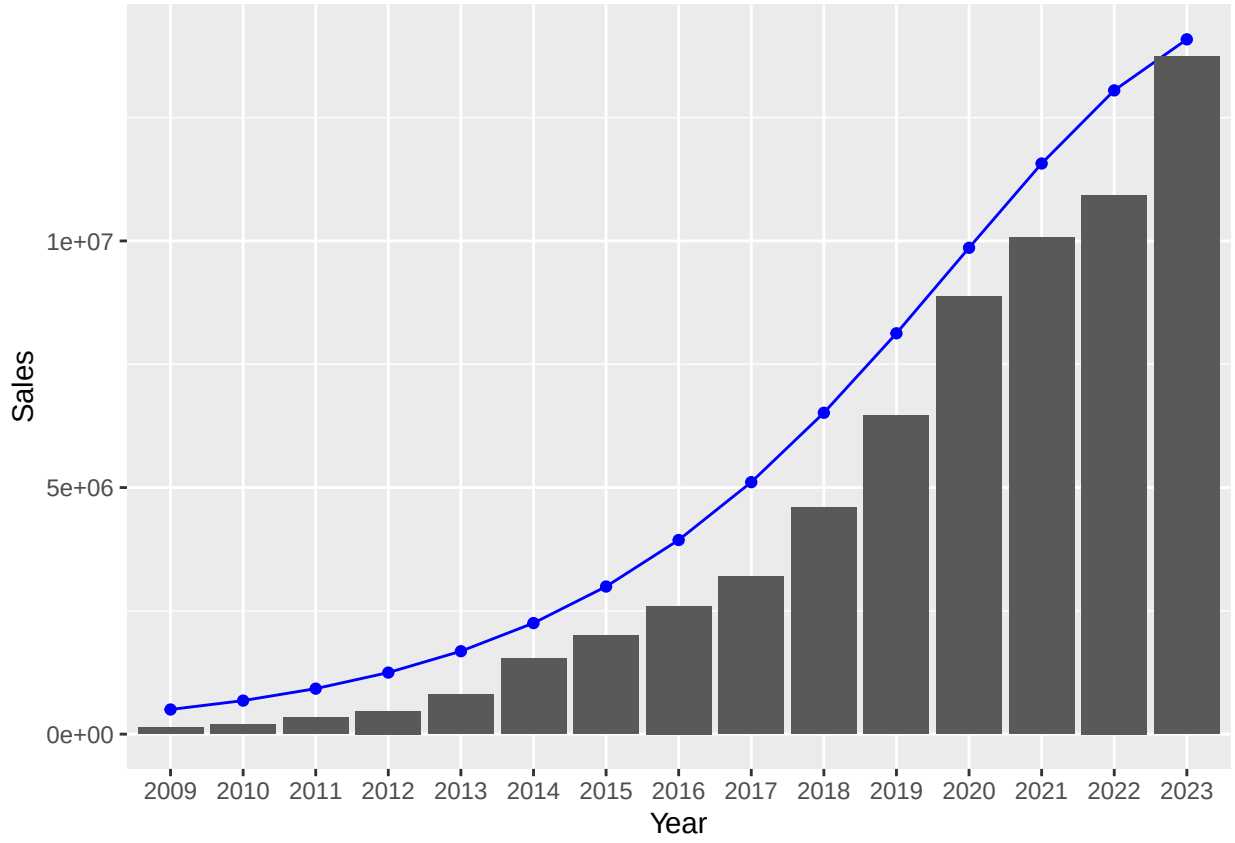
```r
ggplot(data.frame(t = c(1, 15)), aes(t)) +
stat_function(fun = bass.f, args = c(p, q)) +
labs(title = 'Room Air Conditioner - adoptions at time t')
```



So far, we can see that the product is still in its rising stages. It has not reached its full potential and share of the market, so it will probably keep growing before it starts decreasing in time.

Predicting

```
data$preds = bass.f(t, p, q) * m
ggplot(data = data, aes(Year, Sales)) +
    geom_line(aes(Year, preds, group = 1), color = 'blue') +
    geom_point(aes(Year, preds, group = 1), color = 'blue') +
    geom_bar(stat = 'identity')
```



6. Choose a scope (global or country-specific).

As the data is about the audiobook revenue in the UK, I have no choice but to choose the country-specific scope.

7. Estimate the number of adopters by period.

```
years <- seq(1, length(data$Year) + 5) #predicing for the future 5 years too
new_adopters <- bass.f(years, p, q) * m

adopters_data <- data.frame(
  Year = years,
  New_Adopters = new_adopters
)
adopters_data
```

```
##    Year New_Adopters
## 1     1     500055.7
## 2     2     680149.1
## 3     3     922949.3
## 4     4    1248470.7
## 5     5    1681597.3
## 6     6    2251984.8
```

```
## 7     7    2992693.9
## 8     8    3936539.5
## 9     9    5108875.9
## 10   10    6515720.1
## 11   11    8127402.7
## 12   12    9861057.3
## 13   13   11570054.9
## 14   14   13052355.7
## 15   15   14087128.9
## 16   16   14495409.8
## 17   17   14201020.6
## 18   18   13259204.6
## 19   19   11835709.6
## 20   20   10149038.7
```

The diffusion stages by segment:

```
innovators <- m * 0.025
early_adopters <- m * 0.135
early_majority <- m * 0.34
late_majority <- m * 0.34
laggards <- m * 0.16
```

```
# Example of total predicted adopters
total_predicted_adopters <- sum(data$preds)  # This should be the total predicted adopters based on you

# Define the percentages for each segment
percentages <- c(Innovators = 0.025,
                 Early_Adopters = 0.135,
                 Early_Majority = 0.34,
                 Late_Majority = 0.34,
                 Laggards = 0.16)

# Number of adopters in each segment
adopters_by_segment <- total_predicted_adopters * percentages

adopters_df <- data.frame(
    Segment = names(adopters_by_segment),
    Number_of_Adopters = adopters_by_segment
)

# Resulting dataframe
adopters_df
```

```
##                       Segment Number_of_Adopters
## Innovators         Innovators            2063426
## Early_Adopters Early_Adopters           11142500
## Early_Majority Early_Majority           28062592
## Late_Majority   Late_Majority           28062592
## Laggards             Laggards           13205926
```