Master's Thesis of Business Administration

Pricing Subscription Services with Text Data using Hedonic Pricing and Machine Learning

헤도닉 가격모형과 머신러닝을 활용한 구독서비스의 텍스트 기반 가격 책정

February 2023

Graduate School of Business
Seoul National University
Operations Management Major

Jo, Min-hwa

Pricing Subscription Services with Text Data using Hedonic Pricing and Machine Learning

Park, Sang-wook

Submitting a master's thesis of
Business Administration

February 2023

Graduate School of Business
Seoul National University
Operations Management Major

Jo, Min-hwa

Confirming the master's thesis written by
Jo, Min-hwa
February 2023

Chair	Nam, Ick-hyun	(Seal)
Vice Chair	Lim, Michael K.	(Seal)
Examiner	Park, Sang-wook	(Seal)

Abstract

Pricing Subscription Services with Text Data using Hedonic Pricing and Machine Learning

Jo, Min-hwa
Business Administration (Operations Management)
The Graduate School
Seoul National University

Subscription e-commerce market has grown by more than 100 percent a year over the past five years and the importance of study on subscription services is self-evident nowadays. However, prices in subscription services usually have unclear structures even though offering a reasonable price is one of the most important elements of customer relationship management in service area. Thus, the aim of the study is to consider attributes based on customer preferences obtained by user-generated data and predict price for subscription services more objectively in a way that customers accept it reasonable.

Target data are 10,000 web scraped reviews from representative subscription services of video streaming service brands. Using machine learning techniques, topic modeling, vector space model, dimensionality shrinkage and deriving the value of attributes were done, and hedonic pricing model was defined.

In this process, the result shows that a highly predictive value can be obtained by considering the covariance between each value through Partial Least Squares regression that enables supervised Latent Dirichlet Allocation when pricing services with text-based data. Moreover, regarding the result of Netflix, Amazon

Prime Video, and HBO Max being overpriced, Disney+, and Hulu being underpriced, this paper presented insights and implications for each service through considering the relationships between price and attributes that customers value and other situations that the services face.

……

Keyword : Pricing, Subscription Services, Machine Learning, Latent Dirichlet Allocation, Partial Least Squares, Hedonic Pricing Model
Student Number : 2021-27567

Table of Contents

Chapter 1. Introduction

With the development of new technologies and the deregulation of several industries, the number of suppliers of subscription services has grown significantly during the last two decades (Mesak & Darrat, 2022). The importance of studying subscription services is self-evident since, at present, almost every household is involved, in one way or another, in these services (Fruchter et al., 2013).

By a report from McKinsey & Company, the subscription e-commerce market has grown by more than 100 percent a year over the past five years. Fueled by venture-capital investments, start-ups have launched these businesses in a wide range of categories. This strong growth has attracted established consumer brand manufacturers and retailers, making them all launch new subscription businesses (e.g., P&G (Gillette on Demand), Sephora (Play!), and Walmart (Beauty Box)) (Chen et al., 2018).

Opportunities for more subscription-based services will increase, and in this environment, pricing is increasingly recognized as a key element of customer relationship management in subscription services (Crescenzi, 2020). Considering price of subscription services, one of the major challenging questions that managers of business and scholars face is how to price these subscription services over time.

However, prices in services usually have unclear structures. Thinking of a representative subscription service, OTT(Over-the-top) video streaming services, the costs are usually hard to identify clearly, and are subjectively measured. For example, it is quite common for subscription service users to experience unreasonable price derived by companies that measure their price by referring to competitors or unknown variation on prices, which makes customers leave the service after all.

Thus, the aim of the study is to consider attributes based on customer preferences obtained by user-generated data and predict price for subscription services more objectively in a way that customers accept it reasonable. After predicting the price with the attributes obtained, comparison between current price and the obtained price will be done to show the difference.

To obtain the right attributes, online review data will be used. Since online reviews contain valuable information from customers such as satisfaction, loyalty, price sensitivity, willingness to pay, and so on, online reviews written by users of subscription services will be collected through web scraping. With that, the attributes will be defined through topic modeling the text data from online reviews.

In general, text data contains a lot of information that are hard to be reduced. To solve this, machine learning methodologies to discover latent concepts (or topics) inside collection of documents will be used. The most widely used method is Latent Dirichlet Allocation (LDA) among these methods, which is a generative probabilistic model for topic modeling that has also been used to create features for classification and regression purposes (Blei et al., 2003). To figure out the values of the information which comes out as attributes and independent variables, LDA is developed by Partial Least Square (PLS) regression as Supervised LDA. This approach will be the main method for this research to handle text data.

By using the above methods, the attributes obtained here will be made into a regression model by using Hedonic Pricing Model. As these models come with the price regarding the functions of its observable attributes, hedonic models can be applied for the interpretation (Feenstra and Shapiro, 2008). This will be a process to set up hedonic value for the text description by estimating the implicit price of the words providing details on attributes that drive customer preferences.

Chapter 2. Literature Review

1.1. Hedonic Pricing

Hedonic price model was first brought by Waugh (1928) with the idea of finding quality factors that influence vegetable prices, and Court (1939) with defining hedonic price indexed with automotive examples. Later, contributions to this model from Lancaster (1966) with the theory of consumer preferences and Rosen (1974) with examining characteristics from real estate area made the establishment on hedonic pricing model.

Hedonic price model is based on the thought that utility of consumptions on products or services is brought by the attributes or characteristics of the goods, not from themselves. (Liang & Yuan, 2021). The effect of each feature on price regarding utility and satisfaction from consumers is called the hedonic price of this feature. The essence of the hedonic price model is to explain the change in heterogeneous market prices by the quantity change of features (Liang & Yuan, 2021).

From a managerial perspective, it is critically important to understand consumer perceptions of each of the attributes associated with the price (Chen & Rothschild, 2010). In this point, Papatheodorou & Apostolakis (2012) explains basic knowledge and applications on how to analyze hedonic price. They show how hedonic price analysis can estimate price for characteristics that are valued by consumers even for non-market product or under service characteristics. Also, related to research on quality management, Magno et al. (2018) introduced accommodation prices on Airbnb, with the purpose of filling the gap between prices for shared accommodations based on their experience with price management and on the level of market demand by suggesting and testing a comprehensive hedonic pricing model. This shows how managerial decision making and evaluating individual preferences can be done by using hedonic pricing.

For more use of hedonic model, Brynjolfsson & Kemerer (1996) introduced a hedonic model to determine the effects of network externalities, standards, intrinsic features, and a time trend on microcomputer spreadsheet software prices using hedonic regression model with product sample data collected over 6 years. Wang & Zhou (2008) introduces a structural equation model by using hedonic theory to conduct empirical analysis to deduce the degree of residential segregation depends on the degree of consumer preference for location and attributes and those result in different space structures.

Liu & Wu(2009) conducted research to analyze the residential product's value based on structural equation model and hedonic price theory, and Wang & Tong (2010) defined housing characteristics as independent variables for market supply-demand equilibrium model, and conducted an empirical analysis on the housing price in Harbin City based on hedonic model. Zhao & Liu (2010) had gone through a hedonic price study on urban housing by case study to explain the quantitative relationship between housing price and housing characteristics based on hedonic model an analyzed it through OLS regression.

Ivanov & Piddubna (2016) developed hedonic models to analyze determinants of prices, price dynamics and rate parity across distribution channels with data from 150 accommodation establishments' websites. Bacon et al. (2016) also contributed to revenue management area regarding hedonic regression. They investigated the effect of operational attributes and product type on the price that consumers paid in restaurants by testing hypotheses with hedonic regression analysis and experimented with secondary survey data from restaurants in New York City area.

Up to knowing the concept and applications for hedonic pricing, considering then which of the attributes will impact the pricing, research on determinants has also been done by Liang & Yuan (2021). They defined the impact of various quality attributes and network-specific characteristics on the price of word

processing software. Their works are valued for making use of the big data with the pricing model. However, applications on how these attributes can practically be used onto the hedonic model, particularly in service environment have not yet been done, which shows the need for further study.

1.2. Pricing Subscription Services

Several papers have worked on different pricing models for the subscription services. Regarding scales that impact the price of subscription model, Samanta et al. (2007) conducted research on the impact of price, with the application to mobile subscription. They brought up a model that increased significant amount of revenue in the data provided company in India. Their research shows that a pricing strategy based on provisioning cost, when combined with a nonlinear demand model can potentially increase both the penetration for a service and the revenue to the operator (Samanta et al., 2007).

Another factor defining research conducted by Nagaraj et al. (2021) was of measuring scales that affect the willingness to subscribe. The research considered OTT (Over-the-top) video streaming services in India. The study identified reasons for subscribing or un-subscribing OTT services and identified five factors that affect consumer's preferences on online subscription. The data for this research were obtained by conducting survey using a data collection platform. Regarding this, as the authors mentioned, data collection seems to have a room for improvement since OTT services are one of the most subscription services that are known for using bigdata for consumer preferences.

Miao et al. (2022) have worked on profit model for Electronic Vehicle (EV) rental service. EV rental services are not as much popular as other subscription services that people might think of, but scholars and practitioners see it as a promising area for subscription. Mio et al. (2022) constructed a comprehensive

profit model of EV rental service based on queuing theory in order to establish time-based subscription pricing. Blocking level of the service system was considered most important for consumers' decisions. Key factors that impact on profit was also studied.

Other works regarding pricing schemes have been investigated in different way. Bala (2012) deals with strategies that firms, and consumers take under competition. Given optimal pricing by firms, followed by customer type, they try to define the demand of each strategy, buy, rent/subscribe, or do nothing to obtain the product. However, the renting system, or subscription service in this research is mostly considering the characteristics of partly outdated products such as newspaper since it is written in 2012. Thus, more studies considering subscription services up to date seems necessary.

Fruchter et al. (2013) conducted study on dynamic pricing for subscription service. With subscription fee, they also considered activation and cancellation fees. Another modeling approach for optimal pricing was conducted by Danaher (2002), they derived a theoretical model that combined two basic phenomena of subscription services, namely, usage and retention. Regarding this the researcher defined a revenue maximization model regarding usage price and price sensitivity, and data were used to apply from a telecommunications service.

