TESIS

Optimización de Ventas de la Plataforma de Udemy

Torres Ochoa, Daniela Alejandra
Tutor: Rene Caldentey

RESUMEN

En el último año el crecimiento en la demanda de consumidores en busca de aprender a través de las plataformas digitales ha sido exponencial. Debido a la pandemia durante el 2020 el mercado de personas en busca de cursos en múltiples temas se expandió, aumentando la cantidad de instructores que generan contenido digital y aumentando la competencia entre las plataformas ya existentes y las nuevas que se transforman en virtuales debido al contexto mundial. Udemy es una de las principales plataformas que se ha visto favorecida de este crecimiento, obteniendo no solo nuevos clientes pero también nuevos instructores que crean nuevo material para la plataforma.

Este trabajo tiene como objetivo obtener una mayor comprensión sobre los aspectos que influyen en la demanda de los cursos ofrecidos en la plataforma de Udemy en la categoría de 'IT & Software' analizando las tendencias del mercado y sus principales competidores. Se plantean 5 etapas, la primera consiste en recolectar datos de la plataforma sobre los cursos de la categoría y realizar un análisis exploratorio para identificar patrones interesantes y plantear hipótesis sobre cuales pueden ser las variables mas relevantes. En la segunda etapa se van a ejecutar algoritmos de clustering para seleccionar los perfiles principales que se identifican en los cursos. Luego se va a realizar una investigación extensa sobre las tendencias del mercado y los principales competidores de Udemy para entender como estos impactan la demanda. La cuarta etapa consiste en modelar la demanda para cada uno de los cursos de los perfiles identificados. Analizando las diferencias entre los mismos y observando cuáles son los aspectos clave que verdaderamente impactan sobre la demanda. La etapa final consiste en permitir que las investigaciones realizadas y el modelo creado esté disponible para los instructores de la plataforma, creando un sistema de recomendación sobre acciones que pueden realizar los instructores para aumentar la demanda de sus cursos.

ABSTRACT

In the last year the growth of the demand of consumers looking to learn through digital platforms has been exponential. Due to the pandemic during 2020 the market for people looking for courses in multiple subjects expanded, the number of instructors who generate digital content increased and competition between existing platforms and new ones that became virtual due to the context increased as well. Udemy is one of the main platforms that has benefited from this growth, obtaining not only new clients but also new instructors who create new material for the platform.

This paper aims to obtain a better understanding of the aspects that influence the demand for the courses offered on the Udemy platform in the 'IT & Software' category by analyzing market trends and its main competitors. Five stages are proposed to carry out the objective, the first consists of collecting data from the platform about the courses in the category and conducting an exploratory analysis to identify interesting patterns and hypothesize about which may be the most relevant variables. In the second stage, grouping algorithms will be executed to select the main profiles that are identified in the courses. Then there will be extensive research on market trends and Udemy's top competitors to understand how they impact demand. The fourth stage consists of modeling the demand for each of the courses of the identified profiles. Analyzing the differences between them and observing which are the key aspects that truly impact demand. The final stage consists of making the research carried out and the model created available to instructors on the platform, creating a recommendation system based on actions that instructors can take to increase demand for their courses.

Tabla de contenido

1. Introducción

1.1 Motivación

Cada día más personas se suman a el aprendizaje en línea, sumándose a miles de plataformas distintas que ofrecen cursos de todo tipo y cada día más instructores comienzan a crear cursos en formato virtual que luego pueden ser adquiridos a través de estas plataformas. Permitiendo que personas de todo el mundo se conecten con esta nueva forma de aprendizaje.

Udemy es una plataforma digital que ofrece distintos tipos de cursos en todo tipo de áreas. Desde que fue fundada en 2009 su crecimiento ha sido masivo, especialmente durante este último año. A lo largo del tiempo se ha visto una variación en las políticas de precios de Udemy, se han cambiado en varias ocasiones los rangos de precios al que un instructor podía ofrecer su curso, si estos eran dinámicos o fijos, los descuentos para cada curso, entre otras. A medida que ha ido creciendo la plataforma más compleja se ha vuelto, ya que cada vez hay más cursos de distintas categorías, distintas duraciones y con diferentes demandas de parte de los consumidores para cada uno de estos. Encontrar los factores que influyen sobre la demanda es uno de los aspectos clave para mantener el interés tanto de los instructores que producen los cursos como de los consumidores que los adquieren, y que a la larga es lo que más le conviene a la compañía.

Subir material y cursos a Udemy es gratuito, se reparte un porcentaje de las ganancias por venta con la plataforma dependiendo del plan de marketing que seleccione el instructor. Por lo tanto, analizar que factores se pueden optimizar para aumentar la demanda es algo que no solo conviene a los instructores sino también a la plataforma.

1.2 Descripción del Problema

En el último año el crecimiento en la demanda de consumidores en busca de aprender a través de las plataformas digitales ha sido exponencial. Debido a la pandemia durante el 2020 el mercado de personas en busca de cursos en múltiples temas se expandió, aumentando la cantidad de instructores que generan contenido digital y aumentando la competencia entre las plataformas ya existentes y las nuevas que se transforman en virtuales debido al contexto mundial. Jeff Lieberman, director gerente de Insight Parteners, una firma que ha respaldado aproximadamente 24 empresas educativas afirmó:

"Con COVID-19 poniendo en peligro el aprendizaje en la escuela, esperamos que continúe la adopción generalizada del software de tecnología educativa, y esta no es una tendencia a corto plazo. El software remodelará el aprendizaje tanto en el aula como fuera de la pandemia".

Jeff Lieberman, Director gerente de Insight Parteners

Hoy en día según la política de precios actual de Udemy un instructor puede establecer el precio de su curso entre $19.99 y $200.00, colocarle un título y una descripción sin que se le provea ninguna ayuda o guía sobre que factores afectan la demanda del curso. Aquí es donde al instructor le surgen las preguntas ¿Qué puedo hacer para aumentar las ventas de mi curso? ¿Qué le puedo añadir a mi curso para hacerlo más atractivo para los consumidores? Generando incertidumbre para los instructores que no poseen información sobre las características de la demanda ni sobre las ventas de cursos similares a los suyos.

Si bien la plataforma de Udemy ha intentado optimizar su política de precios a lo largo de los últimos años, algunas de estas decisiones no fueron bien recibidas por los instructores. Según Thinkific en 2014 la plataforma cambió su política de precios de poder asignarle un precio a los cursos de hasta $300 a solo poder asignarle $50. Muchos instructores se vieron afectados por esta política ya que al tener cursos más largos y profesionales consideraban que la restricción de precio puesta por la plataforma era muy baja. Respondiendo unos meses después con un nuevo cambio a la política de precio que subió el tope de los cursos a $199.99 USD. Sin embargo, hoy en día la plataforma ha adoptado un sistema de descuento y aproximadamente el 90% de sus cursos se venden entre $12 y $13 USD.

Ante un contexto de competencia creciente y una demanda variante todavía hay mucha incertidumbre sobre cuáles son los principales aspectos que influyen sobre la demanda y sobre cuáles cambios se pueden realizar para mejorarla.

1.3 Objetivo

Según lo desarrollado anteriormente, se sabe que no existe ningún tipo de guía para los instructores al momento de subir sus cursos a la plataforma sobre que factores relacionados al tópico y las características de su curso afectan las ventas de este. Se sabe que la demanda por este tipo de cursos ha venido aumentando, especialmente durante este último año y con ella ha aumentado el nivel de competencia en este sector. El objetivo de esta propuesta es realizar un análisis de los principales factores que afectan la demanda y como se pueden utilizar para

optimizar las ventas de cada curso de Udemy de la categoría de 'It y Software' para maximizar las ventas de los instructores y de la empresa. Analizando e implementando algoritmos de clustering y modelos de regresión para analizar que aspectos aumentan la demanda de los cursos de la categoría y la satisfacción de los clientes en la plataforma.

El proyecto consistirá en construir una herramienta para los instructores de la plataforma, que en base a las características de cada nuevo curso que se quiera subir a la plataforma utilice los datos que se generan diariamente en la misma para predecir su demanda futura y analizar que factores pueden modificar para aumentarla. Es decir, se creará una herramienta de recomendación dinámica y fácil de utilizar en la que los instructores podrán subir sus cursos y las características de estos, aprovechando el modelo creado para analizar cada uno de sus cursos y observar cómo serían aproximadamente sus ventas en el futuro y cuales son los factores que pueden mejorar para aumentar sus ventas. Ofreciendo así una guía visual que los acompañe a lo largo de todo el proceso de dar clases a través de Udemy, haciendo recomendaciones y optimizaciones constantemente para maximizar la satisfacción de los instructores y los clientes. De esta forma Udemy se beneficiará al optimizar sus ventas para obtener las mayores ganancias posibles, al igual que los instructores.

1.4 Metodología

Para alcanzar los objetivos planteados anteriormente, se dividió el proyecto en cinco etapas descritas a continuación:

Etapa de Recolección de Datos

Los datos se van a recolectar a través de una Api suministrada por Udemy, esta será obtenida, procesada y organizada a través de varios scripts en Python y R. En donde se va a realizar una limpieza de los datos, un análisis exploratorio inicial y la ingeniería de atributos, para luego seleccionar las variables más relevantes para las siguientes etapas.

Etapa Descriptiva

Se utilizarán algoritmos de clustering en la data recolectada para identificar los principales perfiles de cursos para así categorizar los cursos dependiendo de sus características similares. La idea es correr distintos algoritmos y quedarse con el que mejor se adapte a los datos para obtener los perfiles principales de cursos.

Etapa de Investigación

Para desarrollar este nuevo modelo de optimización de ventas de la plataforma se analizará el nivel de competencia y sustitución que se encuentra hoy en día en el mercado en otras plataformas similares y como impactan estos aspectos en la demanda de los suscriptores. Igualmente se estudiará la perceptibilidad de los cursos en base a las tendencias del mercado de cursos de 'It y Software', tomando en cuenta el impacto que tendría en la demanda los cursos que pasan de moda o son sustituidos por versiones más modernas.

Etapa de Modelar la Demanda

Se van a generar múltiples datasets, uno semanal por un periodo aproximado de tres meses. De esta forma se obtendrán los suscriptores obtenidos en cada una de estas semanas, para así poder pronosticar la demanda en base a esta serie de tiempo probando distintos modelos de regresión. Con los datos ya procesados con las nuevas variables creadas durante la ingeniería de atributos se analizará cuales son los factores más relevantes que influyen sobre la demanda. Se evaluarán los modelos utilizando como criterios el coeficiente de determinación ajustados (R^2), el error estándar residual y el error de raíz cuadrático medio para la validación cruzada con la finalidad de seleccionar el modelo que mejor se ajuste a cada uno de los perfiles identificados en la etapa anterior.

Etapa de Optimización

En base a los modelos de regresión generados en la etapa de modelado de la demanda se generó un sistema de recomendaciones que pone en práctica todos los modelos y análisis realizados anteriormente. Generando un modelo que categorice cada curso en uno de los perfiles generados en la etapa de clustering y que evalué para cada caso cuales son los factores más relevantes que influyen sobre la demanda. Recomendando que modificaciones se pueden realizar para mejorar la demanda.

2. Los Datos

Los datos que van a ser utilizados para este estudio serán obtenidos de la 'Udemy Affiliate Api' (https://www.udemy.com/developers/affiliate/) otorgada por la misma plataforma sobre los cursos ofrecidos directamente de la página web de Udemy: https://www.udemy.com/courses/it-and-software/. Hoy en día se encuentran disponibles 10,000 cursos de la categoría de 'It y Software' y al obtener los datos de esta forma se aseguran tener datos actuales y reales. Se tiene una base de datos en la que cada fila corresponde a un curso y cada columna serían las características de cada uno de estos. Originalmente la base de datos contaba con 88 columnas, de las cuales se descartaron 38 durante la etapa de tratamiento de '*missings*' y limpieza de datos. Este proceso se explicará a detalle más adelante. En total después de la limpieza de datos, la base principal cuenta con 50 columnas y la complementaria con 6 columnas cuya descripción se puede ver en la Tabla 1 y la Tabla 2 respectivamente.

Tabla 1. Variables de la base de datos principal

	Feature	Descripción	Tipo de dato
1	**campaign_code**	Código de la campaña de descuento	string
2	**caption_languages**	Lenguaje de los subtítulos	string
3	**currency**	Tipo de cambio	string
4	**description**	Descripción	string
5	**estimated_content_length**	Duración en minutos	string
6	**headline**	Titular	string
7	**instructional_level_simple**	Nivel	string
8	**instructors**	Instructores	string
9	**language**	Lenguaje	string
10	**last_update_date**	Fecha de la última actualización	string
11	**objectives**	Objetivos	string
12	**prerequisites**	Prerrequisitos	string
13	**quality_status**	Status de calidad	string
14	**target_audiences**	Audiencia	string
15	**title**	Nombre	string
16	**url**	Url	string

17	**apple_in_app_purchase_price**	Precio en Apple	int
18	**avg_rating**	Average course rating	int
19	**discount_percent**	Porcentaje del descuento	int
20	**discount_price**	Precio descontado	int
21	**google_in_app_purchase_price**	Precio en google apps	int
22	**id**	Identificador único	int
23	**list_price**	Precio de lista	int
24	**num_curriculum_items**	Cantidad de material.	int
25	**num_lectures**	Número de lecturas	int
26	**num_of_published_curriculum_objects**	Cantidad de material publicado	int
27	**num_published_lectures**	Número de lecturas publicadas	int
28	**num_published_quizzes**	Número de quizzes publicados	int
29	**num_quizzes**	Número de quizzes	int
30	**num_reviews**	Número de reseñas	int
31	**num_subscribers**	Número de suscriptores	int
32	**price**	Precio	int
33	**quality_review_score**	Puntaje del status de calidad	int
34	**saving_price**	Dinero ahorrado	int
35	**subcategoryId**	Subcategoría	int
36	**campaign_end_time**	Fecha de fin de la campaña	date
37	**campaign_start_time**	Fecha de inicio de la campaña	date
38	**created**	Fecha de creación	date
39	**published_time**	Fecha de publicación	date
40	**has_certificate**	Si tiene certificado	bool
41	**has_closed_caption**	Si tiene subtítulos	bool
42	**has_discount_saving**	Si tiene descuento	bool
43	**is_available_on_google_app**	Si está disponible en google apps	bool
44	**is_available_on_ios**	Si está disponible en ios	bool
45	**is_banned**	Si está prohibido	bool
46	**is_in_any_ufb_content_collection**	UFB	bool
47	**is_marketing_boost_agreed**	Si está en la campaña de marketing	bool
48	**is_practice_test_course**	Si es un curso de prueba	bool

49	**is_published**	Si ha sido publicado	bool
50	**isPaid**	Si es pago	bool

En la sección de *Anexo 1* se puede visualizar una muestra de los primeros diez registros de todas las columnas del dataset de la primera semana.

Con respecto a la recolección de los datos se creó un 'script' de Python que extraía la información de la api de Udemy , se corrió este script semanalmente y se almacenaron los datos desde el 19 de enero de 2021 hasta el 6 de abril de 2021. Por lo que se contaba con doce dataset, uno por cada una de las doce semanas por las que se recolectó la información y todos con la misma estructura descrita anteriormente.

3. Análisis Exploratorio

Previo a comenzar a trabajar con los datos, se estudió el significado de cada una de las variables del dataset. Tras entender que significaba cada una, se decidió agruparlas según lo que representaban, por ejemplo, las variables que comenzaban por **num** fueron clasificadas como variables que representan la cantidad de distintos atributos de un curso, mientras que variables como **description** o **headline** eran variables descriptivas que aportan las distintas características de un curso en formato de texto. Una de las agrupaciones principales que se hizo de los atributos fue dividirlos en dos grupos, uno que representara aquellos atributos que varían con el tiempo en un mismo registro como por ejemplo **num_subscribers** que muestra el número de suscriptores de un curso y en cada dataset este número variaba dependiendo de las nuevas personas que se suscriben. Luego están los atributos que se categorizaron como variables fijas que suelen permanecer con el mismo valor para todos los datasets, un ejemplo de estas son **title**, **url** o **subcategoryId**. El diferenciar lo que representan los atributos nos permitió en el futuro crear nuevas variables para estudiar y entender las características principales de los cursos de Udemy para la etapa de *clustering*.

Tabla 2. Agrupación de los atributos: fijos y cambiantes

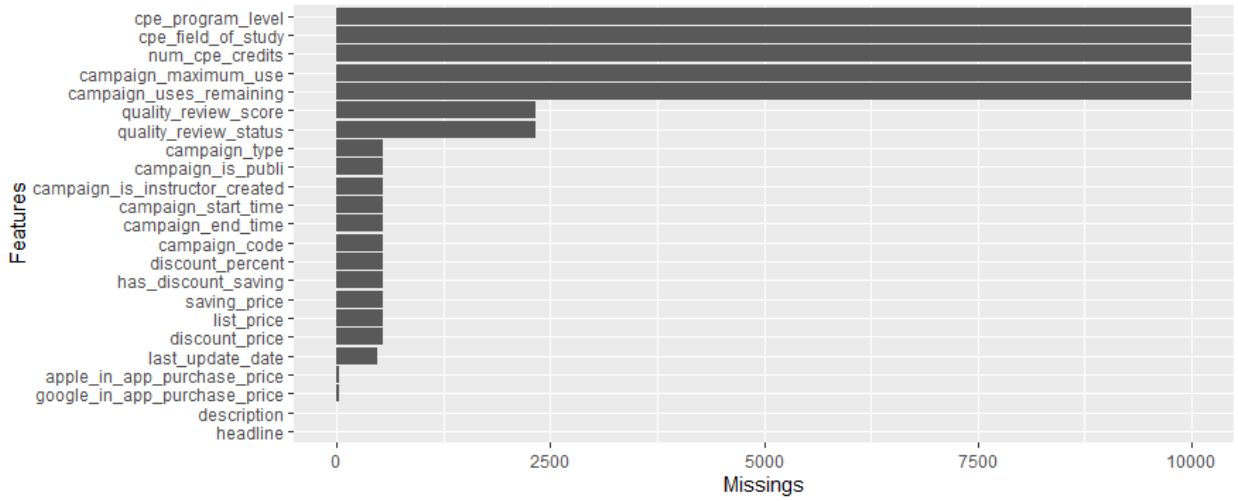
Atributos Fijos	**Atributos Cambiantes**
caption_languages	campaign_code
currency	last_update_date
description	quality_status
estimated_content_length	apple_in_app_purchase_price
headline	avg_rating
instructional_level_simple	discount_percent
instructors	discount_price
language	google_in_app_purchase_price
objectives	list_price
prerequisites	num_curriculum_items
target_audiences	num_lectures
title	num_of_published_curriculum_objects
url	num_published_lectures
id	num_published_quizzes
subcategoryId	num_quizzes
created	num_reviews
has_certificate	num_subscribers
has_closed_caption	price
is_available_on_google_app	quality_review_score
is_available_on_ios	saving_price
is_in_any_ufb_content_collection	campaign_end_time
is_marketing_boost_agreed	campaign_start_time
is_practice_test_course	published_time
isPaid	has_discount_saving
is_banned	-
is_published	-

3.1 Tratamiento de 'missings' y limpieza de datos

Para comenzar se realizó un análisis de la cantidad de missings o NANs por columnas. Inicialmente se contaba con 88 columnas en nuestros datasets originales, entre estas columnas había algunas que se encontraban completamente vacías y otras en las que se presentaba el mismo valor para todos los registros, por lo que se descartaron 38 columnas ya que no contenían información relevante.

Para algunas columnas los valores nulos eran permitidos como por ejemplo para los atributos asociados a precios descontados o código de las campañas de descuento, ya que en algunos casos algunos cursos optan por no participar de estos descuentos y por ende su valor era nulo.

Gráfico 1. Cantidad de valores nulos por columna del dataset Semana 1



Como se puede observar las variables como **cpe_program_level** y **campaign_maximum_uses** contaban únicamente con valores nulos. Variables como estas fueron descartadas de todos los datasets. También se descartaron columnas cuya información ya era suministraba por algún otro atributo.

Después de descartar las columnas que no aportaban al estudio, los datasets contaban con 16 atributos de texto, 19 numéricos, 4 de fecha y 11 que representan variables binarias. De las cuales las columnas de **subcategoryId**, **quality_status**, **intructional_level_simple** y **campaign_code** representan variables de tipo factor cuyas categorías son las siguientes:

- **subcategoryId:**
    - 132: IT Certification
    - 134: Network & Security
    - 136: Hardware
    - 138: Operation Systems
    - 140: Other IT & Software
- **quality_status:** approved, banned, needs_fixes
- **intructional_level_simple:** All Levels, Beginner, Intermediate, Expert
- **campaign_code:** HABITSPAYOFF, NEWLEARNINGS, UDEMYEDUCATION21, WAYSTOLEARN21

Luego se notó la existencia de registros duplicados en los datasets y se procedió a eliminarlos. Se contaban con 12 datasets con 10000 registros cada uno. En la siguiente tabla se pueden observar el porcentaje de registros que estaban duplicados y la cantidad de registros únicos que permanecieron.