Other than this, since the pricing methods are usually differently measured for different characteristics of customers, study on customer segmentation with subscription based online media customers was done by Haatanen (2022), clustering the data of click-data size for the users of the media website.

Regarding these segments, Punj (2013) also studied on the relationship with willingness to pay for different customer characteristics in subscription-based business. Findings indicated that consumers who are most willing to pay for content and those who are not shows definitive contrasts in terms of gender and age and to a lesser degree in terms of income and education. (Punj, 2013)

1.3. Hedonic Pricing on Subscription Services

Hedonic Pricing on areas such as tourism, or housing has been mostly done and mostly the hedonic pricing methods were used with survey data from customers. Gibbs et al. (2018) has gone through a specific analysis on both housing and tourism. They applied a hedonic pricing model for sharing economy and examined through Airbnb listing. For other research in a similar context, Stevens (2014) conducted a study on predicting price by text mining in the area of Real Estate which has more clear attributes to use a hedonic pricing method.

Most attracting study methods and context that will be referred to is of Crescenzi (2020)'s. The researcher established a hedonic pricing model that can be useful in regression approach and by using topic modeling methods through machine learning, the text data were analyzed and measured. Since this study was the only one that can be found as using text based hedonic pricing method through Machine Learning, it is clear that there is not much research on text-based hedonic pricing and for subscription services, and there is huge room for study.

1.4. Data Mining through Machine Learning

Analysis through machine learning technique has been arising these days, and especially for online review has been getting increasingly important, papers using text mining and prediction through machine learning are increasing. However, while there are plenty of methodological and applied contributions on the usage of text data for document clustering and classification (Weiss et al., 2015; Berry & Catellanos, 2004; Berry & Kogan, 2021; Steyvers & Griffiths, 2007), not as much research has been produced using text with regression models for hedonic pricing (Crescenzi, 2020). This gives the reason for using hedonic regression model for text-

mining though this research.

Pricing through machine learning has been done in several studies. Gupta & Pathak (2014) developed framework for the price prediction in order to predict purchases and amount of revenue that can be provided. This was done by using Machine learning technique. They have done data mining and through applying statistical methods in machine learning context, purchase behavior of online customers was predicted by each range of price for customers. The data that can be quantified were mostly collected in the study and they measured price through Dynamic Pricing.

Hernández & Rosales (2021) had introduced models to predict prices for real estate list using ensemble machine learning algorithms by use of regression models with the data of information about properties listed for sale from popular real estate websites.

Archak et al. (2011) also conducted study in this context. Instead of defining actual price, they worked on how much power would pricing have by mining consumer reviews using Machine learning. They used web-scraping on Amazon product review over 15 months and analyzed textual data in order to learn consumers' relative preferences for different product features and also how text can be used for predictive modeling of future changes in sales (Archak et al., 2011).

In data science and management area, Tan & Xiao (2021) also contributed on the research of hedonic pricing model on the purpose of defining effectiveness of online reviews on addressing price endogeneity issue in an application to consumer demand for smartphone. They compared hedonic pricing model and a conditional logit discrete choice model that had come out with data from online reviews.

Pricing through text has many difficulties since in the process of making text into a quantity, most studies go through curse of dimensionality. Referring to studies that resolves the problems occur by using text as data is necessary, and for this, studies of Gupta & Pathak (2014), Archak et al. (2011), Crescenzi (2020) are extremely helpful.

The methods of machine learning that are going to be used in this study such as Topic Modeling are comprehensively formulated in Tong & Zhang (2016) and Crescenzi (2020). They show text mining based on topic modeling methods of LDA, LSI, and analyzing the regression models with Lasso, which are all used as machine learning techniques. With the provided descriptions and approaches, this research is going to define pricing values that can be managerially influential for subscriptions services using supervised LDA, PLS regression and hedonic pricing.

1.5. Research questions derived and organized table of literatures

As a conclusion reviewing all these literatures, the goal of this research came out as two research questions as follows:

- *RQ1. How predictable would pricing subscription services with text be?*
- *RQ2. What more needs to be considered to better predict price for subscription services in addition to the works done before regarding pricing with text?*

The whole papers that have been mentioned in the literature review part are organized in Table1. Checks on which pricing strategy they've used, and whether they made use of text data, use of machine learning technics, or dealt with subscription services can be defined easily on Table 1 on the next page.

Table1. Table of pricing strategies, data, and methods studied in the literatures mentioned in comparison to this study.

	Pricing strategies				Use of text data	Use of Machine learning	Deals with Subscription services
	Hedonic pricing	Dynamic pricing	Linear demand model	Nonlinear demand model			
Punj (2013)	x	x	x	x	x	x	**o**
Danaher (2002)	x	x	x	x	x	x	**o**
Haatanen (2022)	x	x	x	x	x	**o**	**o**
Samanta et al. (2007)	x	x	x	**o**	x	x	**o**
Nagaraj et al. (2021)	x	x	x	x	x	x	**o**
Bala (2012)	x	x	**o**	x	x	x	**o**
Fruchter et al. (2013)	x	**o**	x	x	x	x	**o**
Tong & Zhang (2016)	x	x	x	x	**o**	**o**	x
Archak et al. (2011)	x	x	x	x	**o**	**o**	x
Gupta & Pathak (2014)	x	**o**	x	x	**o**	**o**	x
Wang &Zhou (2008)	**o**	x	x	x	x	x	x
Liang & Yuan (2021)	**o**	x	x	x	x	x	x
Papatheodorou & Apostolakis (2012)	**o**	x	x	x	x	x	x
Bacon et al. (2016)	**o**	x	x	x	x	x	x
Gibbs et al. (2018)	**o**	x	x	x	x	x	x
Ivanov & Piddubna (2016)	**o**	x	x	x	x	x	x
Brynjolfsson & Kemerer (1996)	**o**	x	x	x	△	x	x
Hernández & Rosales (2021)	**o**	x	x	x	x	**o**	x
Magno et al. (2018)	**o**	x	x	x	x	x	**o**
Miao et al. (2022)	**o**	x	x	x	x	x	**o**
Stevens (2014)	**o**	x	x	x	**o**	**o**	x
Crescenzi (2020)	**o**	x	x	x	**o**	**o**	x
Tan & Xiao (2021)	**o**	x	x	x	**o**	**o**	x

Chapter 3. Development Process and Methodologies

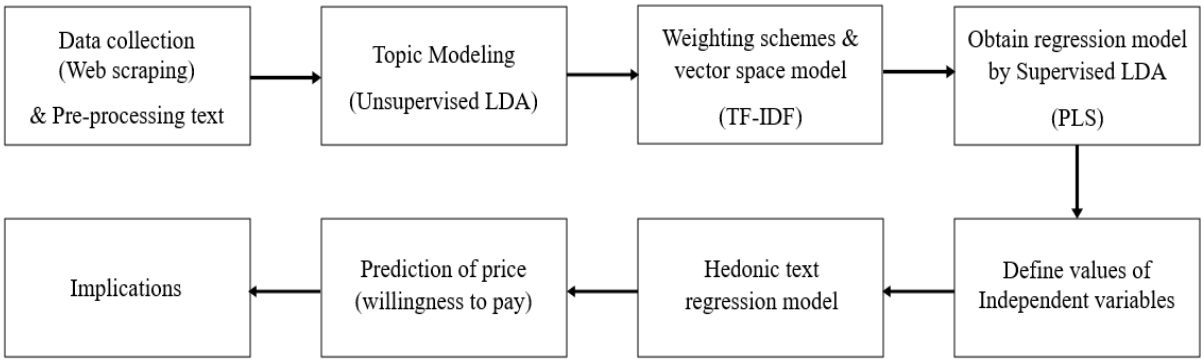


Table2. Model development process

To get the answer for the research questions, the development process and used methodologies would be organized as the table2 above.

Sources from text data are commonly unstructured for statistical applications. These data need specific algebraic model to be used for statistical applications to make a form of a matrix (George et al., 2016). Thus, to make it into a hedonic regression model to get the predicted price value, the text needs to be featured. To do this, the text descriptions of services are going to be text-mined and analyzed with the method of machine learning.

First, the reviews will be scraped, and the scraped data will be pre-processed. Since text data have high dimensionality, the first organization of it will be done by topic modeling. Well known topic modeling methods in machine learning is LDA (Latent Dirichlet Allocation). It is a model for text data where documents are considered as mixtures of topics and each topic is defined as a probability distribution over the vocabular. Text data featured by these methods will be used as attributes for Hedonic Pricing Model.

The topics and weights obtained from LDA seem that this can be used as an equation for obtaining the prices, but since there are no dependent variable set for the result of LDA, the results are regarded 'unsupervised.' To make this into a 'supervised' LDA to work as a regression model, the words abstracted from the topics are set as independent variables, and prices of each service are set as dependent variables. For the coefficients, each review will be weighed by the words from the topic, and the weights are going to be featured as a matrix through TF-IDF (Term Frequency - Inverse Document Frequency) vector space model.

Now, to gain aggregated coefficients and define a regression model, PLS (Partial Least Squares) regression is used. The reasons of using these methods will be explained in the next section with detailed explanation on PLS. Briefly explaining the process, by the outputs of 'LDA, TF-IDF, and dependent variables of prices', PLS regression is applied to predict prices while maximizing the covariance of dependent variables and linear combination of independent variables and weights. Using the results of prediction prices, better coefficients can be tested and since we know the weights here, the optimized independent variable can be defined.

Through this process, the values of independent variables are obtained. These values are going to be the price values of each attribute, and after obtaining the independent variables, the hedonic regression model can be defined finally. With the same process, separate prices of each service are gained, and the implication of these outputs will be discussed. Before moving on to follow up with the actual development process, Hedonic Pricing Model that are going to be used in this research will be introduced.