Tabla 3. Registros duplicados y únicos

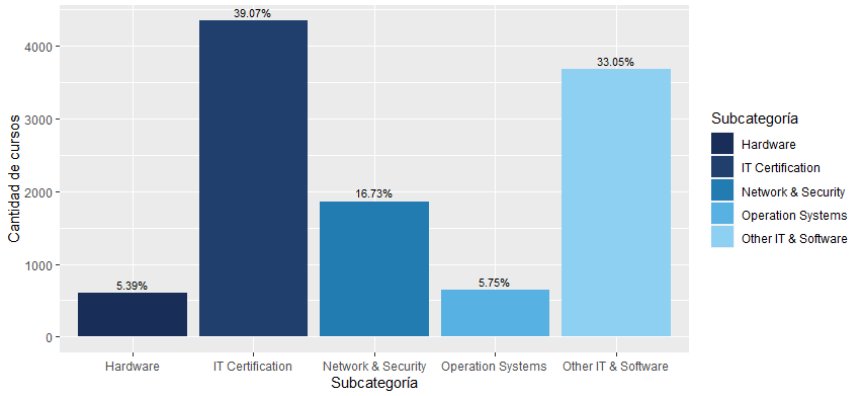
Semana	1	2	3	4	5	6	7	8	9	10	11	12
% duplicados	13.2	13.8	13.0	13.2	14.2	13.2	15.2	12.5	10.8	12.9	12.3	14.1
Registros únicos	8678	8616	8700	8680	8579	8684	8482	8751	8924	8708	8773	8586

3.2 Análisis general de los atributos fijos

Para un análisis general de los atributos que no suelen variar en el tiempo se unieron todos los dataset en uno global tomando las características de los registros únicos. En total se cuentan con 11120 distintos cursos en todos nuestros datasets, tomando en cuenta que hay cursos que fueron creados después de la fecha en la que se comenzaron a recolectar datos.

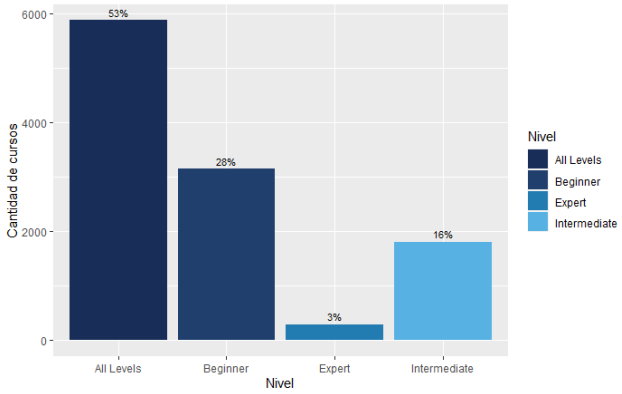
Inicialmente se analizó la distribución de los cursos según la subcategoría de este, notándose que la mayoría de los cursos son de '*IT Certification'* seguido de *'Other IT & Software'.*

Gráfico 2. Cantidad de registros únicos de cada subcategoría



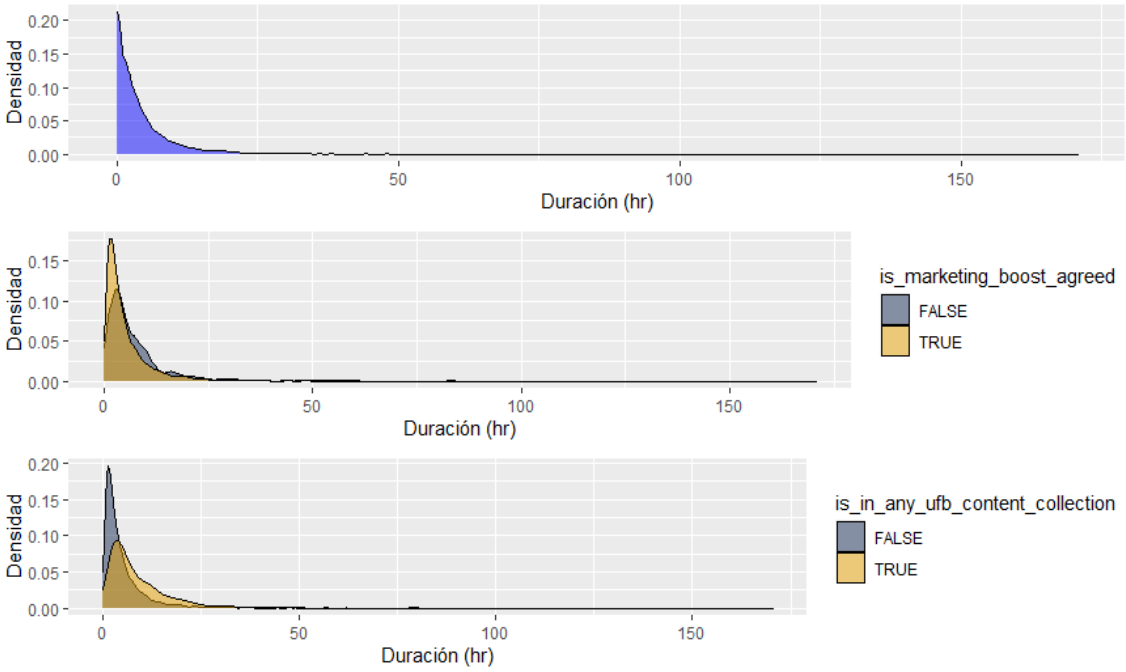
Se observó que más de la mitad de los cursos están calificados para todo tipo de niveles de dificultad y que existen tan solo 3% de cursos para expertos.

Gráfico 3. Cantidad de registros únicos de cada Nivel



Por otro lado, un 99.7% de los cursos se encuentran disponibles en las plataformas de IOS y de Google Apps.

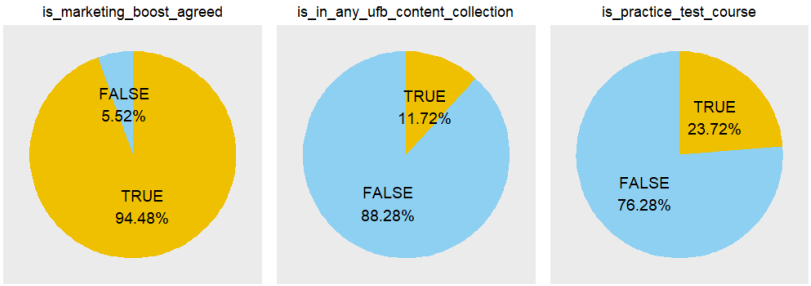
Gráfico 4. Duración del contenido de video de los cursos



La gran mayoría de los cursos tienen videos cuya duración es menor a 25 horas y en promedio estos duran 4 horas y 30 minutos. Cabe destacar que 2610 de los cursos, es decir un 23.6% no cuentan con contenido de video por lo que su duración es cero.

Por otro lado, el 94.48% se encuentra participando de la campaña de marketing, un 88.21% no presenta contenido en UFB y solo un 23.61% son cursos para practicar para pruebas. Udemy realiza un filtrado de sus cursos y crea una lista de los cursos más demandados y de mejor calidad, los cursos calificados como UFB son los seleccionados para esta lista. Esto será explicado con más detalle en los próximos capítulos.

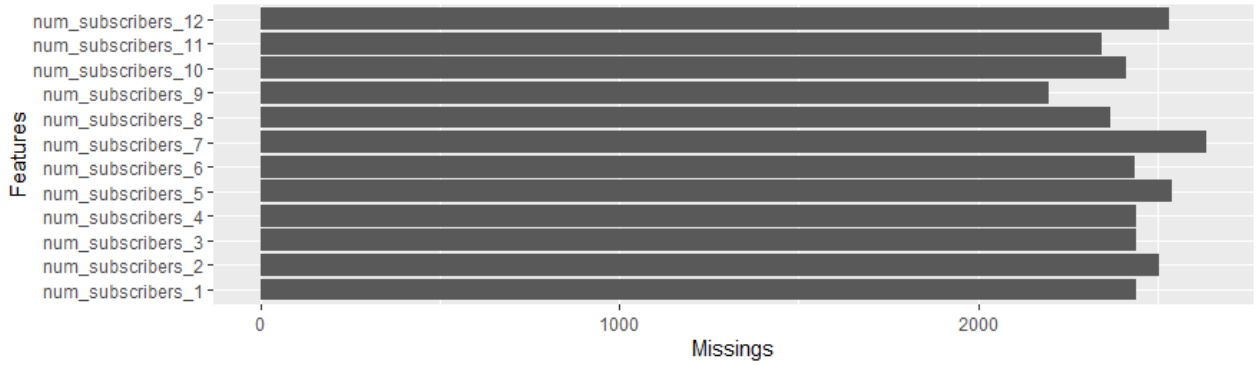
Gráfico 5. Porcentajes de variables binarias principales



3.3 Análisis general de los atributos cambiantes

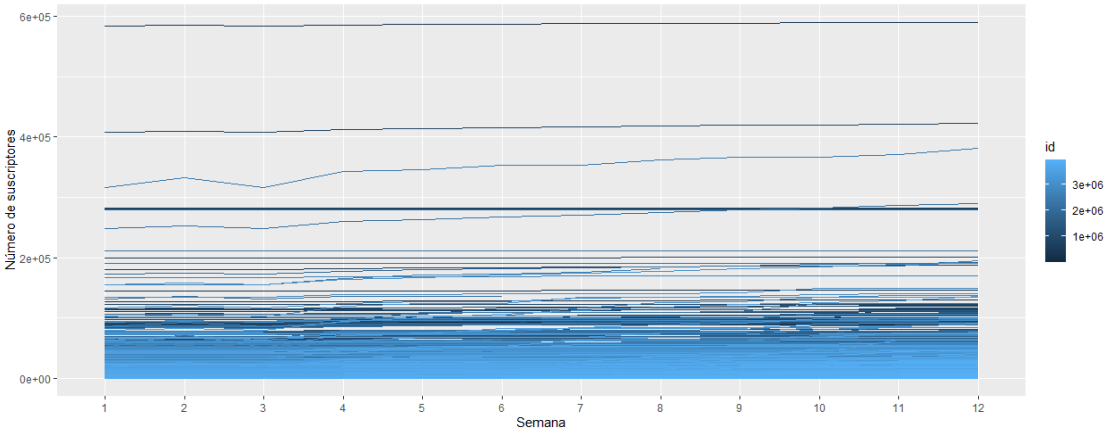
Luego se procedió a analizar los atributos que varían con el tiempo. Es importante aclarar que la evolución que se analiza de estas variables es representativa del periodo en el cual se recolectaron los datos. Para el análisis se crearon distintas tablas para analizar la evolución de estos a lo largo de las doce semanas para cada registro. El primer atributo que se analizó fue el de número de suscriptores (**num_subscribers**), ya que este es elemental para pronosticar la demanda en futuras etapas. En este caso la tabla se observó que existían en promedio 2.441 valores nulos por semana, lo que tiene sentido ya que en promedio semanalmente se recolectaron datos sobre 8.680 y en total en todos los datasets se contaba con 11.120 cursos distintos debido a los cursos nuevos que se iban creando y los que dejaban de existir.

Gráfico 6. Cantidad de missings por semana



La distribución de los suscriptores de cada curso era variada pero como se puede observar en el próximo gráfico la gran mayoría de los cursos tiene menos de 200.000 personas suscritas.

Gráfico 7. Distribución del número de suscriptores por curso a través de las semanas.



Cuando se comenzaron a recolectar datos la cantidad total de suscriptores en los cursos de 'IT & Software' era de 40.366.291 y para la semana 12 se contaba con 43.343.848 suscriptores, aumentando en un 7,4%. Igualmente aumentó el promedio de personas en cada curso en 8,5% de 4.652 suscriptores a 5.048.

Gráfico 8.Total de suscriptores por semana

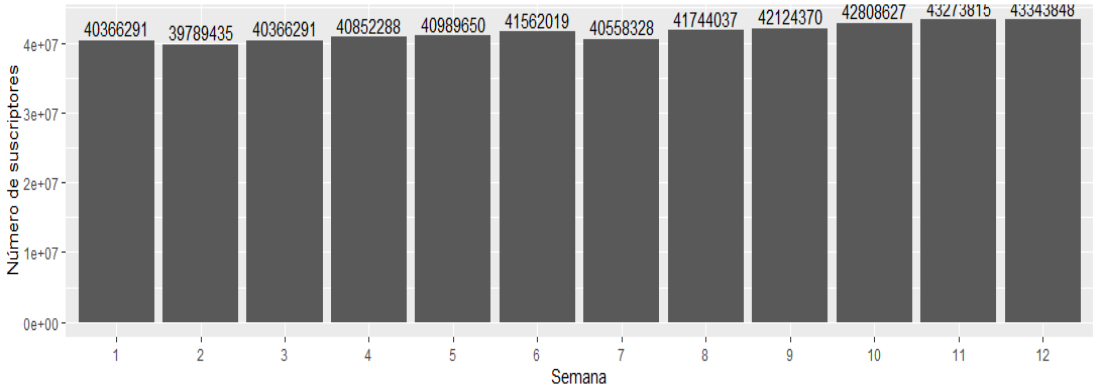
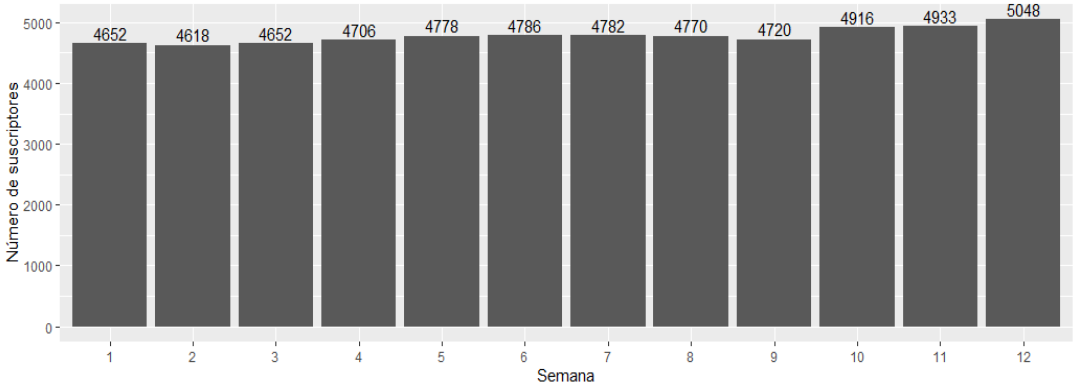
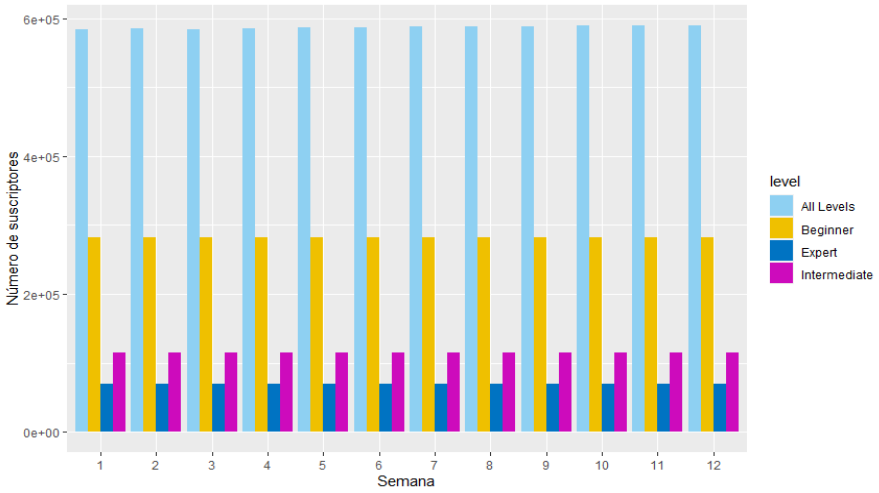


Gráfico 9. Media de suscriptores por curso cada semana



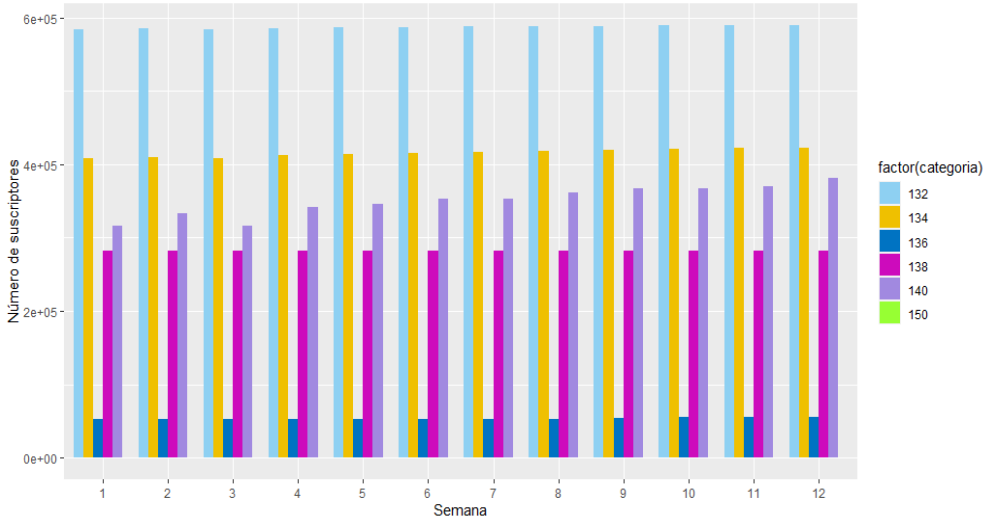
En cuanto al número de suscriptores por nivel estos permanecen relativamente constantes y no se notan cambios significativos de semana a semana.

Gráfico 10. Total de suscriptores por nivel de cada semana



En cuanto al crecimiento del número total de suscriptores por subcategoría se pudo observar un incremento mayor en la categoría de *'Other IT & Software (140)*' que en el resto de las subcategorías en el que el crecimiento fue más leve.

Gráfico 11. Total de suscriptores por subcategoría de cada semana



Luego se estudiaron las variables relacionadas con el precio del curso, se analizaron principalmente la variable del precio completo del curso (**price**) y la variable del precio descontado (**discount_price**) para aquellos cursos que ofrezcan un descuento. En la tabla a continuación se puede visualizar el porcentaje de los cursos cada semana que ofrecían sus cursos con descuento. En promedio un 94,93% ofrece un descuento importante sobre el precio

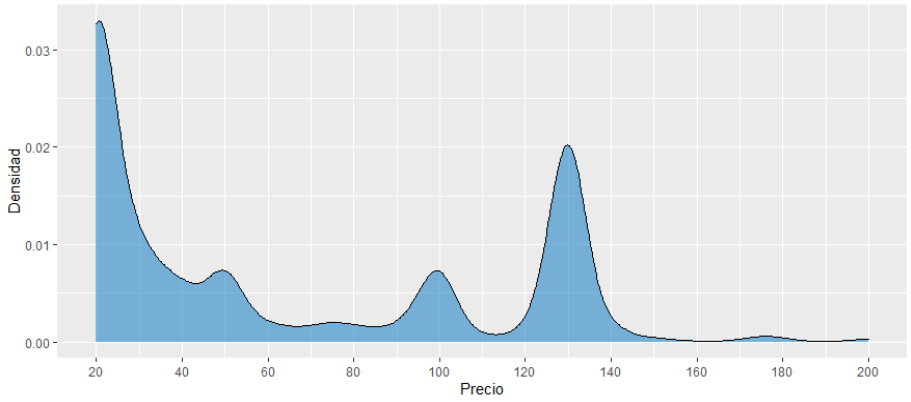
del curso, por lo que se puede concluir que una gran mayoría de los cursos presentan descuentos todas las semanas.

Tabla 4. Porcentaje de cursos con descuentos cada semana

Semana	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
% Descuento	94.69	95.24	94.69	94.85	94.95	94.86	94.93	95.11	94.94	94.99	94.95	95.02

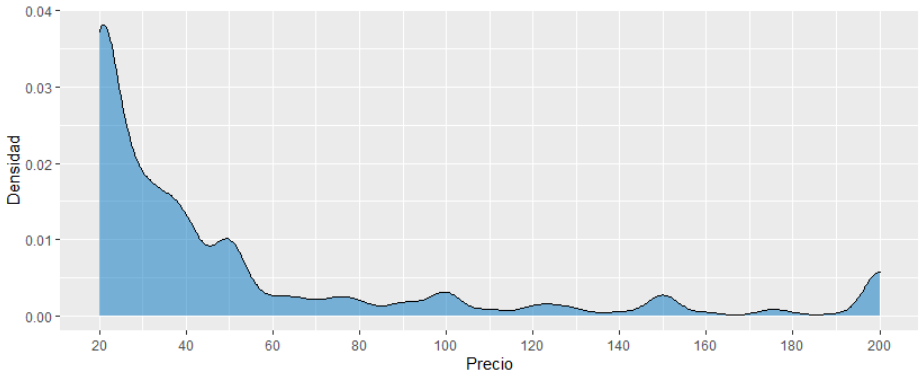
Tomando en cuenta el precio general de todos los cursos sin considerar el descuento este oscila entre los 20 y los 200 dólares acorde a los límites que establece la plataforma a los instructores para seleccionar el precio de un curso. La mayor cantidad de cursos tiene un precio aproximado de 21$ USD.

Gráfico 12. Precio total de los cursos sin considerar el descuento



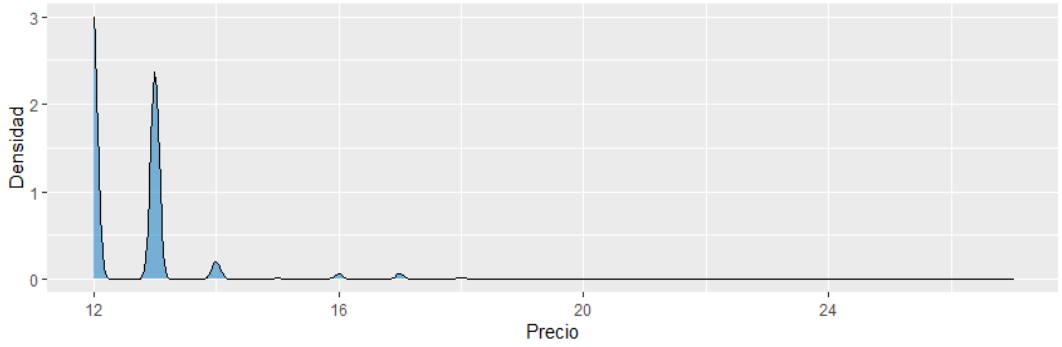
En promedio sólo 5,07% de los cursos en todos los dataset no cuentan con un descuento. En los últimos años Udemy fue alterando su política de precio enfocándose en ofrecer cursos con grandes descuentos como técnica de marketing.