3.1. Hedonic Pricing Model

Hedonic pricing model (HPM) assumes that prices of differentiated goods can be described by a bunch of measurable features/attributes and that the consumer's valuation of a good can be decomposed into implicit values of each product features (Rosen, 1974). As mentioned in the literature review part of the study, the effect of each feature on price regarding utility and satisfaction from consumers is called the hedonic price of this feature. The essence of the hedonic price model is to explain the change in heterogeneous market prices by the quantity change of features (Liang & Yuan, 2021).

The general form of HPM can be described as the form below, where $x_i s$ are attribute as independent variables, and β_i are coefficients of x_is. General form of HPM are as follows:

$$Y_i = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_i X_i + \varepsilon_i$$

For text-based descriptions and weighs of those to fit into the HPM, above equation is revised as follows:

$$p_i = \alpha + w_{i1} X_1 + \cdots + w_{ij} X_j + \varepsilon_i$$

where p_i is dependent variable, which in this case is price, $\boldsymbol{W} \in \mathbb{R}^{D \times V}$ be Document Term Matrix collecting the vector space representation of the descriptions.

The whole development process is the process to obtain the hedonic pricing model. Following the construction of hedonic pricing model's equation, topics from LDA are going to be the independent variables, prices are set as dependent variables, and the coefficients will be obtained by TF-IDF matrix, and PLS.

3.2. Data Collection & Preprocessing

3.2.1. Data

Total 10,000 online reviews were scraped from more than 1,000 pages of online reviews from amazon review pages. Each 2,000 from top 5 popular subscription services in OTT(Over-The-Top) video streaming area, which are Netflix, Disney+, Amazon Prime Video, HBO Max, and Hulu. The reviews were sorted by top reviews that got highest number of helpful buttons from customers starting from the most recent ones of January 2023. Also, the reviews were filtered with verified purchase only, with all stars from 1 to 5.

Example of reviews from amazon review of purchased video streaming subscription services are shown in figure1.

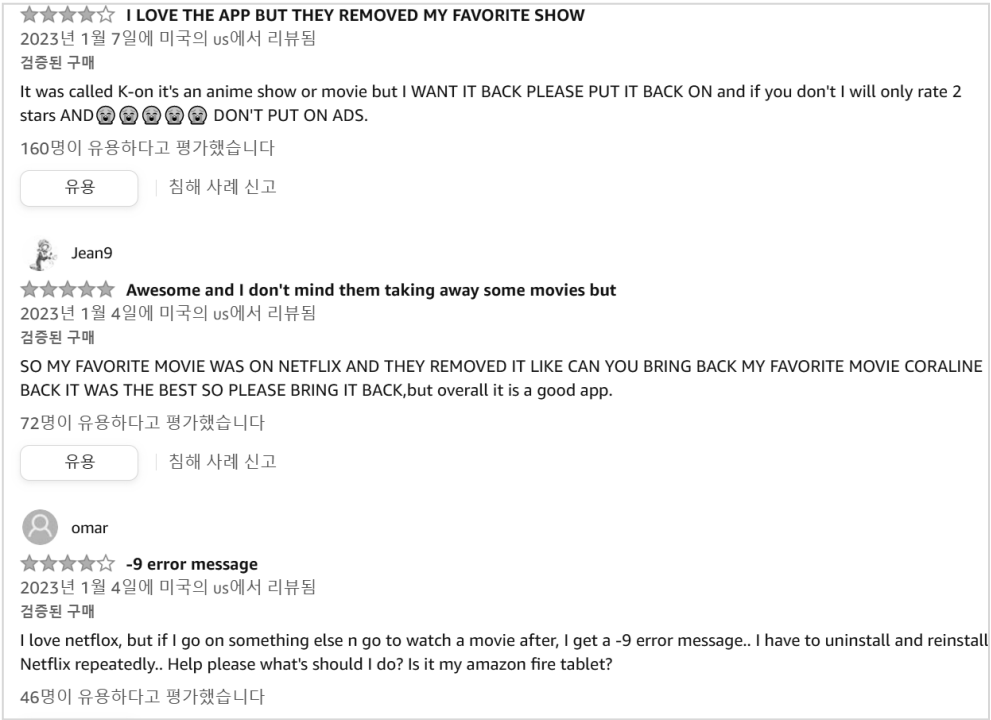


Figure1. Example of review data that has been scraped from amazon website's customer review page.

After preprocessing the reviews, the reviews that contained unverified words were deleted, and the final data of 9,675 reviews were considered. Followings are considered amount from each service. The specific number of each review that were left as useful data after preprocessing are 1,999 reviews of Netflix, 1,960 reviews of Amazon prime video, 1,767 reviews of Disney+, 1,970 reviews of HBO Max, 1,979 reviews of Hulu.

3.2.2. Data scraping & Preprocessing

Online reviews scraped were stored into file and below shows the rare uncleansed scraped reviews that are imported. By preprocessing data into steps of removing stopwords, blanks, punctuations, and then tokenize, go through lemmatization, and then filter nouns, the clean data are imported as right side of figure2 below.

Reviews
I have had this service for awhile and they always have new things being added. Ne
I love having so many great options of shows and movies to watch for a decent pri
Can take up to 3 minutes just to launch the program. Then a long lag time when s
If you love movies, get a subscription. They may not have the one you want, but y
Are your shows on a break? Are you in the mood for an oldie-but-goodie? Date nig
Only had the network for a while it?쇔 great, love the variety of shows.
Me encanta Netflix las mejores series y pel챠culas. La resoluci처n es perfecta.
One of my favorite for viewing my favorite shows and movies
But I need more selections
Heard they plan to raise the rates And if they limit the password sharing does that
Great if you want to watch a movie
Need I say more? If you could only afford one streaming service, this has to be it!
Great streaming application
Favorite platform as far as quality and easy to navigate, fast forward, rewind, etc. W
I use this channel to help fill in extra knowledge in maybe certain situations where
I love Netflix but lately I been having issues with my fire tablet it won't let me log i

unnamed: 0	
1.0	[watch, movie, error, message, reinstall, fire...
2.0	[call, anime, show, movie, rate, star, ad]
3.0	[movie, show, projector]
4.0	[movie, movie, coraline, good]
5.0	[take, cause, cancel, show, thing, show]
...	...
9670.0	[pay, service, ad, time, show, ad, break, watc...
9671.0	[option, drop, cable, issue, time, program, wa...
9672.0	[lord, season]
9673.0	[channel, watch, show, channel, play, empire, ...
9674.0	[stream, issue, cancel, subscription]

9675 rows × 1 columns

Figure2. Web scraped data and preprocessed data

3.2.3. LDA, TF-IDF

1) Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) is a model for text data where documents are considered as mixtures of topics and each topic is defined as a probability distribution over the vocabulary. It is a machine learning tool that can help figure out number of topics that are important among bunch of texts preprocessed as nouns. Below is an image that describes how LDA picks topics and weighs each noun. In this case, we get the weights by term frequency.

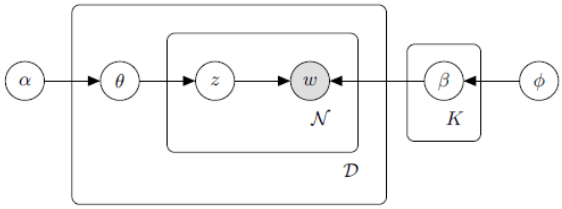


Figure3. The graphical explanation of how topics and weights are driven by LDA process. (Blei et al., 2003)

Figure3 gives graphical explanation of how topics and weights are driven by LDA process. Each node is a random variable and is labeled according to its role in the generative process. Consequently, being mixture of topics, documents are represented as points in the topic simplex. The observed words are w_d, where $w(d,n)$ is the n-th word in document d. (Blei et al., 2003; Crescenzi, 2020).

By using LDA, seeking for 1 aggregated topic from the whole subscription services review can be done. Input of codes to define 1 topic constructed with 30 words and going through 100 times of iteration progressing to define the probability distribution of words to construct the topic were done.

Result of 1 aggregated topic of all OTT services weighed by LDA on 27 words (brand names extracted) was carried out as follows in figure4. This is unsupervised LDA, which has no dependent variable. The words of topics are going to be used as independent variables, and the rest of the process will be regarded to finding out the values of these independent variables.

```
{ 0.039*"movie" + 0.035*"watch" + 0.029*"show" + 0.021*"tv" + 0.019*"video" + 0.016*"stream" +
0.014*"service" + 0.009*"fire" + 0.009*"work" + 0.008*"device" + 0.008*"pay" + 0.008*"content" +
0.008*"download" + 0.008*"quality" + 0.007*"series" + 0.007*"year" + 0.007*"month" + 0.007*"love"
+ 0.007*"issue" + 0.006*"problem" + 0.006*"phone" + 0.006*"tablet" + 0.006*"episode" +
0.006*"day" + 0.006*"selection" + 0.006*"cable" + 0.006*"program" }
```

Figure4. LDA results

In order to apply the review data to Supervised LDA, 9,675 reviews were vectorized and were made into a matrix of weights that show the distribution of words from the above topic in each review sentence. The matrix of weighed reviews will work as the coefficients of the independent variables. This work was done by TF-IDF. Following section is the description of TF-IDF and a brief result of the matrix resulted from the methods is shown in figure5.