Gráfico 13. Densidad del Precio Total de los cursos sin descuento



Como se mencionó anteriormente el 94,93% de los cursos ofrece grandes descuentos todas las semanas. Como se puede observar en el próximo gráfico los precios disminuyen radicalmente cuando el descuento es aplicado, concentrando la mayoría de los cursos en un costo que oscila desde los 12$USD a los 16$USD.

Gráfico 14. Densidad del Precio Total de los cursos con descuento



Estos descuentos de los precios varían de semana a semana debido a la política de precios y descuentos dinámicos de la plataforma. Resulta interesante notar por ejemplo que en algunas semanas la mayoría de los precios descontados era menor a 12.5$$USD como fue en el caso de la primera semana y en otras semanas todos los precios se desplazaban y eran casi todos en su mayoría mayores a este límite como en la segunda semana.

Gráfico 15. Densidad del Precio Total de los cursos con descuento por semana

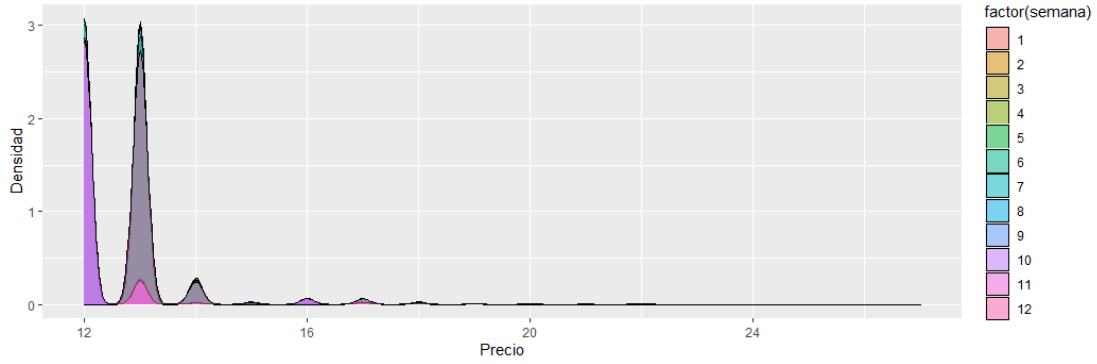
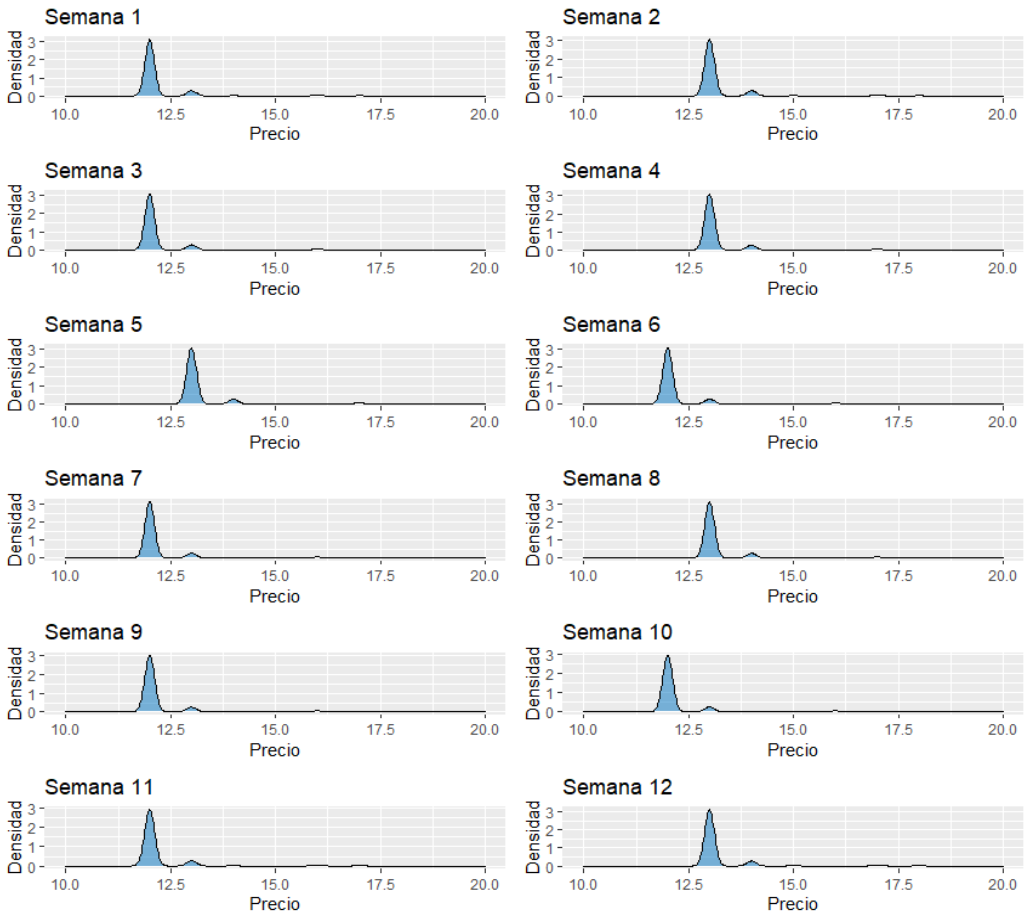


Gráfico 16. Densidad del Precio Total de los cursos con descuento por semana individuales



Igualmente se comparó la distribución de los precios descontados en base al nivel y a la subcategoría del curso. Notando que a pesar de que no se observó ninguna diferencia radical si se percibió que para algunos precios el precio era mayor en algunos niveles. Por ejemplo, los cursos cuyo precio oscila alrededor de 13$USD existen casi el doble de cursos calificados como 'All Levels' que cursos de nivel 'Intermediate'.

Gráfico 17. Densidad del Precio Total de los cursos con descuento agrupados por nivel

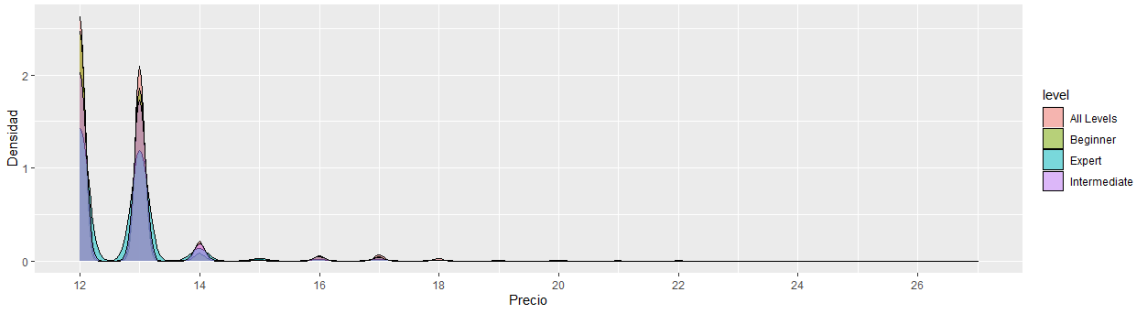
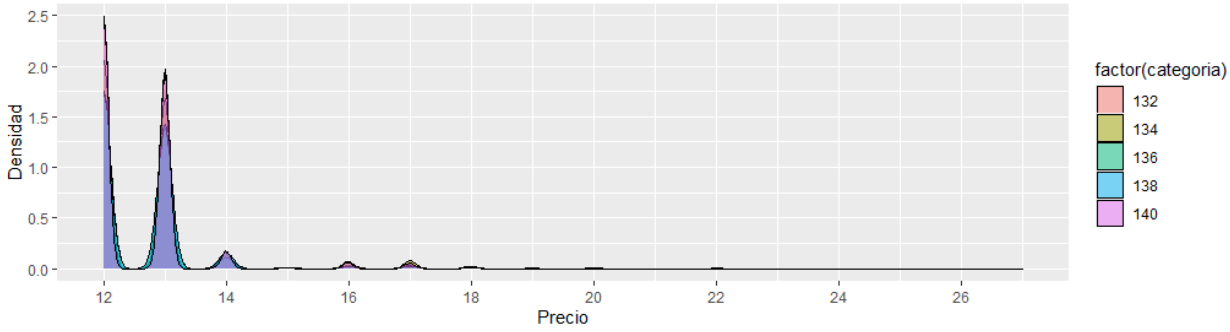


Gráfico 18. Densidad del Precio Total de los cursos con descuento agrupados por subcategoría



4. Etapa de Clustering

Una vez que se tuvo una mayor comprensión de los datos después de su análisis y su limpieza necesaria se procedió con la etapa de clustering. El Clustering consiste, como lo define Max Bramer en su el libro de *Principio de Minería de Datos*, en agrupar objetos que son similares entre sí y diferentes a los objetos que pertenecen a otros grupos. Como se mencionó anteriormente se utilizarán algoritmos de clustering en la data recolectada para identificar los principales perfiles de cursos, para así categorizar los cursos dependiendo de sus características similares.

Primero se analizó que algoritmo de 'clustering' se iba a utilizar en base a los datos con los que se contaba. Debido a que se tenían variables numéricas y variables categóricas se concluyó que la mejor opción era implementar el algoritmo de K-prototypes, el cual permite agrupar conjuntos de datos mezclados integrando los algoritmos de k-means y k-modes.

En el articulo '*Extension to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values*', Zhexue Huang establece que k-means es conocido por su eficiencia agrupando grandes cantidades de datos. Este es un algoritmo de partición o de agrupamiento no jerárquico que utiliza la distancia Euclidiana para calcular los centros de los agrupamientos, dividiendo en partes una cantidad de objetos dados en *k* grupos o 'clusters' minimizando la suma de los errores cuadrados (WGSS) de cada grupo, pero este solo se puede implementar con atributos numéricos. Por otro lado, en su artículo Huang aclara que el algoritmo de k-modes utiliza un método parecido al de k-mean siendo la estructura del algoritmo la misma. Su principal diferencia consiste en que este utiliza una medida de disimilaridad para comprar objetos, reemplazando el uso de promedios por el de modas y utilizando un método basado en

frecuencias para actualizar las modas. Este algoritmo se puede implementar con atributos categóricos.

El algoritmo K-prototype integra estos dos algoritmos permitiendo analizar grandes cantidades de datos con variables continuas y discretas (numéricas y categóricas). Utilizando como medida de disimilaridad el cuadrado de la distancia Euclidiana en el caso de los atributos numéricos y el número de inconcicidencias en el caso de los atributos categóricos. La ventaja de utilizar este algoritmo de clustering es que no es complejo, permite trabajar con grandes cantidades de datos y es mejor que los algoritmos de base jerárquica.

4.1 Selección de variables e ingeniería de atributos

Para implementar el algoritmo de k-prototype era necesario hacer algunas transformaciones y agregar nuevas variables a nuestro dataset. Primero se realizó una imputación para reemplazar los NANs de las variables que variaban a lo largo de las semanas, reemplazandolos por el valor de las variables del mismo curso de la semana anterior. Por ejemplo, en algunos casos se contaba con el número de suscriptores de un curso para todas las semanas menos de la semana 3 ya que por alguna razón este campo había llegado vacío. Por lo tanto, para no romper la línea de tiempo y sabiendo que el número de suscriptores no baja de semana a semana se reemplazó el Nan por el valor de la semana 2. Esta imputación se realizó en las siguientes variables:

- **avg_rating**
- **num_curriculum_items**
- **apple_in_app_purchase_price**
- **google_in_app_purchase_price**
- **num_lectures**
- **num_of_published_curriculum_objects**
- **num_published_lectures**
- **num_published_quizzes**
- **num_quizzes**
- **num_reviews**
- **num_subscribers**
- **price**
- **saving_price**
- **discount_price**
- **discount_percent**
- **num_published_lectures**

Se imputaron los Nans de las columnas de **campaign_code** remplazándolos por una categoría llamada “NONE” la cual representa todos los casos en los que un curso no se encuentra en ninguna campaña de promoción.

Luego de las imputaciones se excluyeron del análisis los cursos que no estuvieron presentes en el estudio desde la primera semana en la que se comenzó a recolectar los datos, optando así por quedaron con 8536 cursos de los que se contaba información todas las semanas. De esta forma se obtuvo un dataset sin ningún valor nulo, el cual era uno de los requisitos para ejecutar el algoritmo k-prototype el cual no admite valores nulos.

Luego se procedió a la creación de variables que se consideraron serían relevantes para el estudio. Se crearon variables que representan los valores máximos de los atributos que representan el precio y el precio de descuento durante las doce semanas denominadas **max_price** y **max_discount_price**. Luego se crearon un conjunto de variables que reportaban el promedio semanal por curso de las variables que varían en el tiempo:

- **avg_price**
- **avg_discount_percent**
- **avg_discount_price**
- **avg_num_published_lectures**
- **avg_num_curriculum_items**
- **avg_num_lectures**
- **avg_num_of_published_curriculum_objects**
- **avg_num_published_quizzes**
- **avg_num_quizzes**
- **avg_num_reviews**
- **avg_num_subscribers**

A continuación, se creo **sum_has_discount_saving** la cual es la suma de las variables **has_discount_saving** reportando la cantidad de semanas en el periodo estudiado en las que el curso tuvo descuento. Unos de los atributos más relevantes creados fueron las variables relacionadas a la demanda semanal. Para esta se creó una variable de demanda para cada semana que equivalía al incremento que había tenido la cantidad de suscriptores ese curso en comparación con la semana anterior. De estas variables surgieron **avg_demand** y **max_demand,** representando la demanda promedio y la demanda máxima semanal por curso.

Luego se crearon variables que representaban el largo de las variables de tipo texto, es decir la cantidad de caracteres de estos atributos por curso:

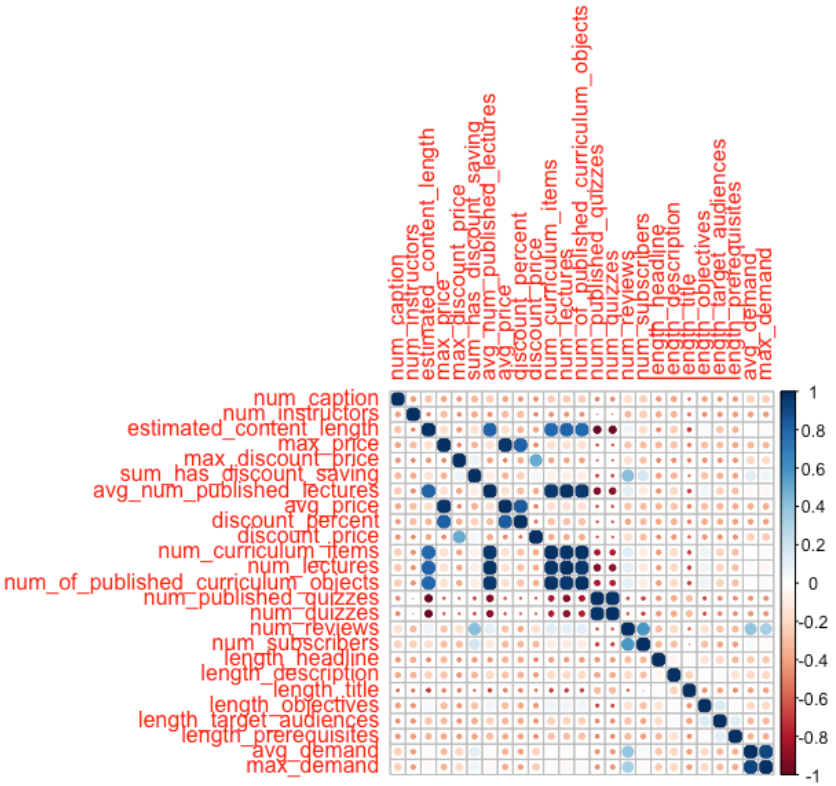
- **length_headline**
- **length_description**
- **length_title**
- **length_objectives**
- **length_prerequisites**
- **length_target_audiences**

Por último, se crearon los atributos **num_instructors** representando la cantidad de instructores por curso y **num_caption** representando la cantidad de lenguajes en la que el curso

ofrece subtítulos. Para que el algoritmo identificara correctamente los atributos a analizar con k-modes y k-mean las variables categóricas se identificaron como variables de tipo factor y el resto como numéricas. Igualmente, como parte del pre-procesamiento de datos se eliminaron algunas columnas originales del conjunto de datos que son irrelevantes para el algoritmo de agrupación de K-Prototype. Las columnas como **is_paid**, **is_available_on_google_app** y **is_available_on_ios** se eliminaron ya que todos los cursos son pagos y se encuentran disponibles en formato de aplicación por lo que estas variables siempre eran verdaderas por lo que no iban a tener impacto. Las columnas como **description** y **headline** presentaban valores largos de texto únicos en su mayoría que representarían una elevada carga computacional y no un gran aporte para el algoritmo. De último se eliminó la columna de ID la cual tiene información sin sentido para el análisis.

Para analizar relaciones interesantes que pudieran surgir en el análisis se estudió la correlación entre las variables numéricas con la que se iba a trabajar en el algoritmo. Para esto se creó una matriz de correlación con la función *cor()* de R. Se puede observar por ejemplo que muchas de las variables **num** que cuantifican el número de ciertos materiales por curso se encuentran fuertemente correlacionadas con la variable **estimated_content_length**. Esta matriz nos va a permitir explicar alguna de las relaciones más interesantes en los cluster creados a continuación.

Gráfico 19. Matriz de correlación



Una vez se tenían todas las nuevas variables era necesario escalar las variables numéricas debido a que todas se encontraban en distintas unidades y esto podría ocasionar un sesgo en los resultados dándole más importancia a los atributos cuyos valores eran mayores. Se utilizó la función de *scale()* de R para escalar todas las variables, esta función centra todos los valores restando a cada uno el promedio de cada atributo.

4.2 Algoritmo K-Prototype

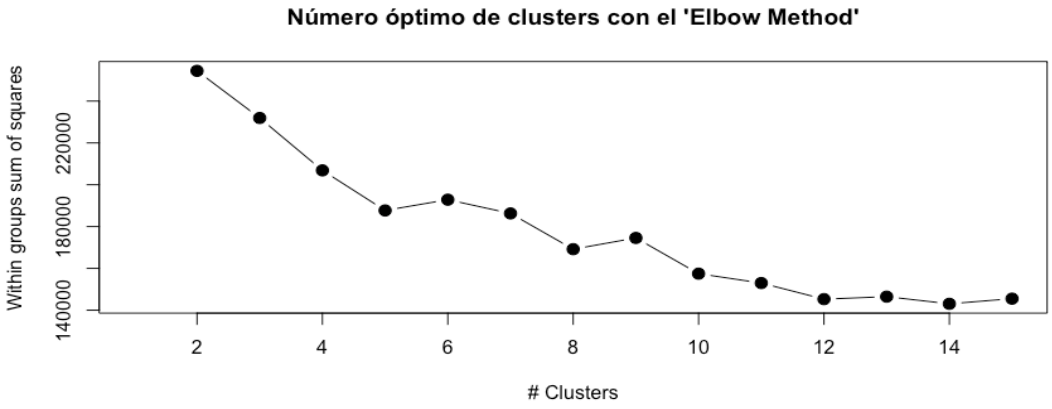
Una vez procesado el conjunto de datos sin valores nulos y con los valores numéricos escalados se procedió con la implementación del algoritmo. Para seleccionar la cantidad de 'clusters' óptimos para nuestro análisis se utilizó el "Elbow Method" o método del codo, el cual consiste en correr el algoritmo con valores crecientes de k, siendo k el número de 'clusters', y evaluar cómo varía la función a minimizar, eligiendo el valor de k en donde aparezca un "codo". Para la implementación se utilizó de la librería clustMixType la función de kproto que computa el algoritmo K-Prototype para datos mixtos en R. Esta proporciona una función de costo que combina el cálculo de las variables numéricas y categóricas denominada *'withinss'* que

representa la suma de las distancias de todas las observaciones que pertenecen a un grupo de su centro. Esta fue la función de costo con la que se compararon los modelos, cuanto menor es esta distancia 'intra-clusters' mejor por lo que se busca con el método del codo un valor de k en el que un incremento de este no represente una mejora sustancial en la distancia media 'intra-clusters'.

Se ejecutó múltiples veces el algoritmo de K-Prototype con los datos procesados, inicialmente sin las nuevas variables y luego añadiendo las nuevas variables de poco para visualizar su impacto. Finalmente añadimos todos los nuevos atributos relevantes a nuestro dataset resultante, el cual una vez eliminados los valores nulos y eliminados los atributos que fueron descartados contaba con 8 variables categorías y 25 variables numéricas.

En este punto se ejecutó el algoritmo de K-Prototype aumentando el número de grupo desde 2 'clusters' hasta 15 y se obtuvo el siguiente gráfico como resultado.

Gráfico 20. Número óptimo de clusters con el Elbow Method



Visualizando estos resultados se optó por seleccionar 5 'clusters' para nuestro análisis ya que era el primer punto donde se evidenciaba una clara forma de codo. También se analizaron resultados con 8 'clusters' pero estos no resultaron tan interesantes en comparación. Al regresar los valores a su escala normal se obtuvieron los siguientes cinco perfiles.

Tabla 5. Perfiles de clusters

Clusters	1	2	3	4	5
instructional_level_simple	All Levels	Beginner	All Levels	All Levels	All Levels
language	en_US	en_US	en_US	en_US	en_US
subcategoryId	140	140	134	132	132
has_certificate	True	True	True	True	False
has_closed_caption	True	True	True	True	False
is_in_any_ufb_content_collection	False	False	False	False	False
is_marketing_boost_agreed	True	True	True	True	True
is_practice_test_course	False	False	False	False	True
num_caption	1	1	1-2	2-3	1
num_instructors	1	1	2	1	1
estimated_content_length	371 min	198 min	283 min	1469 min	74,42 min
max_price	90,31	31,33	110,18	96,63	48,48
max_discount_price	13,30	13,13	13,37	12,52	10,66
sum_has_discount_saving	10,49	10,17	11,00	10,44	7,18
avg_num_published_lectures	50,04	28,24	44,32	241,83	8,01
avg_price	88,35	30,78	108,68	95,84	47,39
discount_percent	76,62 %	50,82 %	84,00 %	69,72 %	46,96 %
discount_price	12,67	12,50	12,69	11,81	10,16
num_curriculum_items	53,20	29,90	47,40	260,72	11,59
num_lectures	50,79	28,53	45,15	249,56	8,14
num_of_published_curriculum_objects	52,38	29,59	46,51	252,61	11,42
num_published_quizzes	1,81	0,92	1,90	10,06	3,38
num_quizzes	1,86	0,94	1,96	10,41	3,42
num_reviews	285,76	155,82	1018,79	3712,96	116,24
num_subscribers	3708	3385	16551	23463	1399
length_headline	92	70	100	96	88
length_description	1902	1051	2427	3529	1607
length_title	48	41	49	51	52
length_objectives	461	252	446	688	221
length_target_audiences	226	131	234	268	153
length_prerequisites	149	92	149	196	116
avg_demand	11,76	22,09	204,60	159,48	13,06
max_demand	27,24	145,28	1297,79	342,93	51,25

4.3 Análisis de los resultados de clustering

Al analizar los resultados del algoritmo de K-Prototype se noto que para algunos atributos se obtuvieron unos resultados muy interesantes. Solo en el caso de los atributos **language**, **is_in_any_ufb_content_collection**, **is_marketing_boost_agreed** y **is_practice_test_course** se obtuvieron los mismos valores para todos los perfiles, lo que es consistente ya que en su mayoría casi todos los cursos presentaban estos valores. Por ejemplo, una vez limpia la base de datos 92,2% de los cursos eran dictados en inglés americano (en_US) y 94,3% de ellos formaban parte del plan de marketing de Udemy, es decir que la variable **is_marketing_boost_agreed** era verdadera.

Analizando primero las variables categóricas en cuanto al nivel del curso cuatro de los perfiles pertenecen al nivel general de 'All level' y uno a el nivel de principiantes 'Beginner', en este caso 52% de los cursos pertenecían al primer nivel y 26% al segundo, por lo que es acorde que estos fueran los dos niveles que resaltaron en los perfiles. Por el lado de la subcategoría, los primeros dos perfiles presentaron en su mayoría pertenecer a "Other IT & Software" (140), el tercero a "Network & Security" (134) y los últimos dos a "IT Certification" (132), quedando solo 12% de los cursos que pertenecían a otras subcategorías. Con el resto de las variables categóricas resulta interesantes remarcar como el perfil 5 se distinguió del resto en especial con las variables con valores binario. Se puede observar que los atributos **has_certificate** y **has_closed_caption** tomaron valores verdaderos para los primeros cuatro perfiles y falso para el quinto y para el atributo **is_practice_test_course** paso igual, pero con los valores invertidos. Dando a entender que puede existir una correlación entre estas tres variables, destacando el perfil cinco como cursos que no tienen un certificado, no tienen subtítulos y son cursos de práctica para pruebas.

Analizando las variables numéricas resulta interesante la variedad observada en cuanto a el largo del contenido de video (**estimated_content_length**), presentando una diferencia significativa entre perfiles. El perfil con menor tiempo es el cinco con un promedio de 74,42 minutos, lo cual es consistente ya que como se mencionó anteriormente en su mayoría presenta cursos de práctica los cuales tienen muy poco contenido visual. Seguido del perfil dos con 198 minutos en promedio lo cual se podría relacionar con que los cursos para principiantes puede que sean más cortos. Luego los perfiles uno y tres, dando un gran salto en el cuarto perfil el cual presenta en promedio los cursos más largos con 1469 minutos. Se observó que los atributos relacionados a la cantidad de material que se ofrece en el curso como lecturas o pruebas

presentaban una distribución similar a la variable **estimated_content_length** entre los perfiles, siendo el cuarto el que presentaba los valores más altos. Las variables en las que se vio este comportamiento fueron **avg_num_published_lectures**, **num_curriculum_items**, **num_lectures**, **num_of_published_curriculum_object**, **num_published_quizzes** y **num_quizzes.** En la matriz de correlación que se mostró anteriormente se observó una fuerte correlación entre estas variables y **estimated_content_length**, es lógico que exista esta relación ya que mientras más largo es un curso más material, pruebas y lecturas este debe tener.

En cuanto a los precios se observó que existía una gran variedad en el precio promedio semanal entre los perfiles, siendo más económico el perfil dos con cursos para principiantes y más costoso el perfil tres. Sin embargo, como se mencionó anteriormente se sabe que la plataforma de Udemy vende la gran mayoría de sus cursos con grandes descuentos que varían de semana a semana. En promedio todos los perfiles terminan teniendo un precio descontado parecido oscilando entre 10$USD y 13$USD como se había observado en la fase de exploración de datos, variando el porcentaje de descuento siendo mayor para los perfiles con un precio más elevado y menor para aquello con precios más bajos, pero al final el precio con el descuento no varía significativamente entre los perfiles.

Luego se analizaron las variables relacionadas al largo de los campos de texto de los cursos. El perfil cuatro volvió a presentar valores más elevados en especial en el largo de la descripción y los objetivos del curso, siendo interesante porque estas variables no habían mostrado una correlación significativa con ninguna otra en la matriz de correlación. El perfil número 2 mostró los valores más bajos de estas variables, esto se puede deber a que los cursos para principiantes capaz requieren mejor pre-requisitos y son más cortos.

Uno de los resultados más interesantes del algoritmo es la distribución de la demanda promedio semanal entre los cursos debido a su variedad y los distintos valores en los que se encuentra cada curso. Los perfiles uno y cinco presentaron la demanda más baja con aproximadamente 12 y 13 nuevos suscriptores respectivamente, seguido del perfil dos con 22, el cuatro con 160 y por último el perfil tres con 205 suscriptores semanales. Esta diferencia demuestra que existe una dependencia o una relación entre la demanda del curso y las características de este. Comparando la demanda semanal con el número de suscriptores totales por curso se observa que no necesariamente tener mas suscriptores lleva a tener una mayor demanda, como es en el caso de los perfiles tres y cuatro en el número de suscriptores del segundo triplica al primero, pero la demanda semanal es mayor en el perfil 3.

Los atributos **subcategoryId, estimated_content_length, avg_demand, num_subscribers, num_reviews** fueron de las variables que resultaron más relevantes para el estudio y presentaron una mayor distinción entre perfiles. Con estos resultados se puede proceder a la fase de pronóstico de demanda utilizando los perfiles obtenidos durante la fase de clustering para aportar información relevante en el pronóstico.

5. Etapa de Investigación

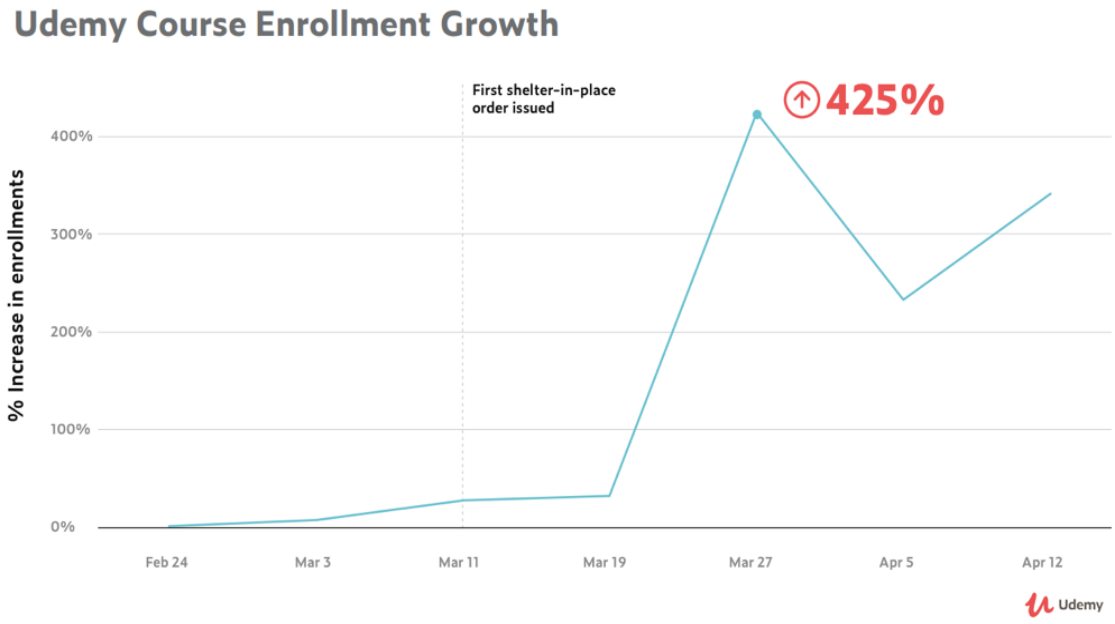
Para optimizar el modelo de precios de la plataforma era importante analizar el crecimiento del sector al igual que el nivel de competencia y sustitución que se encuentra hoy en día en el mercado en otras plataformas similares para observar como afectan estos aspectos a la demanda de los suscriptores. Igualmente se estudiará la perceptibilidad de los cursos en base a las tendencias del mercado de cursos de 'It y Software', tomando en cuenta el impacto que tendría los cursos que pasan de moda o son sustituidos por versiones más modernas en la política de precios.

5.1 Crecimiento y competencia

La compañía '*Research and Markets'* la cual se encarga de conectar a las empresas con los conocimientos y el análisis del mercado que estas requieren para tomar decisiones inteligentes, realiza investigaciones y reportes sobre el crecimiento de muchos de los sectores más relevantes del mercado. En enero de 2021 publicaron el reporte de 'GCC Massive Open Online Course (MOOC)' o Cursos En-Linea Masivo y Abierto en el que analizan el potencial de crecimiento del mercado de cursos en línea enfocados para grandes números de estudiantes que se encuentran geográficamente dispersos, en el que se pronostica que durante el 2021 y 2026 el sector va a tener un crecimiento esperado de 17,23% de la tasa de CAGR, tasa compuesta de crecimiento anual. En el reporte se establece que los cursos online como los ofrecidos en la plataforma de Udemy son escalables para dar clases en línea de forma masiva al poder ser cursados por cualquier individuo que lo desee desde cualquier región en cualquier momento, permitiendo librarse de las limitaciones de la educación típica y formal del tiempo y el espacio. La educación online reduce costos de dar cursos y se convirtió especialmente conveniente dada la situación mundial en el último año con el COVID-19 el cual tuvo un alto impacto en el sector. Para abril de 2020 debido a la pandemia millones de instituciones se vieron obligadas a comenzar a dar sus clases en línea. Para marzo de 2020 la plataforma de Udemy

observó un incremento de 425% de suscriptores con respecto al mes anterior según un reporte realizado por Udemy al igual que un incremento significativo en los cursos de tecnología que eran tendencia.

Gráfico 21. Incremento de suscriptores de Udemy



Fuente: Udemy

Tabla 6. Top tendencias tecnologías según Reporte de Udemy

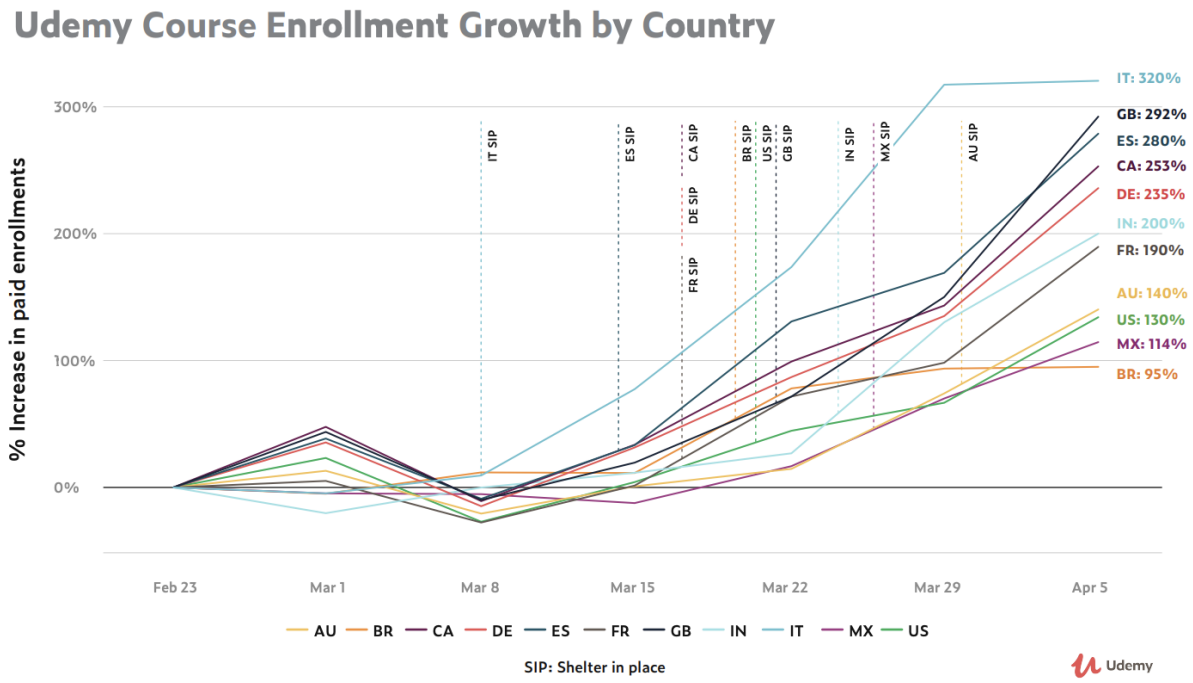
Growth in our top 10 skills

Tech Skills	
1. TensorFlow	↑ 46%
2. Chatbots	↑ 60%
3. Microsoft Azure	↑ 31%
4. OpenCV	↑ 40%
5. Neural Networks	↑61%

Fuente: Udemy

La pandemia ocasionó que millones de personas alrededor del mundo tuvieran que cambiar su estilo de vida incrementando la demanda por los cursos online en muchos países. Como se puede apreciar en el próximo gráfico desde el 23 de febrero al 5 de abril de 2020 Italia tuvo un incremento de suscriptores del 320% y España del 280%.

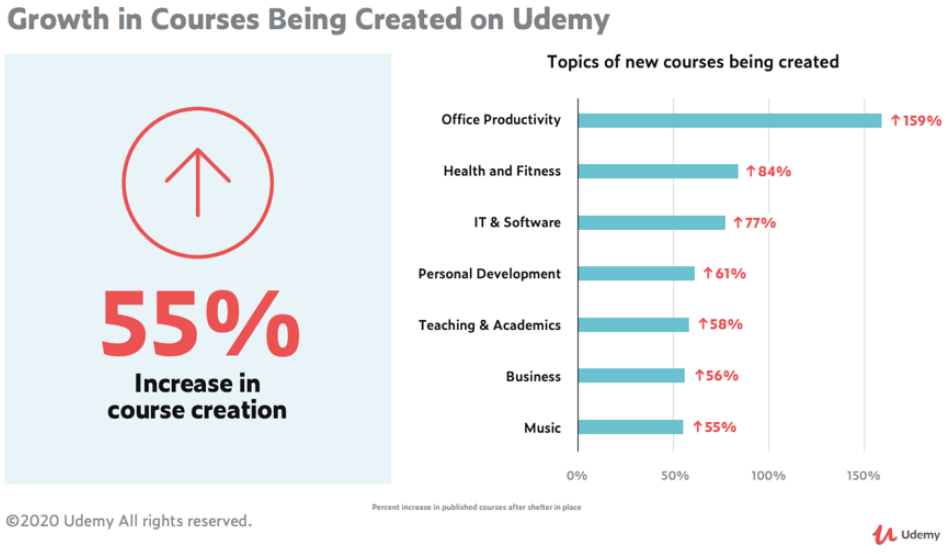
Gráfico 22. Incremento de los suscriptores de Udemy por país



Fuente: Udemy

No solo aumentó el número de suscriptores que comenzaron a utilizar la plataforma, aumentó en 55% el número de cursos creados durante este mes y en 77% la cantidad de cursos creados de 'IT & Software'.

Gráfico 23. Crecimiento de los cursos creados en Udemy



Fuente: Udemy

El crecimiento de la plataforma fue masivo y la compañía cree que aunque este crecimiento fue ocasionado por la pandemia el mismo va a continuar y que la educación online se sostendrá a lo largo del tiempo adaptándose como la nueva forma de educación.

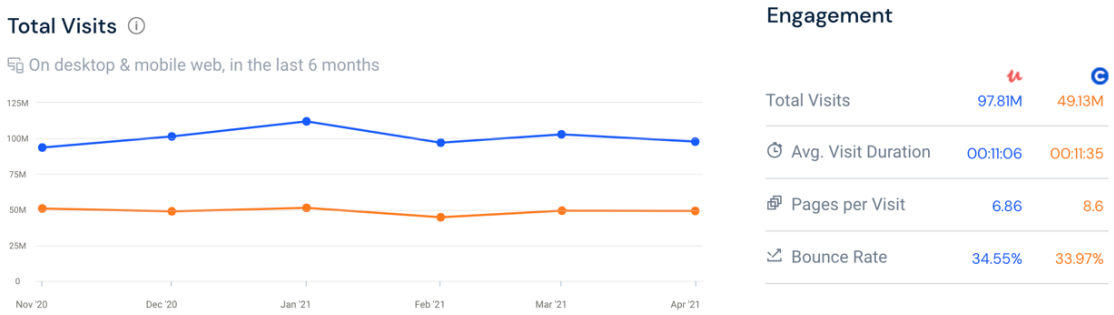
Este crecimiento no fue exclusivo de Udemy, impactó a todos los jugadores de este sector y trajo nuevos jugadores que anteriormente no daban cursos online tuvieron que adaptarse sea por el cambio tecnológico o por la pandemia. En el marco competitivo según el reporte de MOOC los 5 principales jugadores del sector son:

1. Coursera Inc.
2. Edx Inc.
3. Udacity Inc.
4. Canvas Networks Inc.
5. Udemy Inc.

Entre las principales diferencias entre las plataformas de Coursera y Edx con Udemy esta primero los instructores, estas dos plataformas cuentan con contratos con reconocidos instructores de las instituciones y compañías más reconocidas como Harvard, MIT, Microsoft o Google y también cuentan con cursos más largos para realizar especializaciones o maestrías. Mientras que Udemy no cuenta con cursos con colaboraciones con universidad ni instituciones, cualquier persona puede dictarlos solo con registrarse como instructor, brindándole independencia a los instructores de manejar su contenido siendo cursos más cortos y en muchos casos prácticos. En cuanto a los precios, en promedio los cursos de Udemy son más económicos que en las otras dos plataformas, en Udemy los precios oscilan entre 10 y $200 USD pero la mayoría de sus cursos no superan los $20 USD. En la plataforma Edx los precios oscilan entre $25 y $300 USD respectivamente y en Coursera se encuentra cursos certificados desde $39 USD pero esta también cuenta con la opción de una suscripción anual de $399 USD que te permite tomar todos los cursos que desees. Luego en cuanto a las certificaciones en las dos plataformas competidoras ofrecen certificados que se encuentran acreditados por la institución que dicta el curso. En relación a accesibilidad las tres plataformas poseen su página web y aplicación móviles y en cuanto a la variedad de tópicos todas las plataformas presentan una gran variedad, solo Edx presenta cursos enfocados a ciencia.

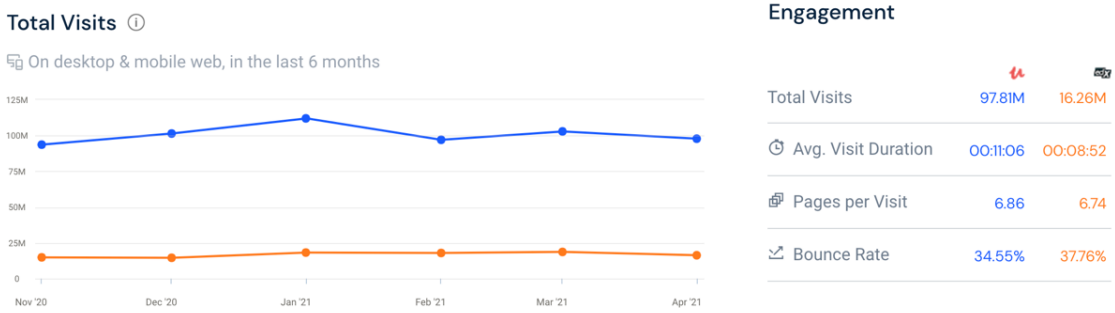
Se comparó la cantidad de visitas estimadas con la plataforma de SimilarWeb que han recibido las plataformas competidoras con Udemy y se obtuvo que en ambos casos la vistas a esta última eran mayores.