2) Term Frequency-Inverse Document Frequency (TF-IDF)

TF-IDF weighting scheme is widely used in the literature on text mining and information retrieval because it puts higher weight on words having high frequency in the document and low frequency in the collection (Crescenzi, 2020). Weights are weighed in the formula as below where tf_{ij} denotes term-frequency, and $\log^{-1}(D/d\,f_j)$ denotes inverse document frequency.

$$w_{ij} = tf_{ij} * \log^{-1}(D/d\,f_j)$$

Figure 5. Partly captured version of matrix obtain by TF-IDF

	movie	watch	show	tv	video	stream	service	fire	work	device	pay	content	downloa	quality	series	year	month	love	issue	problem	phone	tablet	episode	day	selection	cable	program
0	0.180469	0.181091	0	0	0	0	0	0.295018	0	0	0	0	0	0	0	0	0	0	0	0	0	0.327686	0	0	0	0	0
1	0.170346	0	0.186227	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0.232437	0	0.254107	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0.335007	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0.400389	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0.090788	0.182203	0.198504	0	0	0	0	0	0	0	0.150253	0	0	0	0	0	0	0	0	0	0	0	0	0.159192	0	0	0
7	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0.086708	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0.168211	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0.343641	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.632207
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0.161421	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.258417	0	0	0	0	0	0	0
14	0	0.629752	0	0.776796	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.625285
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	0.554171	0.278041	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0.145587	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.420924	0	0	0	0	0	0
20	0.071804	0.144103	0.078498	0.088875	0.06817	0.095383	0	0.195634	0	0.205472	0.118834	0	0	0.197346	0.080562	0	0.082278	0	0	0	0.085653	0.086919	0.044283	0.083936	0.041654	0.043568	0.044033
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
22	0.705889	0.708323	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
24	0	0	0.151377	0	0	0	0	0.226358	0	0	0	0	0	0	0	0	0	0	0.241159	0	0	0.251423	0	0	0	0	0
25	0	0	0	0	0	0	0	0	0	0	0	0.741387	0	0.365571	0	0	0	0	0	0	0	0	0	0	0	0	0
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	0	0	0	0	0	0	0	0	0	0	0.40155	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
30	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
31	0.09892	0.059556	0.108142	0.09795	0.14087	0.052561	0	0	0.031809	0	0	0.033082	0.034809	0.032624	0	0.032571	0	0	0.068912	0.034625	0.035399	0	0	0.03469	0	0	0
32	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
34	0	0	0	0	0	0.245516	0.265109	0	0	0	0	0.309052	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
35	0	0	0.284749	0.322392	0	0	0	0	0.418782	0	0	0	0	0	0	0	0	0.430784	0	0	0	0	0	0	0	0	0
36	0	0	0.203991	0	0	0	0.535299	0	0	0	0	0	0	0	0	0.307202	0	0	0	0	0	0	0	0	0	0	0
37	0.185423	0.372124	0.202709	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
38	0	0	0	0	0	0.324813	0	0	0	0	0	0	0	0.806439	0	0	0	0	0	0	0	0	0	0	0	0	0
39	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
40	0	0	0	0	0	0	0	0	0	0	0	0	0.355578	0	0	0	0	0	0	0	0	0	0	0	0	0	0
41	0.292493	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.512871	0	0	0
42	0	0.229406	0.249931	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
43	0	0.411036	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
44	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.428685	0	0	0	0
45	0.370867	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.613374	0	0	0	0	0	0	0	0	0
46	0.14191	0	0.15514	0	0	0.18851	0	0	0.228166	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
47	0.076197	0.07646	0.083301	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
48	0	0	0.116429	0.131821	0	0	0	0	0	0	0	0	0	0	0.358473	0	0	0	0	0	0	0	0	0	0	0	0
49	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.229141	0	0	0	0	0	0	0	0	0	0	0	0
50	0.103466	0.062294	0.022622	0.230517	0.088407	0.192419	0.029682	0.135312	0.133084	0.106587	0	0	0	0.136495	0	0	0.035567	0	0.07208	0.325943	0.037026	0	0	0	0	0.188339	0

With the specific words of attributes that constructs the topic, the reviews were vectorized as above. Now that we got the weighs, defining values of independent variables can be done. By defining the values, creating a hedonic pricing regression model is available. To do this, dependent variables, in this case the price of each service needs to be applied. For that, supervised LDA is needed, and this can be done by PLS regression algorithm.

3.2.4. Obtain regression model by PLS (Supervised LDA)

Most topic models, such as latent Dirichlet allocation (LDA), are unsupervised: only the words in the documents are modeled. The goal is to infer topics that maximize the likelihood of the collection. The words in the document are obtained by repeatedly choosing a topic assignment from those proportions, then drawing a word from the corresponding topic. (Blei & McAuliffe, 2007).

In supervised latent Dirichlet allocation (sLDA), the key is to add a response variable associated with each document to LDA. The documents and the responses are jointly modeled in order to find latent topics that will best predict the response variables for future unlabeled documents. (Blei & McAuliffe, 2007).

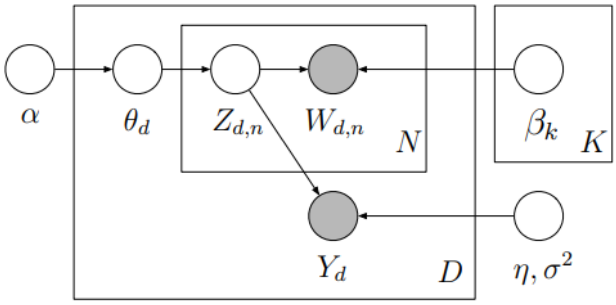


Figure6. Graphical explanation of Supervised LDA
(Blei & McAuliffe, 2007)

In short, supervised LDA is making unsupervised LDA set that did not have dependent variable into a supervised one that has relation to the dependent variable we want to set, which in this case, price. This can be done through using Partial Least Squares (PLS) regression model in machine learning.

A methodology that makes use of response Y_is to infer some traits about the X_is is regarded as supervised. Partial Least Squares (PLS) is a supervised dimensionality reduction technique that seeks

a linear combination of the input variables $\{X_1, X_2, \dots, X_i\}$ to find directions in the data with larger variation that are most correlated with the response Y_is (Crescenzi, 2020; Kim, 2019; Jo, 2021).

The reason that PLS regression should be used in this research is because, PLS is the only supervised dimensionality reduction method that considers correlations between variables (Kim, 2019; Jo, 2021). Considering correlations between variables is extremely important in the text data because the attributes of independent variables were extracted from topic modeling method. Topics are constructed with the words that are most frequently coming out from the customer reviews and those words get related in order to make a topic that gives meaning. Thus, by using PLS as supervised dimensionality reduction technic, the result of left attributes and the weighs become more meaningful to interpretate the whole review data regarding price.

About the use of PLS in this study, PLS originally seeks for variables of coefficients that maximizes covariance of linear combination of X_is and Y_is (Kim, 2019; Jo, 2021). However, in this study, set of coefficients are set with TF−IDF matrix. Thus, we are going to look for values of X_is that maximizes the covariance and correlation of Y_is when the weighs are set.

To make hedonic pricing regression, we'll put PRICE (monthly) of each brand as Y_i s, dependent variable. Insert the weighs (w_{ij}) defined from TF−IDF. To get value of X_is, price value of each word (attributes for hedonic pricing regression), we'll be putting the w values instead of x_is to the equation set for PLS. Below shows how PLS regression is derived (Kim, 2019; Jo, 2021).

$$where\ X = independent\ variables\ (attributes), Y = dependent\ variables\ (prices),$$
$$w = weight\ of\ X, \qquad t = Linear\ combination\ of\ X\ and\ w,$$

$$Maximize\ cov(t, Y) \propto Maximize\ corr(t, Y) var(t)$$
$$where, t = X\omega$$

$$\mathrm{cov}(t, Y) = \frac{\mathrm{cov}(t, Y)}{\sqrt{\mathrm{var}(t)}\sqrt{\mathrm{var}(Y)}}\sqrt{\mathrm{var}(t)}\sqrt{\mathrm{var}(Y)}$$

$$= corr(t, Y)\sqrt{\mathrm{var}(t)}\sqrt{\mathrm{var}(Y)}$$

What we want to obtain are $'X'$ s that maximizes $corr(t, Y) * Var(t)$. This will be defined through machine learning algorithm by keep predicting the values that best fit while satisfying the prediction of Ys considering real price. The distribution of Y predictions after obtaining X came out as shown graph below and the values of predicted results were obtained. The price range of 5 subscription services are about 8~15 dollars, and the values show about the average of those prices. The result of predicted value of each review from PLS regression and the graph that show the distribution is shown in figure 7 below.

```
[[13.64946193]
 [12.21606957]
 [11.76246241]
 [11.2876009]
 [12.59446521
 [11.8032099]
 [12.16546466]
 [15.81675692]
     ...
 [4.837917203]
 [10.93014495]
 [11.12336564]
 [13.80649033]
 [11.98883091]
 [10.96623471]]
```

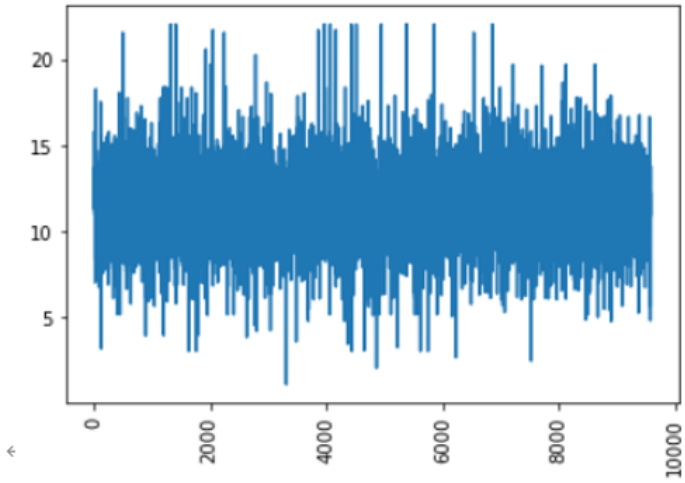


Figure 7. Predicted price of each review by PLS regression and the distribution graph

3.2.5. Define values of attributes & Hedonic regression model

After going through several trials, using PLS, the value of X_is with the maximum covariance obtained from PLS were defined. The organized result is shown in table3 as follows.

movie	**23.11346**	device	**19.69003**	issue	**38.71348**
watch	**13.49595**	pay	**27.53894**	problem	**5.07029**
show	**5.299249**	content	**17.04041**	phone	**10.90359**
tv	**23.26481**	download	**34.45155**	tablet	**13.86824**
video	**13.48982**	quality	**28.76188**	episode	**13.14426**
stream	**13.39209**	series	**9.03454**	day	**20.15484**
service	**30.39508**	year	**11.80395**	selection	**8.203285**
fire	**7.853257**	month	**15.41468**	cable	**32.57922**
work	**29.86588**	love	**26.36383**	program	**6.798889**

Table3. Result of deriving values of independent variables.

Using the values that came out, now hedonic pricing regression is defined:

$$p_{all\ five\ ott\ services} = 23.11346w_1 + 13.49595w_2 + \cdots + 32.57922w_{26} + 6.79889w_{27}$$

With reduced dimensions from PLS, putting above value into coefficients that came as output of PLS dimensionality reduction. the price was finally calculated as ***$ 12.7028.***

Now, using this whole process, any reviews can be turned into data in order to measure prices considering values that customers take important. Application to each OTT service is going to be done and the implications that we can get from predicted prices and topics came out are discussed in the next section.

Chapter 4. Application

4.1. Pricing each OTT service with LDA and PLS

For each brand, the same process has been done. To show the process of applying the whole analysis introduced in the development process section, smaller data of 2,500 reviews of OTT services were separately analyzed as application and comparison of the results are done to give implications.

Thus, 2,500 reviews of each 500 reviews from top5 OTT subscription services were used as data to predict price in the view of customers. To gain more differentiated topics to compare each topic to one another, 5 separate topics from LDA were obtained. The 5 topics from LDA resulted in explanation of each brand since the data were constructed exactly in the same number of reviews.

Example of how LDA result came out containing topics of each brand is shown in figure 8. It describes the result of unsupervised LDA before getting the dimensionality shrinkage.

```
 '0.041*"disney" + 0.036*"stream" + 0.036*"service" + 0.031*"movie" + 0.023*"tv" + 0.
021*"watch" + 0.019*"content" + 0.016*"show" + 0.016*"price" + 0.016*"month" + 0.01
4*"pay" + 0.013*"year" + 0.013*"cable" + 0.013*"quality" + 0.011*"program" + 0.010*"s
ubscription" + 0.010*"love" + 0.008*"war" + 0.007*"issue" + 0.006*"channel" + 0.006*
"offer" + 0.006*"option" + 0.006*"day" + 0.005*"cost" + 0.005*"series" + 0.005*"comp
any" + 0.005*"title" + 0.005*"picture" + 0.005*"deal" + 0.005*"trial"'),↵

 (1,↵

 '0.062*"video" + 0.058*"amazon" + 0.025*"work" + 0.024*"tablet" + 0.023*"device" +
0.022*"download" + 0.021*"fire" + 0.018*"phone" + 0.015*"play" + 0.012*"store" + 0.0
11*"watch" + 0.011*"tv" + 0.011*"issue" + 0.009*"review" + 0.008*"problem" + 0.008*"
stream" + 0.008*"google" + 0.007*"version" + 0.007*"apps" + 0.007*"access" + 0.007*"
service" + 0.007*"purchase" + 0.006*"try" + 0.006*"account" + 0.006*"customer" + 0.0
06*"install" + 0.005*"movie" + 0.005*"stick" + 0.005*"support" + 0.005*"member"'),↵

 (2,↵

 '0.076*"movie" + 0.049*"watch" + 0.023*"hbo" + 0.018*"show" + 0.015*"series" + 0.0
13*"tv" + 0.012*"video" + 0.011*"selection" + 0.010*"max" + 0.010*"stream" + 0.008*"f
amily" + 0.008*"program" + 0.007*"screen" + 0.007*"year" + 0.007*"quality" + 0.007*
"interface" + 0.006*"problem" + 0.006*"search" + 0.006*"love" + 0.006*"pay" + 0.006
```

Figure 8. Example of LDA result, 5 topics of each subscription services

As explained in the process of defining aggregated version of pricing subscription serviced, just by conducting LDA, the output model is considered unsupervised. These results are not able to explain the relationship between price, since they are only showing the probability distribution of words in the topics based on the frequency. Thus, using TF−IDF, and PLS regression, the result of these extracted topics will be made into supervised output that seeks for better coefficients and that considers the relationship between the attributes and dependent variables of price.

However, the result of unsupervised LDA still gives the implication to the corporates by giving insights of which words customers are more frequently putting weight on. To see the organized distribution of the words in the topics, visualization of the relevant topics and words can be shown as below in figure9.

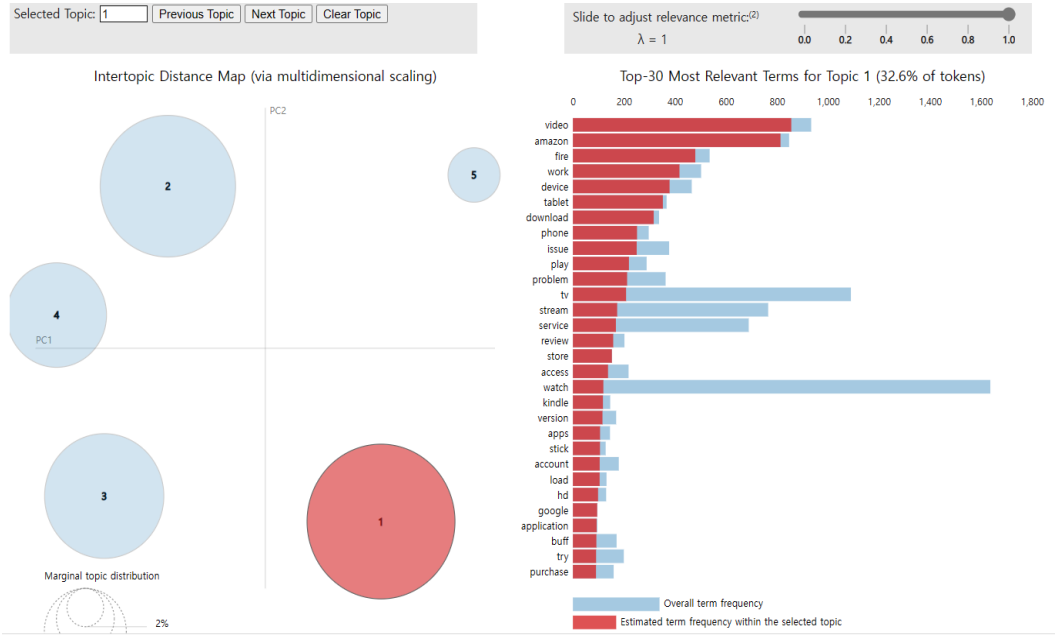


Figure 9. An example of visualized description of LDA results and the frequency relevant words

Using the same process of TF-IDF and PLS regression described in the aggregated version in Chapter 3, pricing results of each subscription service can be defined. Again, for implication, 2,500 data were applied. Characteristic of collected data are same as the ones of 10,000 data on aggregate pricing version.

The organized chart of currently offered prices of each OTT services and the result of PLS regression is shown in table4. More results obtained from TF-IDF, PLS coefficient result, and the trials of other methods are shown in Appendix A, B, C, and D in corresponding order.

Table4. Predicted cost of each subscription service

	Average current price	**Price obtained from PLS**	Dim.red & minimize differences
Netflix Basic with ads*: $6.99/month Basic: $9.99/month Standard: $15.49/month Premium: $19.99/month	$ 13.12	**$ 11.73945**	$ 20.31992
Disney+ $11/month or $110/year	$ 11	**$ 11.57668**	$ 23.86114
Amazon Prime Video $12.99/month or $99/year	$ 12.99	**$ 12.2091**	$ 21.7235
HBO Max $14.99/month or $149.99/year	$ 14.99	**$ 11.80891**	$ 11.53327
Hulu $8/month or $80/year	$ 8	**$ 12.09456**	$ 20.42533

The average price that OTT subscription services offer monthly were about $8 to $15. Whereas the distribution of price obtained from PLS were in smaller differences among all OTT subscription services that were considered. The predicted prices in the view customers were about $11 to $12.

To give more insights on not considering the way of supervised LDA (PLS), trial on dimensionality reduction with only minimization on differences between dependent variable and the linear combination of independent variable along with weighs are considered. The results of the trial were set on the right side of PLS regression result in table 4. Dimensionality shrinkage was done by a machine learning algorithm that does not consider the correlation between data and minimized difference between weight applied attributes and price by OLS (Ordinary Least Squares) regression, which is a simple linear regression method that generally minimizes the differences between the attributes. We can compare the ones with PLS and obtain the answer for the second research question. The results and graphs of these trials are shown in Appendix C and D.

The trial tells that only by minimizing differences without using supervised LDA gives a price range that may not be actually applied both in the customers view, and also in the corporate's view of pricing. Covariances between weighs should also be considered in order to get more predictable price compared to the current price.

Thus, more than shrinking dimensionality or finding a simple way, what was more important in prediction better was to consider correlations and variances between attributes and coefficients that weigh attributes from LDA, TF-IDF by the right supervised method, PLS. By considering variances with regards to the correlation, which is covariance, LDA, TF-IDF results can be more accurately used to predict the price when using the text data.

Further interpretations and implications are discussed in the result and implication part.

4.2. Results and Implications

Implications that can be considered are about the attributes that came out differently among the brands related to price. As you can see in the result of each brand's LDA output, taking out the result of 5 different topics from LDA contains topic regarded to each brand since similar amount of data in each brand are contained. Looking at the differentiated topics, the weighs, and topics that customers considered important comes out differently even though they are all in the same industry of OTT subscription services.

Table 5 in the next page shows the gap between current prices of services and hedonic pricing model results of the services calculated in this research. The table also shows how the attributes were set to give implications on where the corporates should focus on.

The blue colored words are the ones that imply the positive effect on pricing in the customers review, and the red colored words are those that may have caused negative effect. Red colored words were mostly about the issues and prices that needed to be addressed whereas the blue ones were explaining more of the characteristics of the OTT subscription brands.

Further explanations on specific parts of this table and the implications for each OTT services are discussed on 4.2.2. section. Before moving on to that, implications regarding customer review-based pricing will be discussed considering the concept of 'willingness to pay' and other situational factors in section 4.2.1. At last, implications regarding methodologies are discussed before getting onto the conclusion part in section 4.2.3.

Table5. Organized result of pricing each OTT subscription services and words from topics for implications to be discussed.