Gráfico 24. Comparación entre las visitas de Udemy vs Coursera



Fuente: Generado con la plataforma de SimilarWeb

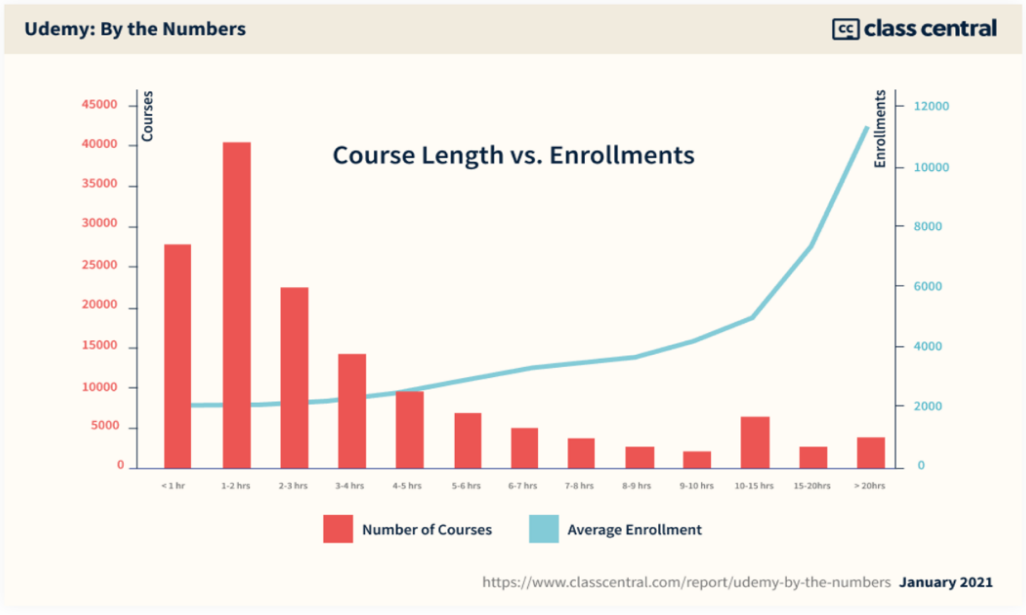
Gráfico 25. Comparación entre las visitas de Udemy vs Edx



Fuente: Generado con la plataforma de SimilarWeb

En el articulo '*157K Courses, 425M Enrollments: Breaking Down Udemy's Massive Catalog* ' un análisis realizado para la página The Report by Cetral Class escrito po Dhawal Shah comenta que la plataforma de Udemy cuenta con el catalogo mas grande de cursos en línea en el mundo con más de 157,000 cursos hasta enero de 2021 con aproximadamente 40 millones de suscriptores. La mitad de los cursos pertenecen a las categorías de Negocios o **Tecnología,** pero estas acreditan en promedio 70% de los suscriptores donde tecnología es la categoría más popular, la cual combina las categorías de Desarrollo y 'It & Software'. En este análisis crearon el siguiente gráfico con los datos de todos los cursos de Udemy para enero 2021, se puede observar que los cursos de menor duración tienden a poseer un mayor número de suscripciones.

Gráfico 26. Comparación entre la duración de los cursos y la cantidad de suscriptores.



Fuente: The Report By Dhwal Shah

En otro artículo escrito igualmente por Dhwal Shah en la misma plataforma llamado '*Udemy vs Coursera: Comparing Online Learning Giants that Might IPO in 2021*' se obtuvo la siguiente tabla Como se puede observar que en enero de 2021 Coursera poseía casi el doble de suscriptores que Udemy pero este tenía un catálogo de cursos que es aproximadamente 2400% mayor que el de Coursera.

Tabla 7. Comparaciones de aspectos claves entre Udemy y Coursera

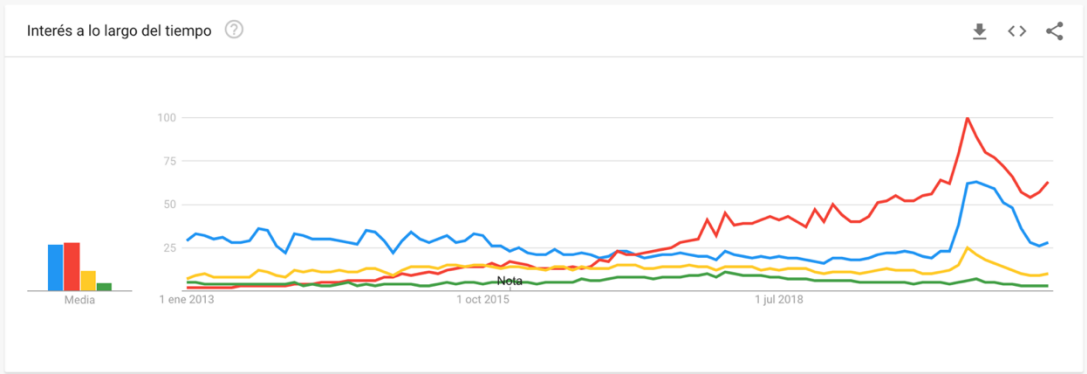
	Coursera	**Udemy**
Users	76 million	40 million
2020 Revenue	>> $200 million	$400 million
Funding	$443 million	$296.5 million
Valuation	$2.5 billion	$3.25 billion
Courses	6,500	157,000
Enrollments	170+ million	425+ million
Enterprise Customers	2,300+	7,000+

Coursera vs Udemy: By the Numbers

Fuente: The Report By Dhwal Shah

Para evaluar el reconocimiento de cada una de las plataformas se analizó las búsquedas realizadas en Google Trends de los últimos años. Como se puede observar Coursera fue la plataforma más buscada hasta que un 2017 Udemy la superó y se ha mantenido de primero hasta el día de hoy. En abril de 2020 se puede evidenciar el pico en la cantidad de búsquedas en todas las plataformas coincidiendo con el comienzo de la pandemia.

Gráfico 27. Búsquedas diarias en google de Udemy y sus competidores en los últimos años.

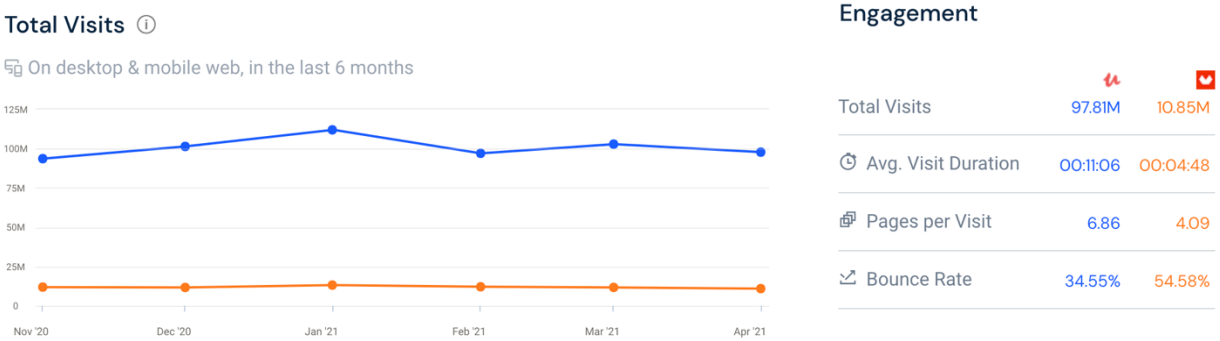


Fuente: Google Trends

Por otro lado, Udemy cuenta con competidores que ofrecen cursos sin certificado y en el mismo rango de precios como Domestika o Skillshare.

En el caso de Domestika la plataforma vende se cursos de una forma muy similar a Udemy con un precio entre los $10 y $40 USD. Los cursos de Domestika en su mayoría se enfocan en temas más relacionados a tópicos árticos y creativos. En cuanto a las visitas la plataforma de Udemy presenta muchas más visitas que Domestika.

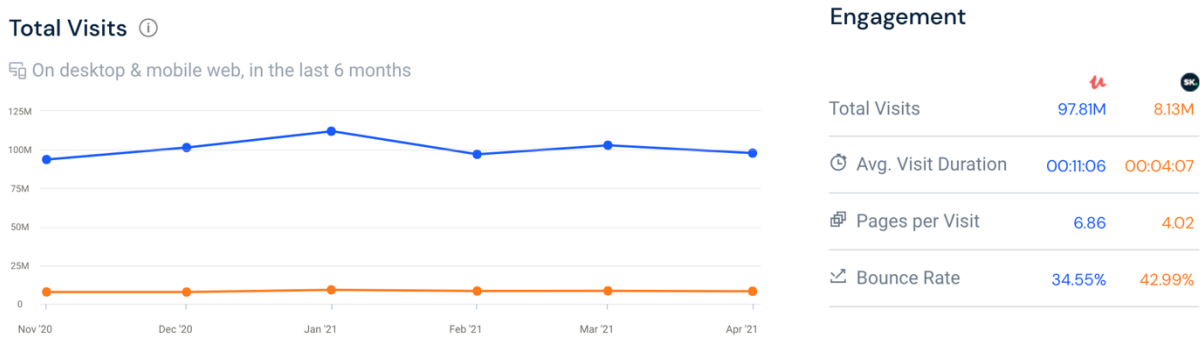
Gráfico 28. Comparación entre las visitas de Udemy vs Domestika



Fuente: Generado con la plataforma de SimilarWeb

Por otro lado, está la plataforma de Skillshare vende cursos similares a los de Udemy con la diferencia que no cobra por curso si no cobra por una suscripción que puede ser mensual o anual, el precio de la suscripción mensual hoy en día es de $19 USD. Udemy supera en casi 6 veces la cantidad de cursos ofrecidos en Skillshare y también lo supera en visitas diarias en la plataforma.

Gráfico 29. Comparación entre las visitas de Udemy vs Skillshare



Fuente: Generado con la plataforma de SimilarWeb

Se concluyo que lo que diferencia a la plataforma de Udemy del resto es que ofrece una gran variedad de cursos que siempre se están actualizando según las tendencias del mercado los cuales generalmente son más económicos y más prácticos que los ofrecidos por sus competidores.

5.2 Tendencias del Mercado

Una vez analizada la competencia se estudió la perceptibilidad de los cursos en base a las tendencias del mercado de cursos de 'It y Software'. El mundo de la tecnología es un mundo muy dinámico que se encuentra en cambio constante, nuevas tecnologías surgen todos los días que requieren ser aprendidas y tecnologías pasan de moda quedando en el olvido. Estos cambios en la tendencia son difíciles de predecir ya que en la mayoría de los casos son rápidos y se pueden ver afectados por factores externos de todo tipo.

En el artículo de MOOC mencionado en el capítulo anterior establecía que el segmento de la tecnología es uno de los segmentos más involucrados en este mercado influenciado por la necesidad de aprender las últimas tecnologías del mercado para las distintas industrias. En el segmento de la tecnología las temáticas principales de los cursos son ciencias de la

computación, análisis y estadística de tecnología e información, ciberseguridad y aprendizaje automático, según el artículo el aprendizaje automático y la inteligencia artificial están entre los temas en los que se pronostica un crecimiento de la demanda.

Con la finalidad crear una base de datos con las tendencias de las temáticas relacionadas con 'IT & Software' se buscaron reportes de fuentes confiables sobre las tendencias del momento. Estos datos nos permitirán observar el efecto de la tendencia del mercado en los cursos.

Future Today Institute es una consultora que investiga, modela y crea prototipos de oportunidades y riesgos futuros. Como consultores líderes en gestión de prospectiva estratégica y futurología para equipos de liderazgo ejecutivo en todo el mundo, la investigación aplicada basada en datos de FTI revela tendencias y calcula cómo alterarán los negocios, el gobierno y la sociedad. Esta consultora realiza reportes periódicos sobre las tendencias en el mundo tech y aconseja utilizar esta información para analizar el impacto de las tendencias de las compañías. Según la consultora, desde los eventos del año pasado relacionados a la pandemia se observaron cambios significativos en las tendencias. En el último reporte, el número catorce de este año, se analizaron más de 500 tendencias tecnologías y relacionadas a la ciencia de las cuales destacaron 12 tópicos que son tendencia hoy en día. Estos reportes fueron creados para ayudar a las compañías a adaptarse a los cambios en de las tendencias y a tener éxito, según la consultora ahora más que nunca las compañías deben examinar el potencial del impacto cercano y a largo plazo que tienen las tendencias del mundo de la tecnología.

Tabla 8. Tópicos tendencia según Future Today Institute

Tópicos Tendencia FTI	
Inteligencia artificial	Gobierno + Política
Puntaje + Reconocimiento	Privacidad + Security
Nuevas realidades	Blockchain, NFTs, Fintech
Trabajo, Cultura + Juego	5G, Robost, Movilidad
Salud + Medicina	Energía, Clima, Spacio
Electrónica de consumo	Biología Sintética, CRISPR, AgTech

Según el reporte las tendencias en el mundo de la tecnología son nuevas manifestaciones que representan las colisiones de nuevos desarrollos y se forman constantemente sobre muchos años. Estas no necesariamente siguen un camino lineal desde la franja hasta la corriente principal y surgen de una integración de las macro fuerzas y una señal. Donde una macro fuerza representa las incertidumbres externas sobre las que no se tiene control pero que influyen directamente sobre el futuro como lo son la educación, la economía o el gobierno.

En otro artículo llamado 'Top Trending Technologies in 2021 You Should Know About' escrito por Rohit Sharma en la plataforma de UpGrad se realizó un análisis sobre las principales tendencias de tecnología para marzo 2021, entre la lista se encontraron los siguientes tópicos:

Tabla 9. Tópicos tendencia según Future Today Institute

Tópicos Tendencia UpGrad 2021	
Inteligencia artificial	Computación Edge
Aprendizaje automático	Blockchain
Ciencia de datos	5G
Desarrollo Full Stack	Ciberseguridad
Robótica y Automatización de procesos	Automatización de procesos

Por último la plataforma de Udemy tiene un sector llamado' Udemy for Business' el cual se enfoca en ofrecer programas de entrenamiento para empresas que quieran entrenar a sus empleados. Para este sector Udemy realiza un reporte anual sobre las tecnologías que fueron más demandadas para ayudar a sus empleados y clientes. En el reporte de 2020 las top 10 tecnologías más demandadas fueron:

1. Tensorflow
2. Chat Box
3. Microsoft Azure
4. OpenCV
5. Redes Neuronales
6. Linux
7. Ethereum
8. Splunk
9. QGIS
10. Kotlin

Igualmente, esta rama de la empresa les brinda a las compañías una selección especial de 5.500 cursos diferenciados de los 155.000 cursos que tiene en total la plataforma. Donde se destacan los cursos más relevantes y de alta calidad del mercado utilizando una serie de criterios, incluidas las calificaciones de los cursos, la participación de los usuarios y la demanda de todos los clientes. Este es un atributo que va a ser interesante analizar en el pronóstico de la demanda.

Se tomaron todas las principales palabras claves del reporte de FTI de cada uno de los tópicos y del artículo mencionado y se colocaron en google Trends para obtener las búsquedas relacionadas con estos temas tendencias hoy en día. Se descargaron todos los datos creando una base de datos con todas las palabras clave de las tecnologías tendencia, esta será utilizada en la etapa de pronóstico de demanda para analizar el impacto de las tendencias tecnológicas.

6. Etapa de Modelar la Demanda

Una vez culminada la etapa de Clustering y de Investigación se comenzó a crear un modelo para estimar la demanda de cada uno de los cursos.

6.1 Limpieza y nuevos atributos

Igual a como se hizo en la etapa de clustering a partir de la variable del número de suscriptores (**num_subscribers)** se calculó la demanda semanal para cada curso y en este caso se filtraron los datos de forma de sólo analizar los cursos de los que se tiene información desde el comienzo del periodo de recolección. Se limpio el data set de todos los valores nulos igual que antes y luego se procedió a crear nuevas variables para tomar en cuenta algunos de los factores investigados en el capítulo anterior.

Para analizar las tendencias del mercado se optó por la creación de tres nuevas variables en nuestro dataset que luego serían utilizadas para analizar su impacto sobre la demanda.

La primera variable creada fue **num_trendKeyWords** la cual representa la cantidad de palabras clave de la base de datos descrita en el capítulo anterior de las tendencias del momento que contiene cada curso en su título. Esta variable permite evaluar si un curso es tendencia o no según los reportes investigados.

Luego se creó **is_trending_topic** la cual es una variable binaria que evalúa la tendencia de un curso comparando el título de los primeros 50 cursos más demandados la semana anterior con los cursos ofrecidos esta semana. Es decir, se creó una lista de los 50 cursos más demandados cada semana en nuestro dataset y se extrajeron las palabras claves de los títulos que fueron tendencia esa semana. Luego la semana siguiente se evaluó que cursos contenían las palabras clave de los cursos más demandados en la semana anterior, en caso de contener alguna de las palabras claves en su título el curso tenía la variable de **is_trending_topic** como verdadera. Esta variable tiene como finalidad captar los cambios en las tendencias más recientes.

Por último se creó la variable de **num_courses_per_topic**, en el que se categorizó cada curso según las tecnologías principales en el mercado y luego se contabilizó la cantidad de cursos en cada categoría. Se espera que esta variable evalué la competencia entre los cursos de la misma temática, ya que al haber más cursos del mismo tema existe una mayor competencia

y también podría significar un crecimiento en la tendencia del tópico. Los tópicos principales se pueden observar en la siguiente tabla:

Tabla 10. Principales tópicos del conjunto de datos

TÓPICOS	% CURSOS
CIBER SECURITY	4,83 %
CLOUD PRACTITIONER	4,77 %
MICROSOFT	4,20 %
AWS	3,61 %
LINUX	3,30 %
PYTHON	3,00 %
TESTING	2,65 %
SAP	2,62 %
AZURE	2,49 %

Luego se analizo como variaba la demanda entre los cursos dentro de los clusters y se observó que se cumplía una especie de Ley de Pareto. El principio de Pareto establece que el 20% del esfuerzo produce el 80% de los resultados, en este caso se observó en los clusters que una minoría de los cursos alrededor del 20% eran los más demandados. Para esto se creó una variable de **cuantil** en las que se categorizó el rango en el que se encuentra cada curso con respecto a la demanda promedio semanal. Dependiendo del cluster y su distribución de la demanda semanal se crearon cinco o seis categorías por cada uno con los datos recolectados excluyendo en cada caso los datos con lo que se iba a validar el modelo para evitar 'Data leakage' debido a que se iba a realizar *cross-validation* o validación cruzada. Este proceso se explicará más adelante.

Primero se creó la variable **cuantil** para todos los datos del conjunto para analizar su distribución general y se probaron distintas posiciones de los cuantiles hasta conseguir la más óptima. Como se puede observar aproximadamente 20% de los datos tienen en promedio una demanda semanal mayor a 30 confirmando la teoría de una distribución Pareto.

Gráfico 30. Histograma de la demanda de todo el data set

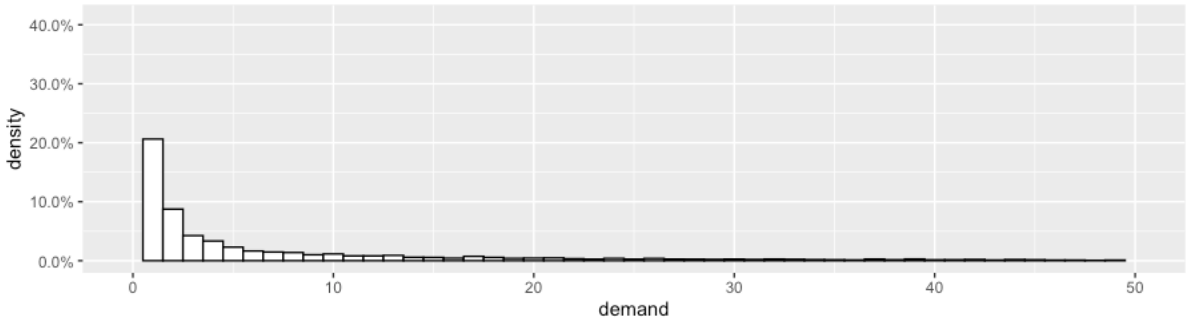


Tabla 11. Cuantiles con todos los datos

Cuantil (Q)	Rango de Demanda Semanal	% Cursos
1	30 ≤ Q	10,7 %
2	10 ≤ Q< 30	9,25 %
3	5 ≤ Q < 10	6,60 %
4	1 ≤ Q < 5	24,9 %
5	Q < 1	48,5 %

Se realizó este procedimiento para cada uno de los clusters probando distintas posiciones para los cuantiles colocándolos en los puntos más interesantes corriendo regresiones lineales para observar qué distribución resultaba más significativa para cada uno de los casos.

6.2 Modelo y Validación

Una vez que ya estaba procesado el conjunto de datos con las nuevas variables se procedió a crear una serie de modelos de regresión lineal con la función lm() de R con el que se evaluó el impacto que tenían las variables sobre la demanda.