Current price	$ 13.12	$ 11	$ 12.99	$ 14.99	$ 8
HPM result	$ 11.73945▼	$ 11.57668▲	$ 12.2091▼	$ 11.80891▼	$ 12.09456▲
	Netflix	**Disney+**	**Amazon Prime Video**	**HBO Max**	**Hulu**
1	video	song	show	movie	game
2	tablet	service	season	series	repeat
3	device	stream	episode	selection	football
4	download	content	movie	family	news
5	phone	movie	ad	show	impression
6	playstore	program	problem	screen	mood
7	issue	quality	search	interface	network
8	version	offer	quality	child	series
9	account	marvel	service	program	character
10	problem	package	option	kid	channel
11	member	option	stream	video	power
12	installation	picture	series	stream	story
13	update	issue	storage	quality	change
14	access	series	device	option	que
15	security	pixar	access	caption	click
16	login	discount	change	song	problem

4.2.1. Implications regarding customer review-based pricing

The results carried out are the predicted prices in the view of customers by taking the customer reviews as data. For Netflix, Amazon Prime Video, and HBO Max, in the view of customers, it came out that they were overpriced. For Disney+ and Hulu, the result showed that current prices were underpriced in the view of customers.

These results are predicted price values obtained by reflecting customers' opinions and considering that, the prices can act as a concept similar to customers' 'willingness to pay.' 'Willingness to pay' is the maximum price a customer is willing to pay for a product or service (Stobierski, 2020). Which means the price predicted is the 'maximum limit' of what customers think about a product or service.

However, pricing is something that companies can choose in order to get the maximum benefit out of it and in fact, corporates do not always take the same price of willingness to pay. Considering this, what is more important to be considered in the implication part is that how much customer is willing to pay with 'what and how much values the customers care when deciding to purchase products or services.' This comes as output of attributes and weighs that are considered in the process of predicting the price.

Thus, by comparing the value of each service that customers think and the actual current price set, this research can give insights to the subscription service brands. This can be done by looking at the prices with the attributes. Considering the prices with the attributes of that has been considered after dimensionality reduction, there are significant implications for the brands to take apart. This can be done by looking at the words of topics that may have caused positive/negative effect to the result of price.

4.2.2. Insights from the prices and attributes carried out from the process of pricing 5 OTT subscription services

Now, by looking into specific outputs of attributes related to price values that came out from data of customer reviews, further implications that each OTT service should consider for better services can be derived. Moreover, situational factors other than the reviews are considered to give descriptions to the prices that came out since whether they are overpriced or not can be affected by numerous factors along with customer's perspective of pricing.

For Netflix, the prices that were adjusted by customer reviews came out about $2 less that the average service price they have been offering. Considering the fact that most of the customers are using $6.99~9.99 price set from Netflix without considering standard and premium version users, this may seem that Netflix has been underpriced. However, considering all the premium customers, looking at the attributes of words from topics explain why the customers are feeling as the service has been overpriced.

The attributes contained a lot of problems and issues that Netflix has faced. What customers took a part as considerable was about playstore issue, account problem, and issues regarding access security. This tells why customers are willing to pay lesser amount than the price they offer in average. The issues and problems need to be considered in order to keep the customers and maintain the value of Netflix as one of the top OTT subscription services.

For Disney+, the price offered came out as a little bit underpriced, but still reasonable to customers. Regarding the attributes, brand contents from Disney has been taken apart in great frequency. Songs from the services, marvel packages they offer, Pixar series discount were considerable in the view of customers that effect higher willingness to pay. The result show that picture issue was considered from customers, which gives insight to Disney+ on getting higher willingness to pay from the customers. Disney+ is one of the latest introduced video streaming services,

and they have promoted discount services in order to catch up with existing subscription services. These strategies of pricing may have caused Disney+ to be considered as underpriced service.

Amazon Prime Video also had good fitness on price. Customers mostly mentioned season episodes, storage of devices, and search quality which gives insights on where the service should focus more on. The result showed that customers' have been struggling with the ad problem, and issues regarding change of the access.

HBO Max had keywords usually related to family and kids. Customers mentioned family show, screen interface, child program, kid video, stream quality and songs in great frequency. Issues and problems were not considered in the most important topic about HBO Max. Regarding price, HBO Max had about $3 gap in between customers considerable price and price they offer. It seems they are overpriced.

Seeking for the reason of this, factors besides text can significantly be considered. By a survey conducted by 'Whip Media's 2022 Streaming Satisfaction Report', HBO Max has won the first on value satisfaction area among all video streaming services. The result of this was interpreted that HBO Max has much more old films and series, and customers seeking for those value HBO Max as one of the most necessary services. Thus, even though the price they offer are overpriced than the predicted willingness to pay price, fans of those contents keep seeking HBO Max.

Hulu had the highest price on willingness to pay from customers with $12. Gap was about $4 with their current price. Brand power may have taken apart on underpricing. If the name value of subscription services is sticky, Hulu may be taking underpricing strategies in order to compete with other OTT subscription services. About attributes, the most frequent topic that has been considered was about game repeat on footballs. The result of LDA shows that customers are more willing to pay on the football shows and news they are offering with channel power also

considered. Click problem that has been mentioned could be taken apart in order to service better quality to the customers.

4.2.3. Implications regarding methodologies

Regarding the second research question, implications regarding methodologies can be considered. PLS regression is well known for supervised LDA method considering the correlation and variance between each attribute, weights, and dependent variables. Use of other methods that does not consider correlation and variances, such as simple linear regression with OLS (Ordinary Least Squares) were tried, but predictable prices came out as less accurate. To gain right value of attributes, accurate price prediction is needed at first.

After getting the higher accuracy on the values of attributes, seeking better coefficients for those values are available to set to define price on customers' view.

Thus, when using the text-based data, it was examined that considering the correlation and variance is needed to get more predictable price.

Chapter 5. Conclusion

Archak et al. (2011) claims that working with descriptions shall not be recommended since the contained texts are too static and give little emphasis on characteristics of the goods. However, findings here provided another empirical evidence that this is not necessarily true. With the proposed framework that has been designed using the powerful techniques of 'Machine Learning, Data Mining, and correctly fitted Statistical Methods' to predict the weighs of important attributed from online customer review, predicting an appropriate price range for customers was available based on Hedonic Pricing Model.

In fact, the set of attributes and the weighs that were obtained from the descriptions can improve the performance of the model if with the right fitted regression models. As the time goes by and more data are collected, with the supervised machine learning technics, descriptions as customer reviews will give more accurate results on determining price value on the view of customers and help prediction the behaviors of the customers that corporates can get insights.

In addition, it was found that more than shrinking dimensionality or finding a simple way, what was more important in prediction better was to consider correlations and variances between attributes and coefficients that weigh attributes from LDA, TF-IDF by supervised method, PLS. The framework introduced in this study can be applied to various industries using online review systems that has intention to reflect customer's view on price of their services. With the process of applying customer reviews to the price, both customer and corporate can be considered in seeking reasonable value of the services.

For future studies as extension, it seems reflecting brand effect would contribute on better pricing. On the interpretation part of the attributes that came out, there seemed more factors that

affect the price besides the words given by the reviews, and the most effective part may be the brand power. Thus, more works on putting the brand power and name value of the subscription services into account would give more predictability to price on the view of customers.

Bibliography

Archak, N., Ghose, A., & Ipeirotis, P. (2011). Deriving the pricing power of product features by mining consumer reviews. *Management Science, 57*(8), 1485-1509. 10.1287/mnsc.1110.1370.

Bacon, D.R., Besharat, A., Parsa, H.G. & Smith, S.J. (2016). Revenue management, hedonic pricing models and the effects of operational attributes. *International Journal of Revenue Management*, *9*(2-3), 147-164.

Bala, R. (2012). Pricing online subscription services under competition. *Journal of Revenue and Pricing Management, 11*(3), 258-273.

Berry, M., & Castellanos, M. (2004). Survey of text mining. *Computing Reviews, 45*(9), 548.

Berry, M., & Kogan, J. (2010). Text mining: Applications and theory. *john wiley & sons.*

Blei, D., Gerrish, S., & Randanath, R. (2014). Black box variational inference. *Seventeenth International Conference on Artificial Intelligence and Statistics.*

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*(1), 993-1022

Blei, D. M., McAuliffe., J. D. (2007). Supervised Topic Models. *Advances in Neural Information Processing Systems (NIPS), 20.*

Brynjolfsson, E. & Kemerer, C. F. (1996). Network externalities in microcomputer software: An econometric analysis of the spreadsheet market. *Management Science*, *42*(12), 1627-1647.

Chen, T., Fenyo, K., Yang, S., & Zhang, J. (2018). Thinking inside the subscription box: new research on e-commerce consumers. *McKinsey & Company.*

Chen, C., & Rothschild, R. (2010). An Application of Hedonic Pricing Analysis to the Case of Hotel Rooms in Taipei. *Tourism Economics, 16*. 10.5367/000000010792278310.

Chin, T. A., Govindasamy, U., Sulaiman, Z. & Tat, H. H. (2016).

Factors affecting the consumers proneness to buy 99-ends products. *Advanced Science Letters, 22*(12), 3991-3994.

Court, A. (1939). Hedonic Price Indexes. *The Dynamics of Automobile Demand,* 99-119.

Crescenzi, F. (2020). Text Based Pricing Modelling: an application to the fashion industry. *University of Bologna*. DOI: 10.6092/unibo/amsdottorato/9524.

Danaher, P. (2002). Optimal Pricing of New Subscription Services: Analysis of a Market Experiment. *Institute for Operations Research and the Management Sciences, Marketing Sciences, 21*(2), 119-138. http://www.jstor.org/stable/1558063.

Feenstra, R. & Shapiro, M. (2007). Scanner data and price indexes*. University of Chicago Press, 64*.

Fruchter, G. & Sigue, S. (2013). Dynamic pricing for subscription services. *Journal of Economic Dynamics & Control, 37*, 2180-2194. http://dx.doi.org/10.1016/j.jedc.2013.05.003.

George, J. & Gupta, M. (2016). Toward the development of a big data analytics capability. *Information & Management, 53*(8), 1049-1064.

Gibbs, C., Guttentag, D., Gretzel, U., Morton, J. & Goodwill, A. (2018). Pricing in the sharing economy: a hedonic pricing model applied to Airbnb listings, *Journal of Travel & Tourism Marketing, 35*(1), 46-56, 10.1080/10548408.2017.1308292.

Gupta, R., & Pathak, C. (2014). A Machine Learning Framework for Predicting Purchase by Online Customers based on Dynamic Pricing*. Procedia Computer Science, 36,* 599-605. https://doi.org/10.1016/j.procs.2014.09.060.

Haatanen, H. (2022). Customer segmentation with subscription based online media customers. *University of Helsinki.*

Hernández, C. & Rosales, I. (2021). Building Models to Predict Real Estate List Prices using Ensemble Machine Learning Algorithms. *Proceedings of the International Conference on Industrial Engineering and Operations Management,* 2481-2490.

Ivanov, S. & Piddubna, K. (2016). Analysis of prices of accommodation establishments in Kiev: Determinants, dynamics and

parity. *International Journal of Revenue Management*. *9*(4), 221-251.

Jo, Y. H. (2021, Mach 14). *[ML] Partial Least Squares*. TISTORY. https://techblog-history-younghunjo1.tistory.com/174

Kim, S. B. (2019, June 10). *[Core Machine Learning] Partial Least Squares (PLS).* [Video]. YouTube. https://youtu.be/OCprdWfgBkc

Lancaster K. (1966). A new approach to consumer theory. *Journal of Political Economy, 74*(2), 132-157.

Liang, J. & Yuan, C. (2021). Data Price Determinants Based on a Hedonic Pricing Model. *Big Data Research, 25*. https://doi.org/10.1016/j.bdr.2021.100249

Liu, Y. & Wu, Y. X. (2009). Analysis of residential product's value based on structural equation model and hedonic price theory. *International Conference on Management Science and Engineering - 16th Annual Conference Proceedings (ICMSE),* 5317673, 1950-1956

Magno, F. & Cassia, F., Ugolini, M.M. (2018). Accommodation prices on Airbnb: effects of host experience and market demand. *TQM Journal*, *30*(5), 608-620.

Mesak, H. & Darrat, F. (2002). Optimal pricing of new subscriber services under interdependent adoption processes. *Journal of Service Research, 5,* 140-154.

Miao, R., Guo, P., Huang, W. & Zhang, B. (2022). Profit model for electric vehicle rental service: Sensitive analysis and differential pricing strategy. *Energy,* 249. https://doi.org/10.1016/j.energy.2022.123736

Nagaraj, S., Singh, S. & Yasa, V. (2021). Factors affecting consumers' willingness to subscribe to over-the-top(OTT) video streaming services in India. *Technology in Society, 65.* https://doi.org/10.1016/j.techsoc.2021.101534.

Papatheodorou, A., Lei, Z., & Apostolakis, A. (2012). Hedonic Price Analysis. *Handbook of Research Methods in Tourism: Quantitative and Qualitative Approaches*, 170-182. 10.4337/9781781001288.00015

Punj, G. (2013). The relationship between consumer characteristics and willingness to pay for general online content: Implications for content providers considering subscription-based business models. *Marketing Letters, 26*(2), 175-186, DOI 10.1007/s11002-013-9273-y

Rosen, S. (1974). Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of political economy, 82*(1), 34-55.

Samanta, S., Woods, J. & Ghanbari, M. (2007). Impact of price on mobile subscription and revenue. *Journal of Revenue and Pricing Management, 7*(4), 370-383.

Stevens, D. (2014). Predicting Real Estate Price Using Text Mining: Automated Real Estate Description Analysis. *Tilburg University School of Humanities.*

Steyvers, M. & Griffiths, T. (2007). Probabilistic topic models. *Handbook of latent semantic analysis, 427*(7), 424-440.

Stobierski, T. (2020). Willingness To Pay: What it is & How to calculate. Business Insights. *Harvard Business School Online,* URL: https://online.hbs.edu/blog/post/willingness-to-pay.

Tan, X. & Xiao, Y. (2021). The effect of online reviews on addressing endogeneity in discrete choice models. *Data Science and Management*, *2*, 1-11.

Tibshirani, R., Johnstone, I., Hastie, T. & Efron, B. (2004). Least angle regression. *The Annals of Statistics, 32*(2), 407-499. 10.1214/009053604000000067.

Tong Z. & Zhang, H. (2016). A Text Mining Research Based on LDA Topic Modelling. *Computer Science & Information Technology, 6*, 201-210. 10.5121/csit.2016.60616.

Waugh, F. (1928). Quality Factors Influencing Vegetable Prices. *American Journal of Agricultural Economics, 10*(2), 185-196. https://doi.org/10.2307/1230278.

Weiss, S., Indurkhya, N., & Zhang, T. (2015). Fundamentals of predictive text mining. *Springer.*

Wang, W., Tong, G. J. & Zhang, H. R. (2010). Empirical analysis on the housing price in Harbin City based on hedonic model.

International Conference on Management Science and Engineering (ICMSE), 5720005, 1659-1664.

Wang, Y. & Zhou, Y. (2008). A study of mechanism of urban residential segregation based on housing consumer preference. *International Conference on Management Science and Engineering 15th Annual Conference Proceedings (ICMSE),* 4669136, 1713-1719.

Zhao, Y. & Liu, X. J. (2010). Hedonic price study on urban housing: The case of Shijiazhuang city. *International Conference on Management and Service Science (MASS),* 5577066.

Appendices

Appendix A

Sample of TF-IDF result conducted for the application part

watch	movie	show	video	tv	stream	service	work	device	pay	fire	year	download	content	quality	tablet	month	episode	issue	problem	series
0.21045	0.43496	0.23454	0	0.26351	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.35453
0	0	0.40741	0	0.22887	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.06315	0.10876	0.11729	0.14697	0.10542	0.05423	0	0.03264	0	0	0	0.03264	0.03688	0.03437	0.03394	0	0	0	0.07014	0.03587	0
0	0.42633	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0.18212	0	0	0	0.227	0	0	0	0	0	0.27327	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.66854
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.07	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.11928	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.69965	0
0	0	0.28386	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.06447	0.11104	0.02395	0.09003	0.24217	0.19377	0.03006	0.13329	0.10563	0	0.13906	0	0	0	0.13859	0	0.03546	0	0.07161	0.32955	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0.43955	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0.12794	0	0	0	0	0	0.07679	0	0	0	0.0384	0	0	0	0	0.08172	0	0	0.04219	0
0	0	0	0	0	0	0	0	0	0	0	0.40167	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0.29044	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.16357	0.33807	0	0	0.20481	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0.0855	0	0	0	0	0	0	0.1981	0	0	0	0	0	0	0	0	0.20865	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.02932	0.1515	0.16338	0.04094	0.03671	0.2266	0.12304	0.04547	0.09608	0	0	0.04547	0	0.04787	0	0	0.14514	0	0.0977	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.23901	0	0.26637	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.19183	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0.24392	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.4132	0	0	0
0	0	0.17715	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.02687	0	0.02994	0.11256	0	0	0.15033	0	0.35217	0	0.04347	0	0.23537	0.08773	0	0.04657	0.04433	0	0.08953	0.13734	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0.3171	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Appendix B

Result of coefficients brought by Y prediction and dimensionality reduction with PLS algorithm conducted for the application part