Según el libro de 'Forecasting: Principles and Practice 'escrito por Rob J. Hidman y George Athanasopuolos el primer paso para realizar un pronóstico es definir cual es el problema que se desea abordar. Es este caso se desea pronosticar la demanda semanal de cada curso y analizar cuáles son los aspectos que tienen una mayor influencia sobre la misma para optimizar los servicios ofrecidos por la plataforma. El segundo paso consiste recopilar información por lo que se utilizó la data recolectada durante tres meses seguido de un análisis exploratorio de los

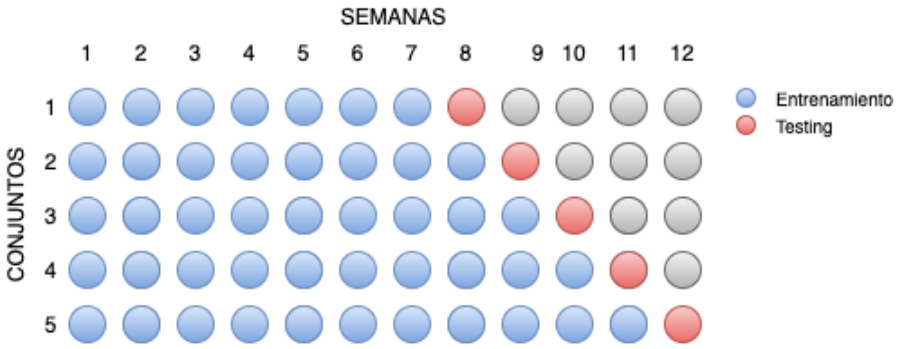
datos para observar si existe algún patrón, tendencia o generar una hipótesis de cuáles pueden ser nuestras variables más relevantes.

El cuarto paso del proceso para definir un pronóstico consiste en seleccionar el modelo que se va a utilizar y ajustar sus variables, en este caso se optó por realizar el análisis con modelos de regresión lineal tomando la demanda como la variable a pronosticar y el resto de los atributos como las variables predictoras. Del análisis exploratorio se observó que la demanda de los cursos con descuento y sin descuento presentaban un comportamiento diferente por lo que se evaluaron dos regresiones separadas para estos casos. Igualmente se generó un modelo de regresión por clúster para los cursos con descuento para observar las diferencias en el impacto de la demanda en cada uno de los conjuntos. Inicialmente se decidió correr una regresión para cada uno de los casos mencionado con todos sus atributos para realizar un paneo de cuáles eran las variables más relevantes de cada conjunto. Para ajustar los modelos y seleccionar cuáles eran las variables más relevantes se corrieron múltiples veces cada una de las regresiones descartando las variables que no resultaban significativas y comparando los modelos con distintas combinaciones de los atributos en función del coeficiente de determinación y el error estándar residual para así conseguir el modelo más óptimo. El coeficiente de determinación indica cual es la proporción de la variabilidad de la variable independiente que está siendo explicada en el modelo. Por otro lado, el error estándar residual está relacionado al tamaño del error promedio del modelo, por lo que se comparó los modelos en base a su R^2 quedándose con el que tenía el valor más cercano a uno y el menor error estándar residual.

El último paso es el de implementar y evaluar el modelo, una vez que el modelo ha sido seleccionado y estimado se optó por utilizar una validación cruzada para series de tiempo tipo ‘rolling’ la cual es una versión más sofisticada de la validación cruzada tradicional. La técnica de validación cruzada es una de las técnicas más aceptadas en el mundo de Aprendizaje Automático para validar y evaluar los modelos, esta consiste en tomar muestras al azar de los datos disponibles dividiendo los mismos en un conjunto de entrenamiento con el que se entrena al modelo y un conjunto de prueba con el que se validan los resultados obtenidos. Esta forma de hacer validación cruzada no puede ser utilizada para series de tiempo ya que al separar los datos de forma completamente al azar se estaría pronosticando con datos del futuro valores del pasado lo cual no tiene ningún sentido. Por lo tanto, para validar los modelos se utilizó una versión de validación cruzada la cual consiste en seleccionar una serie de conjuntos de prueba, cada uno de los cuales consta de una única observación por registro. El conjunto de entrenamiento correspondiente sólo contiene observaciones que ocurrieron antes de cada conjunto de prueba. Por lo que no se pueden utilizar observaciones futuras para realizar el pronóstico. Para los

modelos se crearon cinco conjuntos de entrenamiento con su correspondiente conjunto de prueba comenzando desde la octava semana hasta la doceava como se muestra a continuación.

Gráfico 31. Conjuntos creados para la validación cruzada



Dado que no es posible obtener un pronóstico confiable basado en un pequeño conjunto de entrenamiento, las primeras observaciones no se consideran conjuntos de prueba. Se tomó el error de raíz cuadrático medio (RMSE) de cada uno de los conjuntos para evaluar los modelos, el RMSE representa la desviación estándar de los errores de predicción.

Una vez que se tenían los modelos ajustados, estimados y validados se construyó la siguiente ecuación que incluye los factores más relevantes de los modelos:

$$Curso_i : \{D_{it}\}\,;\ \{P_{it}\}\,;\ \{S_{it}\}\,;\ \{Q_{it}\}\,;\ \{C_{it}\}\,;\ \{TT_{it}\}\,;\{K_{it}\}\,;\ \{U_{it}\}\,;\{NT_{it}\}\,;\{NI_{it}\}\,;\{NC_{it}\}\,;\{W_{it}\}\,;\{SC_{it}\}\,;$$

$$D_{it} = K_i \times P_{it}^{\varepsilon_p} \times S_{it}^{\varepsilon_s} \times \delta_1^{Q_{it}^1} \times \delta_2^{Q_{it}^2} \times \delta_3^{Q_{it}^3} \times \delta_4^{Q_{it}^4} \times \rho_1^{C_{it}^1} \times \rho_2^{C_{it}^2} \times \rho_3^{C_{it}^3} \times \rho_4^{C_{it}^4} \times \rho_5^{C_{it}^5} \times TT_{it}^{\varepsilon_{TT}} \times \gamma_{true}^{K_{it}} \times \gamma_{false}^{K_{it}} \times U_{it}^{\varepsilon_u}$$
$$\times NT_{it}^{\varepsilon_{NT}} \times NI_i^{\varepsilon_{NI}} \times NC_i^{\varepsilon_{NC}} \times W_t^{\varepsilon_w} \times \beta_{142}^{SC_{it}^{132}} \times \beta_{134}^{SC_{it}^{134}} \times \beta_{136}^{SC_{it}^{136}} \times \beta_{138}^{SC_{it}^{138}} \times \beta_{140}^{SC_{it}^{140}}$$

D_{it} : Demanda acumulada en la semana t del curso i,

P_{ij} : Precio total a pagar con descuentos aplicados para la semana t del curso i.

S_{it} : Porcentaje de descuento ofrecido semana t del curso i.

Q_{it} : Cuantil al que pertenece en la semana t del curso i.

C_{it} : Cluster al que pertenece en la semana t del curso i.

TT_{it} : Si el curso i era calificado como tópico tendencia (**isTrendingTopic**) en la semana t.

K_{it} : valor de la variable **num_trendKeyWords** en la semana t del curso i.

U_{it} : valor de la variable **is_in_any_ufb_content_collection** en la semana t del curso i.

NT_{it} valor de la variable **num_courses_per_topic** en la semana t del curso i.

NI_{it} : Número de instructores (**num_instructors**) en la semana t del curso i.

NC_{it} : Número de subtítulos (**num_caption**) disponibles del curso i.

SC_{it} : Subcategoría (**subcategoryId**) a la cual pertenece el curso i en la semana t-

W_{it} : Número de semana desde la cual se comenzó a recolectar información.

Debido a que se contaba con un set de datos relativamente grande y algunas variables muy sesgadas con respecto al conjunto donde existía una relación no lineal entre estas y la variable independiente se optó por realizar una transformación logarítmica sobre las variables numéricas. Esto permitió mejorar el ajuste del modelo al transformar la distribución de las variables en una curva de campana con una forma más normal. En cuanto a las variables categóricas se realizó 'one hot enconding' para evaluar el modelo para cada uno de los posibles valores. Al plantear la siguiente ecuación se tomó la primera categoría de cada una de estas variables como base. Una vez realizadas estas transformaciones se planteó la ecuación final que iba a ser ingresada en la función de lm() de R para obtener nuestros resultados.

$$log\ (D_{it}) = log\ (K_i) + \varepsilon_p log(P_{it}) + \varepsilon_s log(S_{it}) + Q_{it}^1 \times log(\delta_1) + Q_{it}^2 \times log(\delta_2) + Q_{it}^3 \times log(\delta_3) + Q_{it}^4 \times log(\delta_4)\ ) + C_{it}^1 \times log(\rho_1)\ + C_{it}^2 \times log(\rho_2)\ + C_{it}^3 \times log(\rho_3)\ + C_{it}^4 \times log(\rho_4) + \varepsilon_{TT} log(TT_{it}) + K_{it}^{true} \times log(\gamma_{true}) + \varepsilon_u log(U_{it}) + \varepsilon_{NT} log(NT_{it}) + \varepsilon_{NI} log(NI_{it}) + \varepsilon_{NC} log(NC_{it}) + \varepsilon_W log(W_{it}) + SC_{it}^{134} \times log(\beta_{134}) + SC_{it}^{138} \times log(\beta_{138}) + SC_{it}^{140} \times log(\beta_{140})$$

Esta fue la ecuación base para las dos regresiones generales de los cursos con descuento y sin descuento, luego esta fue ligeramente modificada para las regresiones de cada uno de los clusters ya que en este caso no se tomaba en cuenta la variable **C_{it}** por ser todos los datos de cada conjunto de un cluster exclusivamente. Siendo la ecuación para los análisis de cada clúster la siguiente:

$$\begin{aligned} log\ (D_{it}) = {} & log\ (K_i) + \varepsilon_p log(P_{it}) + \varepsilon_s log(S_{it}) + Q_{it}^1 \times log(\delta_1) + Q_{it}^2 \times log(\delta_2) + Q_{it}^3 \times log(\delta_3) \\ & + Q_{it}^4 \times log(\delta_4) + \varepsilon_{TT} log(TT_{it}) + K_{it}^{true} \times log(\gamma_{true}) + \varepsilon_u log(U_{it}) + \varepsilon_{NT} log(NT_{it}) \\ & + \varepsilon_{NI} log(NI_{it}) + \varepsilon_{NC} log(NC_{it}) + \varepsilon_W log(W_{it}) + SC_{it}^{134} \times log(\beta_{134}) + SC_{it}^{138} \times log(\beta_{138}) \\ & + SC_{it}^{140} \times log(\beta_{140}) \end{aligned}$$

De las regresiones lineales se va a obtener el valores de los coeficientes para cada uno de los conjuntos.

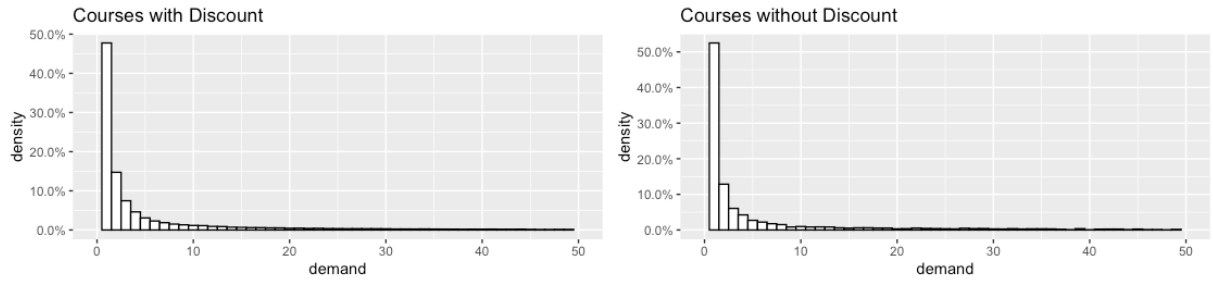
6.3 Análisis de los Resultados de los Modelos Generales

Una vez ajustados y validados los modelos se ejecutaron las regresiones con las ecuaciones finales que habían sido trabajadas y optimizadas con los datos de todas las semanas.

Los primeros dos modelos que fueron entrenados son los generales con todos los datos recolectados. El modelo 1 contiene todos los registros de los cursos con descuento cada semana, fue entrenado con el 95% de los datos con un total de 81.946 registros. El modelo 2 por el otro lado consta de todos los registros de los cursos que no ofrecían ningún tipo de descuento, este fue entrenado con el restante 5% con un total de 4325 registros.

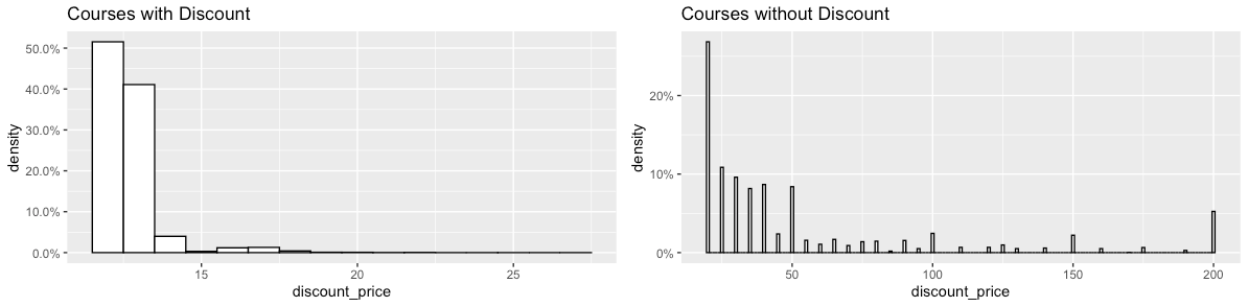
La distribución de la demanda para cada uno de los conjuntos analizados se puede visualizar en los siguientes histogramas.

Gráfico 32. Histogramas de la demanda de los registros de cursos con descuento y sin descuento.



Igualmente se analizó la distribución del precio de estos dos conjuntos y se observó una diferencia significativa. Los cursos con descuento no superan los $25 USD y más del 90% de los mismos cuestan entre $12 y $13 USD. Mientras que los cursos sin descuento pueden llegar a costar hasta $200 USD y presentan una distribución mucho más variada.

Gráfico 33. Histogramas del precio cursos con descuento y sin descuento



En cuanto a la variable de ‘cuantil’ en este caso se tomaron 5 cuantiles distribuidos de la siguiente forma:

Tabla 12. Distribución de los cuantiles en los casos generales.

Cuantil (Q)	Rango de Demanda Semanal	% Cursos Con Descuento	% Cursos Sin Descuento
1	30 ≤ Q	12,9 %	9,34%
2	10 ≤ Q< 30	10,60 %	11,10%
3	5 ≤ Q < 10	7,05 %	5,41%
4	1 ≤ Q < 5	25,1%	22,20%
5	Q < 1	44,4 %	51,9%

Se obtuvieron los siguientes resultados de estas dos regresiones. Se debe destacar que el coeficiente de **discount_price** representa el valor final a pagar por el curso.

Tabla. Resultados de los modelos generales con y sin descuento.

Tabla 13. Resultados de las regresiones de los modelos con y sin descuento.

Resultados

	Dependent variable:	
	log(demand)	
	Modelo 1	Modelo 2
log(discount_price)	0.752***	-0.079***
log(discount_percent)	0.038***	
log(num_trendKeyWords)	0.022***	-0.021
isTrendingTopic	0.176***	0.102***
log(num_courses_per_topic)	0.011***	-0.009
q1	2.847***	3.294***
q2	1.903***	2.215***
q3	1.403***	1.492***
q4	0.686***	0.791***
cluster2	-0.056***	-0.139
cluster3	-0.096***	0.136

cluster4	0.279***	0.053
cluster5	-0.074***	-0.062
log(week)	0.275***	0.263***
log(num_instructors)	0.076***	0.146***
log(num_caption)	0.302***	0.021
is_in_any_ufb_content_collection	0.668***	0.310***
subcategoryId134	-0.059***	-0.001
subcategoryId136	-0.084***	0.080
subcategoryId138	-0.069***	-0.070
subcategoryId140	-0.040***	0.008
Constant	-2.488***	0.098
Observations	81,946	4,325
R^2	0.683	0.780
Adjusted R^2	0.683	0.779
Residual Std. Error	0.871 (df = 81924)	0.649 (df = 4304)
Note:	*p**p***p<0.01	

Como se puede observar en base al p-valor todos los coeficientes del modelo uno resultaron ser significativos mientras que del modelo dos solo nueve de ellos. Los resultados completos con todos los valores de error por coeficiente se encuentran en los anexos.

Primero se analizará el primer modelo, de los resultados de la regresión uno de los coeficientes que inicialmente llamó la atención fue el de *log(discount_price)* este representa la elasticidad del precio en función de la demanda. Indicando que ante un aumento del 1% del precio la cantidad demandada aumenta en 0,752% lo cual es muy raro y solo sucede en caso muy particulares. En este caso se concluyó que esto sucedía debido a la poca dispersión que había de los cursos en cuanto al precio para los casos en los que eran ofrecidos con descuento. Como se mencionó anteriormente más del 90% de estos cursos costaban $12 y $13 USD y el tener tan poca dispersión hace que el modelo asocie relaciones erradas. Se notó que existían algunos cursos dentro del conjunto que eran más largos, más especializados y más demandados, estos contaban con un precio mayor al promedio por lo que se concluye que esto

ocasionó un sesgo en el modelo. Asociando estos cursos que tienen mayor demanda y mayor precio con que un aumento en la demanda está relacionado con un aumento de precio. Cuando en realidad esto no es así, existen muchos factores como las tendencias del momento o la competencia que limitan los precios. Como se mencionó en la etapa de investigación la plataforma de Udemy compite contra plataforma que tiene precios parecidos y se debe tratar con mucho cuidado los aumentos de precio ya que pueden ocasionar que clientes opten dejar la plataforma e irme por la competencia. Por lo que el precio para este tipo de cursos se encuentra limitado, igualmente se recomienda realizar A/B testing variando los precios ligeramente para obtener una mayor dispersión y analizar el impacto en la demanda para así poder predecir el precio óptimo con los modelos. Por otro lado, el coeficiente de la variables **discount_price** el cual indicaba el porcentaje de descuento que ofrecía un curso dio positivo indicando que antes un 1% de aumento la variable la demanda aumenta en 0,038%.

En cuanto a los coeficientes relacionados a los temas tendencia y la competencia del momento se obtuvo resultados muy interesantes. En cuanto a la variable creada **isTrendingTopic** esta evalúa la tendencia comparando los cursos más demandados de la semana anterior con la demanda de hoy en día y la variable **is_in_any_ufb_content_collection** que refleja si un curso pertenece a la colección de más de 5500 cursos que selecciona la plataforma como los cursos de mejor calidad y más demandados. Se obtuvo que, si un curso es calificado como tema tendencia, es decir que la variable **isTrendingTopic** es verdadera, entonces la demanda aumenta en un 17,6% aproximadamente. Mientras que, si un curso es calificado como parte de la colección de UFB, es decir la variable **is_in_any_ufb_content_collection** como verdadera la demanda aumenta en 66,8%. Estos son dos de los coeficientes que presentaron un mayor impacto sobre la demanda. También se tienen los coeficientes de **num_trendKeyWords** y **num_courses_per_topic** que evalúan la cantidad de palabras clave de tópicos tendencia en el título del curso y el número de cursos por tópico respectivamente. En estos casos se obtuvo que ante un aumento del 1% del **num_trendKeyWords** la demanda aumenta en un 0,022% lo que no es un cambio tan grande en comparación con los anteriores, pero en conjunto con otras variables va sumando valor. En cuanto al **num_courses_per_topic** ante un 1% de aumento la variable de la demanda aumenta en un 0,011%. Esta variable tenía como intención inicial medir la competencia entre cursos del mismo tópico, sin embargo, se obtuvo que mientras más cursos existan de un mismo tópico más demanda indicando que la hipótesis planteada anteriormente sobre que la creación de más cursos sobre un tópico puede indicar que ese tópico se está volviendo tendencia, haciendo que

no solo exista una tendencia en los usuarios a adquirir este curso si no también una tendencia en los instructores a crearlos. Todos estos coeficientes indican lo esperado, mientras un curso sea tendencia más personas desean adquirirlo y por ende aumenta la demanda.

Existe un factor relevante que no se toma en cuenta en estas predicciones que es el destaque que se realiza en la pagina principal de la plataforma recomendando los cursos mas relevantes del momento. La plataforma ofrece distintas listas de recomendaciones en su pagina entre las que se pueden destacar la lista de los cursos mas buscados por sus estudiantes, los cursos mas vendidos por tópico, los cursos tendencia del momento, entre otras. Por otro lado, también se ofrecen recomendaciones de cursos personalizadas. Es decir en base a los cursos adquiridos y buscados la plataforma genera listas de recomendación que se ajustan a cada usuario según sus necesidades. Este es un aspecto que tiene un peso importante en la demanda de los cursos ya que son los primeros cursos que visualizan los usuarios al ingresar a la plataforma. Cuando los usuarios seleccionan un tópico del curso que desean realizar lo primero que se encuentran son estas listas de recomendación.

Si bien parte del efecto de esta sección de destacado en la pagina principal es capturado por las variables de **isTrendingTopic** y **num_trendKeyWords,** ya que si estos dos atributos reflejan que un curso es tendencia la demanda tiende a aumentar notablemente. Se podría concluir que los cursos que sean calificados como tendencia probablemente estarán presentes en alguna de las listas destacas mencionadas anteriormente, lo que a su vez posiblemente brindaría una demanda potencial futura debido a la correlación entre estas variables y el factor de estar presente en la pagina principal. Sin embargo existe una parte de esta sección destacada que no esta siendo analizada en este modelo, principalmente debido a que no se posee información sobre cuales fueron los cursos en estas listas destacadas durante el periodo en el cual se recolectaron los datos y que las listas de recomendación personalizadas tampoco se encuentran disponibles. Se recomienda en un futuro incorporar estos datos al modelo para poder medir el impacto que tiene sobre la demanda las secciones de cursos destacados de la pagina principal de una forma mas directa. De forma inicial se podría añadir un nuevo atributo binario que indique si un curso se encuentra presente en la pagina principal de la plataforma, luego si se desea profundizar se podrían crear múltiples atributos que indiquen si el curso se encuentra en alguna de listas de recomendación especifica. Esto permitiría a su vez evaluar cuales de las listas destacadas tiene un mayor impacto en la demanda, cosa que no se puede determinar con los datos actuales.