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
0	-0.5572	1.37513	1.39475	3.32253	-0.9089	-2.1466	-3.5782	0.89629	-0.81	-1.4888	0.95667	-1.251	0.85022	0.08515	0.99077	-2.7142	1.6904
1	0.00952	0.35288	-0.0863	-0.4963	0.10341	0.6854	0.9453	-0.1387	0.00859	-0.39	-0.7575	0.09339	-0.0668	-0.2145	-0.3702	0.25986	0.1386
2	-0.0381	0.30854	-0.1093	-0.4617	0.00955	0.84601	1.31089	-0.2154	-0.0058	-0.3162	-0.8474	0.03369	-0.0498	-0.3862	-0.4801	0.46152	0.15034
3	0.82512	-0.1876	-0.1552	-0.2693	-0.2465	0.33084	0.07557	0.41902	-0.0354	-0.4886	-0.6808	0.01434	-0.0347	0.00703	-0.2319	-0.1844	0.33826
4	-0.8895	0.93691	-0.0145	-0.7391	0.47139	1.09804	1.95306	-0.7603	0.0544	-0.2707	-0.8554	0.17052	-0.0993	-0.4834	-0.5385	0.77689	-0.0779
5	0.14021	0.47454	-0.0233	-0.5912	0.36091	0.24477	-0.0577	0.07178	0.04806	-0.5923	-0.5109	0.25718	-0.1137	0.25643	-0.0687	-0.2934	0.10637
6	-1.3677	0.48032	0.28938	-1.9979	-0.5544	-0.2941	-0.1225	-0.9842	0.96422	1.35364	0.52092	-2.0073	1.62224	-0.6821	0.33233	0.76301	-0.2211
7	2.18468	-1.5021	-0.4171	0.36959	-1.4524	0.50169	0.3401	1.1083	-0.201	-0.2829	-1.018	-0.4677	0.12211	-0.488	-0.5558	0.03213	0.79857
8	0.31748	0.30315	-0.0574	-0.5079	0.20369	0.26705	-0.0232	0.16166	0.02647	-0.5654	-0.5548	0.19432	-0.0933	0.19188	-0.111	-0.2652	0.16639
9	0.14021	0.47454	-0.0233	-0.5912	0.36091	0.24477	-0.0577	0.07178	0.04806	-0.5923	-0.5109	0.25718	-0.1137	0.25643	-0.0687	-0.2934	0.10637
10	0.22522	0.40779	0.17276	-0.6951	0.01835	-0.6985	-1.1363	0.67873	-0.3818	-0.043	-0.466	0.0647	-0.4176	0.62974	-0.1288	0.29814	-0.1706
11	1.83522	-1.2913	0.01117	1.38603	-2.0037	0.00276	0.43304	-0.0853	-0.4153	-0.4505	-2.2089	0.07138	2.263	0.96122	2.89929	-0.8996	-1.7885
12	0.14021	0.47454	-0.0233	-0.5912	0.36091	0.24477	-0.0577	0.07178	0.04806	-0.5923	-0.5109	0.25718	-0.1137	0.25643	-0.0687	-0.2934	0.10637
13	-0.2249	-1.4263	-1.4226	1.18694	-1.1767	-0.3374	-0.169	0.33318	-1.2353	0.76452	-1.3977	-0.3447	0.40969	0.38891	-1.0554	-0.4537	0.25306
14	3.09466	-3.212	-5.99	-7.2796	6.19899	1.41205	-1.8252	-1.5925	0.41917	8.86017	-1.9795	4.15944	-0.2508	1.95829	1.82959	0.84475	1.4088
15	1.12179	-0.6002	0.14465	1.03783	-1.3615	-0.0819	0.29247	-0.4359	-0.3256	-0.5573	-2.0179	0.31978	2.15683	1.20652	3.03233	-1.0036	-2.0031
16	0.14021	0.47454	-0.0233	-0.5912	0.36091	0.24477	-0.0577	0.07178	0.04806	-0.5923	-0.5109	0.25718	-0.1137	0.25643	-0.0687	-0.2934	0.10637
17	0.14021	0.47454	-0.0233	-0.5912	0.36091	0.24477	-0.0577	0.07178	0.04806	-0.5923	-0.5109	0.25718	-0.1137	0.25643	-0.0687	-0.2934	0.10637
18	1.41371	-0.7312	0.08252	-0.2305	-1.2102	-1.1719	-1.6201	1.64944	-0.8006	0.48712	-0.7178	-0.4627	-0.4853	0.46094	-0.438	0.86479	0.03219
19	-0.9161	-2.3093	-1.2501	-1.1937	3.76904	0.08996	-0.0755	-1.3045	-0.273	-1.6939	0.99051	-2.425	3.394	3.12617	-2.4631	6.0404	-4.1627
20	-0.6252	-1.6327	0.70562	1.02492	0.31797	-0.9343	-1.4887	-1.198	0.33331	-0.8121	3.27417	-0.3908	2.75804	-0.4044	-0.0534	1.371	-1.2872
21	-3.6203	-0.877	-2.6314	-0.5878	-2.4739	-2.6203	-1.7859	-2.2777	4.54119	-0.559	-0.1959	0.50355	-3.0494	1.28309	-3.1232	-2.6939	3.11525
22	1.94135	-1.2018	0.52426	-0.3505	-2.3616	-3.5457	-4.3187	3.35928	-1.938	1.93912	-0.6799	-1.065	-1.2268	1.30293	-0.6656	2.42736	-0.5712
23	0.14021	0.47454	-0.0233	-0.5912	0.36091	0.24477	-0.0577	0.07178	0.04806	-0.5923	-0.5109	0.25718	-0.1137	0.25643	-0.0687	-0.2934	0.10637
24	-2.0946	1.71579	-0.3965	2.22206	-1.0174	-1.7525	-1.4114	0.1037	-0.3228	-2.1738	1.42818	-0.8864	0.62569	-1.3975	2.18208	-2.349	0.49594
25	1.08479	-3.3736	-1.1401	-4.1953	4.70337	-0.7917	1.07799	4.54219	-2.8151	-1.2498	4.82089	-2.4299	1.65901	-1.8464	0.21735	-3.7985	1.02944
26	0.14021	0.47454	-0.0233	-0.5912	0.36091	0.24477	-0.0577	0.07178	0.04806	-0.5923	-0.5109	0.25718	-0.1137	0.25643	-0.0687	-0.2934	0.10637
27	-3.2294	0.43901	1.05636	-3.6453	-0.0703	1.36999	-0.9004	-3.013	4.40039	2.65356	2.01571	-3.5122	4.58462	-1.8677	1.06	-0.6437	0.04931
28	0.14021	0.47454	-0.0233	-0.5912	0.36091	0.24477	-0.0577	0.07178	0.04806	-0.5923	-0.5109	0.25718	-0.1137	0.25643	-0.0687	-0.2934	0.10637
29	0.14021	0.47454	-0.0233	-0.5912	0.36091	0.24477	-0.0577	0.07178	0.04806	-0.5923	-0.5109	0.25718	-0.1137	0.25643	-0.0687	-0.2934	0.10637
30	0.14021	0.47454	-0.0233	-0.5912	0.36091	0.24477	-0.0577	0.07178	0.04806	-0.5923	-0.5109	0.25718	-0.1137	0.25643	-0.0687	-0.2934	0.10637

Appendix C

Result of trials on getting values of independent variables using OLS (Ordinary Least Squares) regression

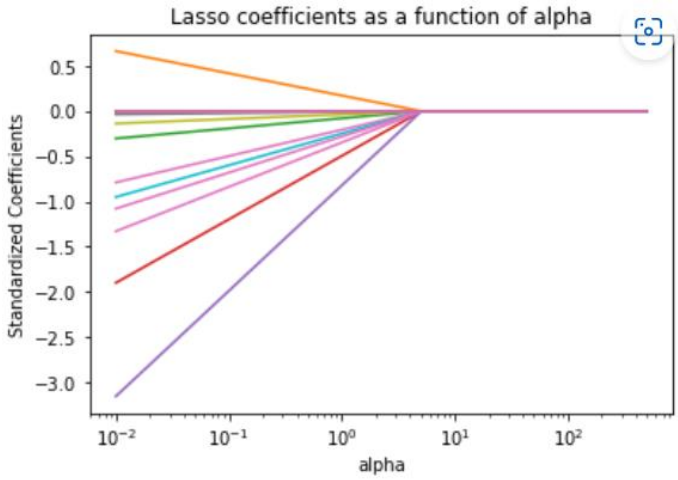
attributes	movie	watch	show	tv	video	stream	service	fire	work
OLS coef	16.9296	14.8160	11.3209	6.8240	9.4899	12.5808	10.8588	14.5468	15.4635
std.err	0.5420	0.6610	0.6110	0.8150	0.9100	0.8590	0.9680	1.2180	1.0650

attributes	device	pay	content	download	quality	series	year	month	love
OLS coef	8.3271	9.0568	18.4437	12.2326	11.1474	13.2816	12.9133	11.3989	11.8292
std.err	1.0650	1.1470	1.0190	1.0620	0.8290	1.0830	1.2920	1.2140	0.6030

attributes	issue	problem	phone	tablet	episode	day	selection	cable	program
OLS coef	11.1542	12.9692	9.8094	5.0343	9.9272	10.4202	7.6724	6.7348	12.3617
std.err	1.2260	1.2290	1.3650	1.3200	1.1970	1.2770	0.9840	1.1860	1.0280

Appendix D

Trials on testing coefficients to reduce dimensionality with other methods, LASSO regression.




```
[(0.0, array([0.        , 0.        , 0.2134947 , 0.        , 0.        ,
       0.        , 0.53594298, 0.        , 0.        , 0.        ,
       0.        , 0.29705471, 0.        , 0.        , 0.        ,
       0.        , 0.        , 0.        , 0.        , 0.        ,
       0.        , 0.        , 0.        , 0.        , 0.        ,
       0.        , 0.        , 0.        , 0.        , 0.        ,
       0.        , 0.        , 0.        , 0.        , 0.        ,
       0.        , 0.        , 0.        , 0.        , 0.        ,
       0.        , 0.        ])), (1.0296742378540982, array([0.37949616, 0.19608845, 0
, 0.        ,
       0.        , 0.        , 0.        , 0.        , 0.        ,
       0.        , 0.        , 0.        , 0.        , 0.        ,
       0.        , 0.        , 0.        , 0.        , 0.        ,
       0.        , 0.        , 0.        , 0.        , 0.        ,
       0.32006023, 0.        , 0.        , 0.        , 0.32975471,
       0.        , 0.        , 0.        , 0.        , 0.        ,
       0.        , 0.        , 0.        , 0.        , 0.        ,
       0.        , 0.        , 0.        , 0.        , 0.        ,
```

요약 (국문초록)

구독형 전자상거래 시장은 지난 5년간 매년 100% 이상 성장세를 이어오고 있으며, 이에 따라 구독 서비스에 대한 연구의 중요성은 커지고 있다. 그러나, 중요성에 비해 구독서비스의 가격 책정은 대개 구조가 불분명하며, 합리적 가격제시는 고객관계관리의 가장 중요한 요소 중 하나임에도 소비자는 불명확한 가격 책정으로 불만을 호소하고 있다. 따라서 본 연구는 사용자가 생성한 고객 데이터를 바탕으로 고객의 선호도를 반영한 속성들을 고려하고, 고객이 합리적으로 수용할 수 있는 가격 책정 방식을 머신러닝과 쾌락적 가격책정 모델을 통해 보다 객관적으로 예측해 보고자 한다.

데이터는 10,000건의 온라인 리뷰를 활용했으며, 대표적인 5개의 동영상 스트리밍 구독 서비스들의 리뷰로 구성했다. 이 데이터를 가지고 머신러닝을 활용해 토픽모델링, 리뷰의 벡터화, 차원 축소 및 속성들의 가치 값을 도출했고, 이를 통해 헤도닉 가격 모형을 도출해 서비스의 합산 가격을 정의했다.

이 과정에서 텍스트 기반 데이터로 서비스 가격 책정 시 지도학습된 잠재디리클레할당 이 가능한 부분최소제곱법을 통해 각 값들 간 공분산을 고려함으로 예측도가 높은 값을 얻을 수 있음을 논했다. 또한, Netflix, Amazon Prime Video, HBO Max는 고객의 평가보다 높게, Disney+, Hulu는 낮게 가격책정 된 것으로 결과값이 도출된 것에 대하여 본 논문은 가격과 속성과의 관계를 통해 고객이 중요시 생각하는 속성들과 상황적 고려 등을 통해 각 서비스들에 대한 인사이트와 시사점을 제시했다.

.....

주요어: 가격책정, 구독서비스, 머신러닝, 잠재디리클레할당, 부분최소제곱법, 헤도닉가격모형

학 번: 2021-27567