Luego entre las variables más relevantes se encuentran las relacionadas a los cuantiles, donde se ve una gran diferencia entre las 5 categorías. Se tomó como categoría base el cuantil 5 el cual contaba con 51,9% de los datos. Se observa que los cursos que pertenecen al cuantil uno el cual representa a los cursos con una la demanda semanal mayor a 30 se presenta un aumento en la demanda del 284,7% y los que pertenecen al cuantil dos presentan de un aumento del 190,3%. Este porcentaje decrece a medida que se va aumentando el número de cuantil. Indicando que los cursos que logran mantener una demanda promedio semanal elevada aumentan de forma muy significativa sus ventas.

Se observó que en el caso de la subcategoría la 132 se tomó como base y los cursos con cursos de otras subcategorías tienden a bajar la demanda, por ejemplo, en el caso en el que un curso era de la subcategoría 134 la demanda decrecía en un 5,9% y para la 138 decrecía en 8,4%. Tiene sentido ya que como se había mostrado en el análisis exploratorio la subcategoría 132 era la más demandada. Por otro lado, se tiene la variable relacionada a la semana, esta indica ante un aumento del 1% la demanda aumenta en 0,275%. Se interpreta que la demanda tiene una tendencia a aumentar con el tiempo.

Por último, se tienen los coeficientes relacionados a los clusters de los cuales se observó una diferencia que resultó interesante analizar. Se tomó como base el cluster 1 y se obtuvo que si un curso pertenecía al cluster 2, 3,o 5 su demanda tendía a decrecer en 5,6%, 9,6% y 7,4% respectivamente, mientras que si el curso pertenece al cluster 4 su demanda tendía aumentar en un 27,9%. Debido a estas diferencias se decidió correr las regresiones en cada clusters de forma separada para observar en que difería el impacto de las variables a la demanda en cada uno de los casos.

Luego al analizar el modelo dos se observó que en este caso la elasticidad del precio en función de la demanda si es negativa indicando que ante aumento del 1% del precio la demanda decrece en un 0,079%. En este caso el modelo se entrenó con datos que contaban con una mayor dispersión de la variable de precio. Sin embargo, se puede observar que la elasticidad es muy cercana a cero indicando que se tiene una elasticidad casi inelástica, lo que implica que ante un cambio en el precio la demanda se mantiene casi constante. Igual que en el modelo anterior la demanda se encuentra igualmente muy afectada por las tendencias y la competencia.

En cuanto a coeficientes de las variables relacionadas a las tendencias del momento se obtuvo que estas continuaban teniendo un gran impacto, en este caso si un curso era calificado

como tópico tendencia (**isTrendingTopic**) la demanda aumentaba en un 10,2%, mientras que si era seleccionado en la colección de cursos UFB de Udemy esta aumentaba en un 31%.

Igualmente, las variables relacionadas con los cuantiles que categorizaban el rango en el que se encontraba la demanda semanal resultaron muy significativas y de alto impacto. Para los cursos pertenecientes al cuantil 1,2,3 y 4 la demanda tiende a aumentar en un 329,4%, 221,5%, 149,2% y 79,1% respectivamente, tomando los cursos pertenecientes al cluster 5 como caso base. También se noto que en este el impacto de añadir un nuevo instructor al curso era de casi el doble del modelo anterior, aumentan en 14,6% la demanda con la nueva adición.

En cuanto a las variables que no resultaron significativas en este modelo se tienen los clusters, lo cual tiene sentido ya que los cursos que son ofrecidos sin descuento en la plataforma son cursos que se distinguen mucho del resto. Estos suelen ser más largos, más especializados y en su mayoría presentan una distribución de la demanda muy distinta. Por esto se optó por hacer el análisis por clusters solo con los cursos con descuento. Igualmente, en este caso en cuanto a las subcategorías no se noto ningún cambio significativo entre las mismas.

En la validación cruzada de estos dos modelos se obtuvo que en promedio el modelo uno tenía un error de raíz cuadrático medio de **498,31** y el modelo dos presenta un error de **31,03**. Tiene sentido que el error sea mayor en el modelo uno porque se cuenta con muchos más datos y el modelo dos fue entrenados con datos que presentan una mejor dispersión. En cuanto al coeficiente de determinación ajustado se obtuvieron buenos resultados, 0,683 en el primer modelo y 0,779 en el segundo.

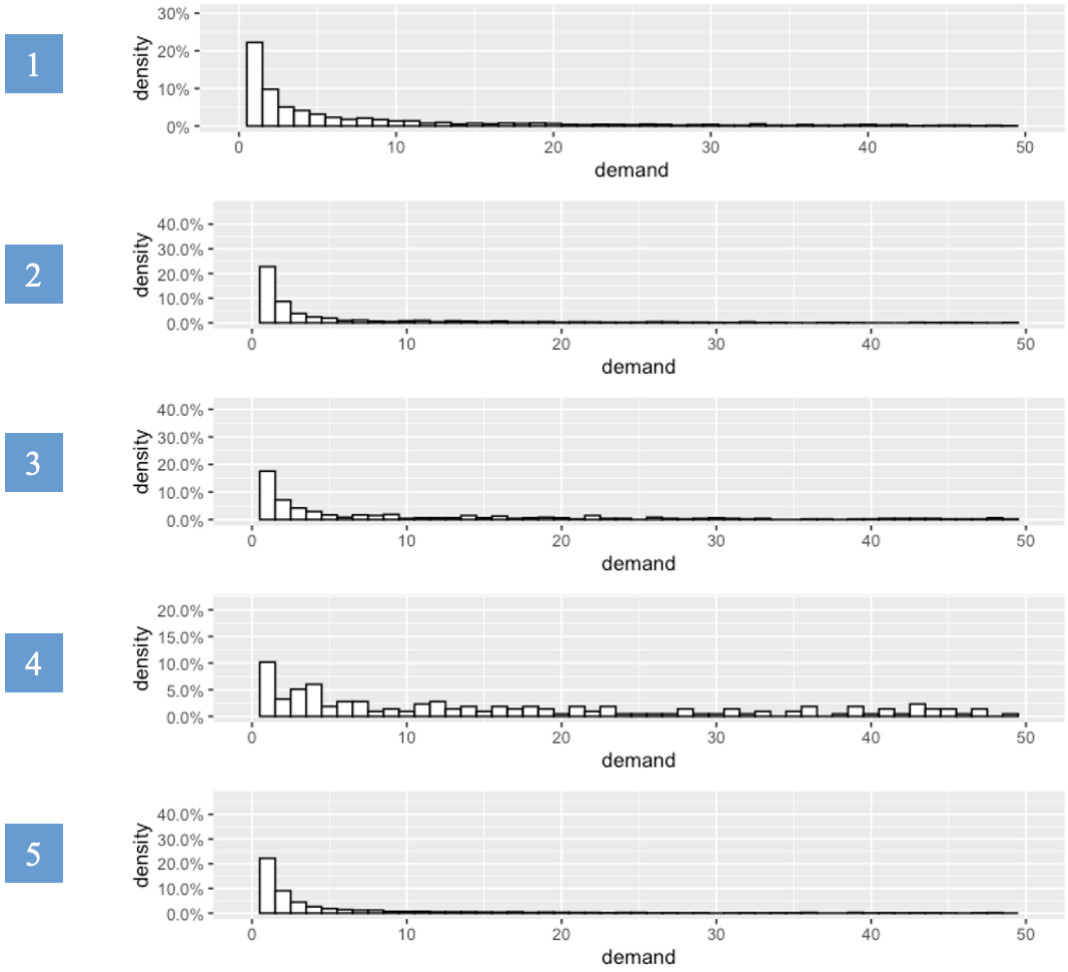
6.4 Análisis de los Resultados de los Modelos por Cluster

Como se mencionó anteriormente, al observar que en el conjunto de los cursos con descuento (Modelo 1) resultaron muy significativas las variables relacionadas a los clusters creados se decidió correr un modelo para cada uno de estos. Los 81,946 registros se distribuyeron de la siguiente manera entre los cinco modelos:

- Modelo del cluster 1: 13.488 (16,45%)
- Modelo del cluster 2: 30.299 (36,97%)
- Modelo del cluster 3: 21.596 (26,35%)
- Modelo del cluster 4: 6,.531 (7,97%)
- Modelo del cluster 5: 3.622 (4,42%)

La distribución de la demanda en cada uno de los clusters se puede visualizar en los siguientes gráficos.

Gráfico 34. Histograma de la distribución de la demanda por cluster.



La variable de ‘cuantil’ en este caso en algunos cluster tuvo 5 categorías y en otros 6 como se muestra en la siguiente tabla. Se aseguró que la distribución de los cursos en cada categoría contará con una cantidad de registros significativos para evaluar los puntos más interesantes de la distribución.

Tabla 14. Distribución de los cuantiles por cluster.

Cluster	Cuantil (Q)	Rango de Demanda Semanal	% Cursos
1	1	$30 \leq Q$	9.50 %
	2	$10 \leq Q < 30$	11.4 %
	3	$5 \leq Q < 10$	9.57 %
	4	$2 \leq Q < 5$	15.0 %
	5	$0 \leq Q < 2$	54.5 %
2	1	$15 \leq Q$	10.5 %
	2	$5 \leq Q < 15$	7.71 %
	3	$2 \leq Q < 5$	10.6 %
	4	$1 \leq Q < 2$	13.2 %
	5	$0 \leq Q < 1$	58.0 %
3	1	$1000 \leq Q$	7.09 %
	2	$200 \leq Q < 1000$	11.4 %
	3	$50 \leq Q < 200$	8.01 %
	4	$15 \leq Q < 50$	10.6 %
	5	$2 \leq Q < 15$	16.8 %
	6	$0 \leq Q < 2$	46.1 %
4	1	$400 \leq Q$	9,17 %
	2	$100 \leq Q < 400$	15,5 %
	3	$30 \leq Q < 100$	24,9 %
	4	$10 \leq Q < 30$	21,2%
	5	$2 \leq Q < 10$	8,02 %
	6	$0 \leq Q < 2$	21,2 %
5	1	$10 \leq Q$	13.1 %
	2	$5 \leq Q < 10$	5.2 %
	3	$3 \leq Q < 5$	5.32 %
	4	$1 \leq Q < 3$	21,6 %
	5	$0 \leq Q < 1$	54,8 %

Los resultados obtenidos de estos cinco modelos fueron los siguientes:

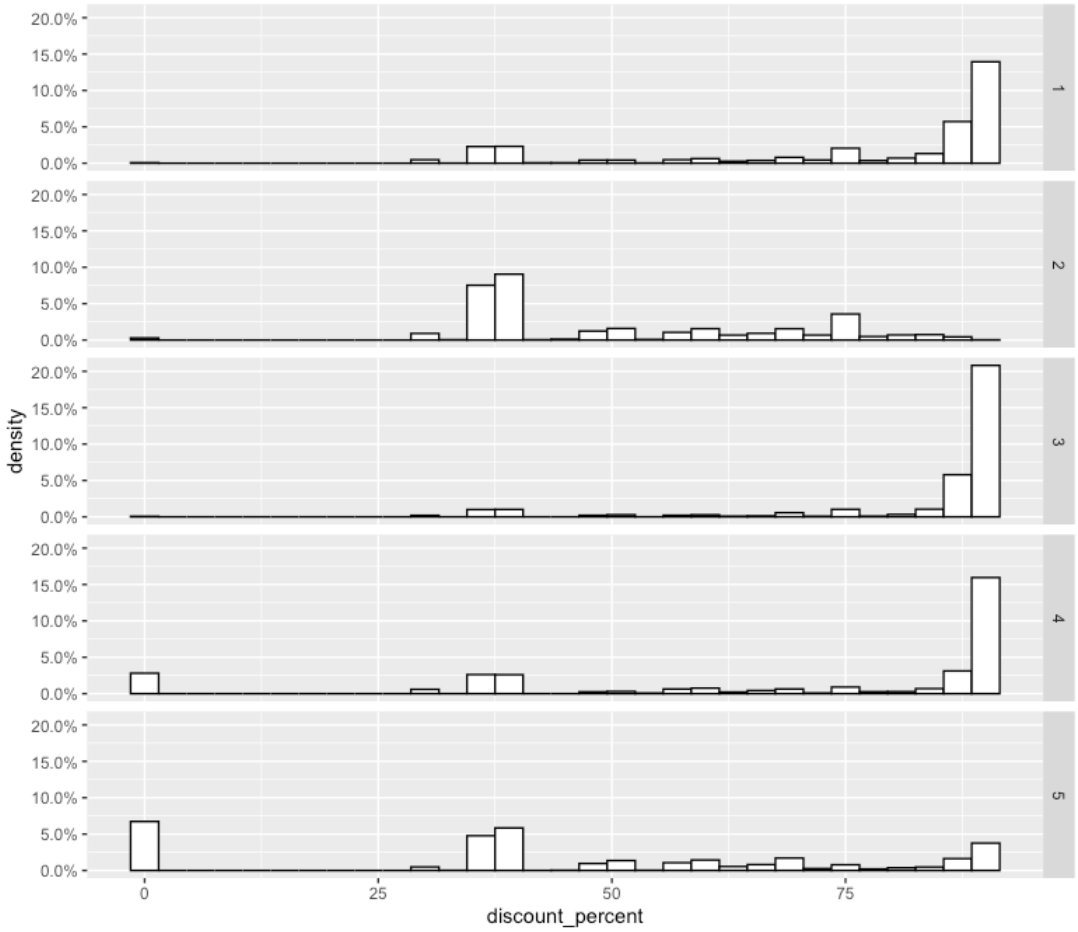
Tabla 15. Resultados de las regresiones de los modelos por cluster.

Resultados

	Dependent variable: log(demand)				
	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
log(discount_price)	0.722***	0.690***	0.616***	1.321***	0.465***
log(discount_percent)	0.038***	0.027	0.377***	0.035	-0.096***
log(num_trendKeyWords)	0.014*	0.040***	0.057	0.120***	0.071***
isTrendingTopic	0.146***	-0.012	0.005	-0.044	-0.058**
log(num_courses_per_topic)	0.010***	0.007**	0.013	0.010	0.006
q1	3.061***	2.251***	3.183***	4.981***	2.212***
q2	2.041***	1.503***	2.353***	3.870***	1.491***
q3	1.383***	0.910***	2.450***	2.942***	1.090***
q4	0.810***	0.506***	1.534***	1.820***	0.573***
q5			1.005***	0.875***	
log(semana)	0.314***	0.245***	0.499***	0.857***	0.288***
log(num_instructors)	-0.134***	-0.033	0.179***	-0.072	0.522***
log(num_caption)	0.108***	0.106***	0.428***	0.142***	
is_in_any_ufb_content_collection	0.332***	0.708***	1.237***	0.425***	1.071***
subcategoryId134	-0.009	-0.006	-0.257***	-0.040	0.062*
subcategoryId136	0.058***	0.015	-0.277***	-0.114	-0.047
subcategoryId138	0.007	-0.029	-0.163**	-0.041	-0.037
subcategoryId140	0.002	0.011	-0.086	-0.071	-0.288***
Constant	-2.357***	-2.151***	-3.907***	-4.673***	-1.159***
Observations	32,936	23,460	7,105	3,950	14,495
R^2	0.741	0.621	0.552	0.745	0.632
Adjusted R^2	0.740	0.620	0.551	0.744	0.631
Residual Std. Error	0.711 (df = 32918)	0.756 (df = 23442)	1.437 (df = 7086)	1.100 (df = 3931)	0.813 (df = 14478)
Note:	*p**p***p<0.01				

Al correr estas regresiones se observó que algunas variables tienen un mayor impacto en algunos clusters que en otros e incluso llegan a tener efectos inversos sobre la demanda. Por ejemplo, el porcentaje del descuento tiene un efecto diez veces mayor en el cluster 3 que en el resto de los cluster donde ante un aumento del 1% del porcentaje de descuento del curso la demanda aumenta en un 0,377% mientras que en el cluster cinco la elasticidad del descuento en función a la demanda dio negativa indicando que en este caso un aumento del porcentaje de descuento impacta de forma negativa en la demanda. Al analizar la distribución de los porcentajes de descuento del cluster 5 en comparación con los otros se observó que en promedio presenta descuentos menores concluyendo que los cursos de este sector no son lo que tienen la mejor respuesta ante los descuentos.

Gráficos 35. Histogramas del porcentaje de descuento en cada cluster.



En el caso de la variable **is_in_any_ufb_content_collection** esta resultó significativa para todos los clusters. La variable tiene un impacto mayor en los clusters 3 y 5 con un aumento de la demanda del 123,7% y 107,1% respectivamente en los casos en que la variable es positiva,

pero igual presenta un impacto relevante en el resto de los clusters. El coeficiente de log(**num_trendKeyWords)** varió entre 0,014 y 0,12, indicando por ejemplo que para el cluster 4 ante un aumento el 1% de la variable la demanda aumenta en 0,12%.

Se observó un comportamiento inesperado de la variable **isTrendingTopic** en donde para algunos clusters se obtuvo una elasticidad negativa con respecto a la demanda, aunque en estos casos solo fue significativo el coeficiente del cluster 5. En este caso se interpreta que para este conjunto puede haber una demanda que es menos constante a lo largo de las semanas y que tópicos de los cursos más vendidos cada semana varían de semana a semana.

En el caso de las variables de 'cuantil' se observó que se mantenía el mismo comportamiento en todos los clusters ya que a medida que aumenta el número de la categoría del cuantil disminuye el efecto sobre la demanda. Los cuantiles del cuarto cluster presentaban valores más elevados que el resto en todas las categorías, lo que tiene sentido porque en el histograma de la demanda de este cluster mostrado anteriormente se observó que existía una mayor dispersión de la demanda en este caso.

En cuanto a los coeficientes relacionados a las subcategorías se observó que en el cluster tres estos atributos tienen un mayor impacto que en los otros clusters, en donde se refleja que, si un curso pertenece a la subcategoría 134, 136 o 138 la demanda tiende a disminuir en 25,7%, 27,7% y 16,3% respectivamente. Por otro lado, en relación con el atributo semana el cual mide el paso del tiempo se observa que el cluster cinco seguido del cluster 4 son los que presentan un mayor incremento en la demanda de semana a semana. Mientras que los cluster dos y cinco son los que presentan un crecimiento más lento.

En cuanto al número idiomas en los que se encuentra disponibles los subtítulos se obtuvo que todos los cursos del cluster 5 tenían una sola opción por lo que el coeficiente era irrelevante, para el resto de lo clusters se mostró una elasticidad positiva (menos que proporcionalmente) con un valor entre 0,1 y 0,45.

En la validación cruzada de estos modelos se obtuvo el promedio que tenían del error de raíz cuadrático es el que se muestra en la siguiente tabla.

Tabla 16. RMSE de los modelos de cada Cluster.

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
51,39	409,65	1328,62	618,16	170,56

Se esperaba que el Cluster 1 fuera el mejor modelo dado que presentaba el menor error estándar residual de todos los modelos y uno de los coeficientes de determinación más altos con 0,74. Por el otro lado, el cluster 3 había presentado el error residual standard más elevado y el coeficiente de determinación más bajo, por lo que tiene sentido que este haya sido el que tuvo el mayor error de validación.

Sin embargo, se considera que en general todos los modelos entienden de forma óptima el comportamiento de los datos, en la realidad utilizado datos del día a día llegar a tener modelos con un coeficiente de determinación igual a 1 o con un error igual cero es algo que no es factible. Especialmente en un tema como el que se está tratando en el que influyen muchas acciones externas que llegan a impactar sobre la demanda de los cursos.

7. Etapa de Optimización

Una vez culminadas todas las etapas anteriores solo queda poner en práctica todos los resultados obtenidos. Se categorizaron todos los cursos en cinco perfiles durante la etapa de clustering y luego se modeló la demanda para cada uno de los cursos de estos perfiles. Analizando las diferencias entre los mismos y observando cuáles eran los aspectos clave que verdaderamente impactan sobre la demanda. Se investigó sobre la competencia de Udemy y sobre fuentes confiables que ofrecían reportes de los temas claves que eran tendencia del momento. Ahora solo falta permitir que esta información esté disponible para que los instructores de la plataforma, creando un sistema o un dashboard de recomendación.

Para los modelos de regresión se creó una función de R llamada trendingWords la cual se encarga de recolectar las palabras claves de los temas más demandados de un conjunto, con esta se calculaba el valor de la variable **isTrendingTopic**. Con esta función se puede obtener por ejemplo las palabras clave de los primero 10 cursos más demandados del cluster 1 en la semana 12:

- Obs
- Basics
- Guide
- Hero
- Linux
- Server
- Studio
- Ubuntu
- Ultimate
- Active
- Addressing
- Administration
- Amazon
- Beginner
- Cloud

- Computing
- Crack
- Cyber
- Desktop
- Google
- Helpdesk
- Interviews
- Ip
- Learning
- Livestreaming
- Microsoft
- Nginx
- Sde
- Security
- Streamlabs
- Subnetting
- Technicians
- web

Del modelo de regresión lineal del cluster 1 se obtuvo que si un curso de este cluster contenía alguna de las palabras clave de los cursos más demandados de la semana anterior su demanda podría llegar a incrementar en un 14,6%. Por otro lado, se tiene la base de datos creada con las palabras clave que son tendencia hoy en día según los reportes investigados y las búsquedas asociadas en Google Trends. Si un curso es categorizado en el perfil 1 (cluster 1) durante la doceava semana se podría ofrecer este tipo de información aconsejando colocar las palabras clave que se relacionen con su curso en el título ya que del modelo de este cluster se obtuvo que esto se reflejaba en un aumento de la demanda del curso. No solo se trata de crear un curso de calidad de Udemy si no también ofrecerles a los instructores las palabras clave más buscadas por los clientes para que los mismos puedan conseguir su curso en la plataforma y sentirse tentados a comprarlo.

Con esta función también se le puede suministrar información a los instructores de Udemy que deseen crear nuevos cursos sobre los tópicos que según los estudios realizados se encuentran entre las tendencias del momento. Esta función permite buscar las palabras claves más demandadas semanalmente por tópico lo que permite ayudar a los instructores a colocar las palabras más acertadas en sus títulos. Por ejemplo, para la semana 12 las palabras clave de los cursos más demandados del tópico de AWS son:

- aws
- certified
- architect
- associate
- solutions
- ultimate
- practice
- developer
- exam

Igualmente se puede ofrecer una información más actualizada sobre los tópicos presentes en la colección de UFB, ya que se observó que los cursos pertenecientes a esta colección

presentan un aumento importante en su demanda. Udemy suele liberar un reporte anual sobre los temas tendencia del año pero al ser la tendencia tecnológica un tema tan dinámico se propone informar a los instructores lo más seguido posible para que estos puedan continuar creando contenido de los temas más demandados del momento.

De los modelos creados en la etapa anterior se creó esta tabla con los principales aspectos en orden de prioridad a tomar en cuenta para mejorar la demanda de un curso cuando este es categorizado en alguno de los cinco perfiles.

Tabla 16. Acciones a tomar para mejorar la demanda dependiendo del cluster.

Prioridad	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
1	Utilizar palabras clave de la lista de 'Trending Topic'	Añadir subtítulos en más idiomas	Añadir subtítulos en más idiomas	Añadir subtítulos en más idiomas	Considerar añadir un instructor al curso
2	Añadir subtítulos en más idiomas	Analizar las palabras clave de la base de datos de Tópicos Tendencia	Aumentar el porcentaje de descuento	Analizar las palabras clave de la base de datos de Tópicos Tendencia	Analizar las palabras clave de la base de datos de Tópicos Tendencia
3	Aumentar el porcentaje de descuento	Aumentar el porcentaje de descuento	Considerar añadir un instructor al curso	Aumentar el porcentaje de descuento	Añadir subtítulos en más idiomas
4	Analizar las palabras clave de la base de datos de Tópicos Tendencia	-	Analizar las palabras clave de la base de datos de Tópicos Tendencia	-	.

De esta forma los instructores pueden tener acceso a información sobre la demanda, las tendencias y los aspectos que hacen que un curso tenga una mayor demanda. Suministrar esta información permite que los instructores creen contenido más actualizado para la plataforma y que se ajuste más a las necesidades de los clientes. Descubrir que aspecto como aumentar la cantidad de idiomas en lo que se encuentran disponibles los subtítulos o la presencia de ciertas palabras claves en los títulos son detalles que se pueden mejorar en los cursos sin mucho esfuerzo pero que pueden representar un aumento en la demanda.

Igualmente, el análisis y los modelos generados podrían ser utilizados para optimizar las listas de recomendaciones ofrecidas de forma destacada en la pagina principal al usuario final, es decir los estudiantes que adquieren los cursos. Mejorando y actualizando de forma mas seguida los

cursos mas relevantes y con mayor tendencia del momento en base a la investigación y el análisis realizado. Al optimizar estas recomendaciones los usuarios encontraran los últimos cursos mas eficientemente aumentando así la demanda de los mismo, satisfaciendo la búsqueda de los usuarios y aumentando las ventas de los instructores.

No solo se podrían optimizar las listas que se habían mencionado anteriormente que ya se encuentran presentes en la plataforma como las listas de los cursos mas vendidos o mas vistos por los estudiantes. También se podrían crear nuevas listas de recomendación como por ejemplo basándose en la información recolectada por el atributo de **isTrendingTopic** se podría añadir una sección con las palabras clave mas relevantes del momento que muestren los cursos mas vendidos de cada tópico.

8. Conclusiones

Este trabajo tiene como finalidad obtener una mayor comprensión sobre los aspectos que influyen en la demanda de los cursos ofrecidos en la plataforma de Udemy en la categoría de 'IT & Software' analizando las tendencias del mercado y sus principales competidores.

Inicialmente se recolectaron datos sobre los cursos de esta categoría semanalmente por tres meses a través de la API Udemy Affiliate que suministra la misma plataforma. Se procesaron todos los datos y se realizó un análisis exploratorio para tener un mayor conocimiento y crear hipótesis sobre que variables podrían potencialmente tener un mayor efecto sobre la demanda. Se realizó ingeniería de atributos para crear nuevas variables en función de nuestros datos que pudieran resaltar características interesantes de los mismos. Luego se entrenó un algoritmo de K-Prototype para categorizar los datos en cinco perfiles de curso con distintas características tomando en cuenta la subcategoría de los cursos, el largo de los videos, la cantidad de material de apoyo que suministran, demanda semanal promedio, entre otros.

A continuación, se realizó una investigación extensa sobre el crecimiento del mercado en el último año y que tan sostenible era que continuará creciendo de la misma manera. Se investigó sobre los tópicos que eran tendencia en el mundo tecnológico y sobre reportes de buena fuente que realizan estudios periódicos sobre los cambios de la tendencia. De estos reportes se extrajeron las palabras claves tendencias hoy en día y se busco cada término en la plataforma de Google Trends para recolectar las búsquedas más frecuentes asociadas a los términos. Con

estos datos se creó una segunda base de datos de las palabras clave de las tendencias del día de hoy. Durante esta etapa también se investigó a la competencia de la plataforma y los efectos que tenía esta sobre la demanda y los precios de Udemy.

En base a esta investigación se volvió a hacer ingeniería de atributos añadiendo nuevas variables que midieran las tendencias y la competencia de los cursos. De aquí se armaron modelos de regresión lineal general con todos los datos ya procesados y para cada cluster. Se ajustaron todos los modelos y se validaron con validación cruzada para series de tiempo. De estos modelos se obtuvieron que factores tienen un efecto positivo sobre la demanda y cuales no. Las variables asociadas a la tendencia de los tópicos de los cursos y a la distribución de la demanda promedio semanal (cuantiles) resultaron ser de las más relevantes en todos los modelos. Se considera que en general todos los modelos entienden de forma óptima el comportamiento de los datos. Especialmente en un tema como el que se está tratando en el que influyen muchos factores externos que llegan a impactar sobre la demanda de los cursos.

Por último, en función al estudio realizado se recomendó a la plataforma suministrarle esta información a los instructores para que estos puedan aumentar las ventas de sus cursos. Utilizando los modelos para categorizar los cursos en perfiles y calcular cuales son los aspectos más relevantes en cada caso. También se recomendó utilizar esta información para optimizar la generación las listas de recomendación destacadas en la pagina principal de la plataforma lo que permitiría recomendar a los usuarios los cursos con mayor tendencia en el mercado según el análisis realizado.

Este estudio propone un modelo para tener conocimiento constante y actualizado sobre los factores que afectan la demanda. Por un lado, es un modelo que requiere de mantenimiento, por ejemplo, es necesario mantener actualizada la base de datos de temas tendencia los cuales cambian muy a menudo hoy en día y el tener que almacenar toda esta data en una base de datos puede representar un costo. Sin embargo, es un modelo que es sumamente escalable y permite que en un futuro se añadan nuevos atributos para analizar su impacto. Al ser un modelo que se nutre de la información añadir mas datos al mismo permitiría evaluar mas aspectos como por ejemplo se podría añadir información sobre si un curso es mostrado en alguna zona destacada en la pagina principal de la plataforma, lo que permitiría analizar el impacto en la demanda de los cursos que tienen una mayor visibilidad.

Incluso para futuros análisis se recomienda realizar A/B testing en la plataforma realizando variaciones en los precios y descuento para poder tener una mayor dispersión de los

datos. Esto permitiría evaluar con más detalle la influencia del precio en la demanda y realizar otros estudios sobre el posible precio óptimo. Por otro lado, se observo que las recomendación obtenidas para mejorar la demanda eran muy parecidas para cada uno de los clusters por lo que se recomienda en un futuro implementar distintos tipos de modelos analizando los clusters de forma separada para poder resaltar sus diferencias e implementar el mismo procedimiento en cada uno de los casos.

Este trabajo busca analizar que factores se pueden optimizar para mejorar la demanda de los cursos combinando un análisis de clustering con distintos modelos de regresión lineal.

9. Bibliografía

Bramer, Max (2007) Principles of data mining. Vol. 180. London: Springer.

Tan P.N., Steinbach M., and Kumar V (2005) Introduction to Data Mining, Pearson.

Luis Fonseca (2019) Clustering Analysis in R using K-means, https://towardsdatascience.com/clustering-analysis-in-r-using-k-means-73eca4fb7967
Audhi Aprillant (2021) The k-prototype as Clustering Algorithm for Mixed Data Type (Categorical and Numerical), https://towardsdatascience.com/the-k-prototype-as-clustering-algorithm-for-mixed-data-type-categorical-and-numerical-fe7c50538ebb

ZHEXUE HUANG (1998) Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values, https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.15.4028&rep=rep1&type=pdf

Future Today Institute (2021) The Future Today Institute's 14th Annual Tech Trends Report**,** https://futuretodayinstitute.com/trends/
Dhawal Shah (2021) Udemy vs Coursera: Comparing Online Learning Giants that Might IPO in 2021, https://www.classcentral.com/report/udemy-vs-coursera/

Dhawal Shah (2021), 157K Courses, 425M Enrollments: Breaking Down Udemy's Massive Catalog, https://www.classcentral.com/report/udemy-by-the-numbers/#:~:text=On%20Udemy%2C%20anyone%20can%20offer,have%2035%20million%20registered%20users.
SimilarWeb (2021) https://www.similarweb.com/website/udemy.com/?competitors=edx.org
Baidhurya Mani (2021) Top 75 Online Learning Stadistics & Trends for 2021, https://sellcoursesonline.com/online-learning-statistics

Udemy (2020) Online Educations Steps Up, https://research.udemy.com/wp-content/uploads/2020/04/Udemy_OnlineLearning_Report_4.30.pdf
Rohit Sharma (2021) Top Trending Techonologies in 2021 You Shouyld Know About, https://www.upgrad.com/blog/top-trending-technologies-in-2021/
Soumya Shrivastava (2020) Cross Validation in Time Series, https://medium.com/@soumyachess1496/cross-validation-in-time-series-566ae4981ce4
Thinkific Team (2018) Udemy´s Pricing Model: How To User It To Your Advantage As An Online Course Creator ,https://www.thinkific.com/blog/how-to-use-udemys-new-pricing-model-to-your-advantage

Udemy For Buisness (2020) Top 10 Skills in 2020, https://info.udemy.com/rs/273-CKQ-7053/images/Udemy_2020_Top_10_Skills_Global_Infographic.pdf
Rob J Hyndman and George Athanasopoulos (2021) Forecasting Principles and Practice, https://otexts.com/fpp3/tscv.html

10. Anexos

10.1 Primero 10 Registros del Conjunto de Datos Crudos.

id	title	url	isPaid	price	currency	instructors	descriptio	headline	is_practice
362328	AWS Certif	/course/av	TRUE	129.99	USD	['/user/rya	Note: Our	Want to pa	FALSE
857010	Learn Ethic	/course/le	TRUE	179.99	USD	['/user/zai	Welcome t	Become a	FALSE
2196488	Ultimate A	/course/av	TRUE	129.99	USD	['/user/ste	Welcome!	Pass the A	FALSE
393306	AWS Certif	/course/av	TRUE	129.99	USD	['/user/rya	Amazon W	Do you wa	FALSE
2359992	TOTAL: Co	/course/ne	TRUE	129.99	USD	['/user/duc	Hey, Mike	Course 1:	FALSE
614772	The Compl	/course/th	TRUE	119.99	USD	['/user/nat	Learn a pa	Volume 1 :	FALSE
1921420	Ultimate A	/course/av	TRUE	179.99	USD	['/user/ste	Welcome!	Become a	FALSE
1203374	Cisco CCNA	/course/cc	TRUE	19.99	USD	['/user/10	If you wan	The top ra	FALSE
1596138	TOTAL: Co	/course/cc	TRUE	129.99	USD	['/user/duc	Welcome	Everything	FALSE
437490	The Compl	/course/pe	TRUE	174.99	USD	['/user/ern	Gain the al	Learn how	FALSE

num_subs	discount_	list_price	saving_pri	has_disco	discount_	campaign_	campaign_	campaign_	caption_la
583907	12	129.99	118	TRUE	91	UDEMYED	2021-01-2	2021-01-1	['pt_BR', 'e
408209	17	179.99	163	TRUE	91	UDEMYED	2021-01-2	2021-01-1	['en_GB', '
248711	12	129.99	118	TRUE	91	UDEMYED	2021-01-2	2021-01-1	['pt_BR', 'f
199295	12	129.99	118	TRUE	91	UDEMYED	2021-01-2	2021-01-1	['en_US', '
85507	12	129.99	118	TRUE	91	UDEMYED	2021-01-2	2021-01-1	['pt_BR', 'c
180111	12	119.99	108	TRUE	90	UDEMYED	2021-01-2	2021-01-1	['ro_RO', '
172759	17	179.99	163	TRUE	91	UDEMYED	2021-01-2	2021-01-1	['pt_BR', 'c
85172	12	19.99	8	TRUE	40	UDEMYED	2021-01-2	2021-01-1	[]
106194	12	129.99	118	TRUE	91	UDEMYED	2021-01-2	2021-01-1	['ro_RO', 'c
280512	16	174.99	159	TRUE	91	UDEMYED	2021-01-2	2021-01-1	['ro_RO', '

avg_rating	num_revie	num_quiz	num_lectu	num_publ	num_publ	num_curri	num_of_p	quality_st	is_banned
4.52972	196682	25	266	136	11	291	147	approved	FALSE
4.557983	90011	0	145	138	0	145	138	approved	FALSE
4.703689	58620	26	309	307	25	335	332	approved	FALSE
4.362069	38827	21	121	118	12	142	130	approved	FALSE
4.710022	35221	22	131	131	22	153	153	approved	FALSE
4.543549	32942	0	124	124	0	124	124	approved	FALSE
4.69336	32026	28	362	361	27	390	388	approved	FALSE
4.741818	29439	0	309	309	0	309	309	approved	FALSE
4.61343	27896	10	121	121	10	131	131	approved	FALSE
4.294239	27567	0	121	113	0	121	113	approved	FALSE

is_publish	has_certif	subcatego	is_in_any_	language	has_close	created	instructior	estimated	is_availabl
TRUE	TRUE	132	FALSE	en_US	TRUE	2014-12-0	All Levels	1076	TRUE
TRUE	TRUE	134	TRUE	en_GB	TRUE	2016-05-2	All Levels	879	TRUE
TRUE	TRUE	132	TRUE	en_US	TRUE	2019-02-0	All Levels	1450	TRUE
TRUE	TRUE	132	FALSE	en_US	TRUE	2015-01-1	All Levels	978	TRUE
TRUE	TRUE	132	TRUE	en_US	TRUE	2019-05-0	Beginner	1049	TRUE
TRUE	TRUE	134	TRUE	en_US	TRUE	2015-09-1	All Levels	725	TRUE
TRUE	TRUE	132	TRUE	en_US	TRUE	2018-09-1	All Levels	1763	TRUE
TRUE	TRUE	132	TRUE	en_US	FALSE	2017-05-0	Beginner	2295	TRUE
TRUE	TRUE	132	TRUE	en_US	TRUE	2018-03-1	Intermedia	1137	TRUE
TRUE	TRUE	134	FALSE	en_US	TRUE	2015-03-0	All Levels	1501	TRUE

is_availabl	google_in	apple_in_a	quality_re	published	is_marketi	prerequisi	objectives	target_au	last_updat
TRUE	11.99	11.99	65	2015-01-0	TRUE	['You will n	['Pass the A	['AWS Abs	2020-11-1
TRUE	16.99	16.99	35	2016-06-2	TRUE	['Basic IT S	['135+ ethi	['Anybody	2020-10-2
TRUE	11.99	11.99	80	2019-02-1	FALSE	['Know the	['FULLY UP	['Anyone w	2021-01-1
TRUE	11.99	11.99	65	2015-02-0	TRUE	['You will n	['Pass the A	['AWS abs	2020-12-1
TRUE	11.99	11.99	80	2019-06-0	TRUE	['There are	['How to p	['Anyone k	2020-08-3
TRUE	11.99	11.99	90	2016-06-1	TRUE	['A basic ur	['An advan	['This cour	2020-08-0
TRUE	16.99	16.99	0	2018-09-1	FALSE	['Know the	['Pass the A	['Anyone w	2021-01-1
TRUE	11.99	11.99	45	2017-08-3	TRUE	['You'll nee	['Get what	['Anyone w	2021-01-1
TRUE	11.99	11.99	80	2018-04-0	TRUE	['Basic fam	['This is a c	['This cour	2020-03-1
TRUE	15.99	15.99	0	2015-03-1	TRUE	['Reliable a	['Answers	['You can b	2017-11-2

10.2 Resultados completos de las Regresiones Generales

Results

	Dependent variable:	
	log(demand)	
	(1)	(2)
log(discount_price)	0.752***	-0.079***
	(0.044)	(0.016)
log(discount_percent)	0.038***	
	(0.010)	
log(num_trendKeyWords)	0.022***	-0.021
	(0.006)	(0.020)
isTrendingTopic	0.176***	0.102***
	(0.009)	(0.028)
log(num_courses_per_topic)	0.011***	-0.009
	(0.002)	(0.006)
q1	2.847***	3.294***
	(0.013)	(0.052)
q2	1.903***	2.215***
	(0.012)	(0.045)
q3	1.403***	1.492***
	(0.013)	(0.050)
q4	0.686***	0.791***
	(0.008)	(0.026)
cluster2	-0.056***	-0.139
	(0.009)	(0.090)
cluster3	-0.096***	0.136
	(0.016)	(0.187)
cluster4	0.279***	0.053
	(0.016)	(0.084)
cluster5	-0.074***	-0.062
	(0.011)	(0.077)
log(semana)	0.275***	0.263***
	(0.005)	(0.017)
log(num_instructors)	0.076***	0.146***
	(0.017)	(0.055)

log(num_caption)	0.302***	0.021
	(0.010)	(0.039)
is_in_any_ufb_content_collection	0.668***	0.310***
	(0.011)	(0.041)
subcategoryId134	-0.059***	-0.001
	(0.010)	(0.032)
subcategoryId136	-0.084***	0.080
	(0.014)	(0.056)
subcategoryId138	-0.069***	-0.070
	(0.014)	(0.047)
subcategoryId140	-0.040***	0.008
	(0.009)	(0.025)
Constant	-2.488***	0.098
	(0.123)	(0.114)
Observations	81,946	4,325
R^2	0.683	0.780
Adjusted R^2	0.683	0.779
Residual Std. Error	0.871 (df = 81924)	0.649 (df = 4304)
F Statistic	8,402.396*** (df = 21; 81924)	765.024*** (df = 20; 4304)
Note:		*p**p***p<0.01

10.3 Resultados Completos de las Regresiones por Cluster

Results

	Dependent variable:				
	log(demand)				
	(1)	(2)	(3)	(4)	(5)
log(discount_price)	0.722***	0.690***	0.616***	1.321***	0.465***
	(0.052)	(0.081)	(0.224)	(0.199)	(0.126)
log(discount_percent)	0.038***	0.027	0.377***	0.035	-0.096***
	(0.012)	(0.016)	(0.080)	(0.052)	(0.020)
log(num_trendKeyWords)	0.014*	0.040***	0.057	0.120***	0.071***
	(0.008)	(0.010)	(0.036)	(0.036)	(0.014)
isTrendingTopic	0.146***	-0.012	0.005	-0.044	-0.058**
	(0.010)	(0.012)	(0.045)	(0.051)	(0.023)
log(num_courses_per_topic)	0.010***	0.007**	0.013	0.010	0.006
	(0.003)	(0.003)	(0.012)	(0.012)	(0.005)
q1	3.061***	2.251***	3.183***	4.981***	2.212***
	(0.018)	(0.021)	(0.087)	(0.089)	(0.026)
q2	2.041***	1.503***	2.353***	3.870***	1.491***
	(0.015)	(0.020)	(0.070)	(0.077)	(0.030)
q3	1.383***	0.910***	2.450***	2.942***	1.090***
	(0.014)	(0.017)	(0.082)	(0.067)	(0.031)
q4	0.810***	0.506***	1.534***	1.820***	0.573***
	(0.012)	(0.015)	(0.073)	(0.061)	(0.018)
q5			1.005***	0.875***	
			(0.052)	(0.077)	
log(semana)	0.314***	0.245***	0.499***	0.857***	0.288***
	(0.006)	(0.008)	(0.029)	(0.032)	(0.014)
log(num_instructors)	-0.134***	-0.033	0.179***	-0.072	0.522***
	(0.039)	(0.031)	(0.066)	(0.057)	(0.042)
log(num_caption)	0.108***	0.106***	0.428***	0.142***	
	(0.014)	(0.025)	(0.037)	(0.027)	
is_in_any_ufb_content_collection	0.332***	0.708***	1.237***	0.425***	1.071***
	(0.014)	(0.021)	(0.057)	(0.050)	(0.032)
subcategoryId134	-0.009	-0.006	-0.257***	-0.040	0.062*
	(0.012)	(0.019)	(0.052)	(0.054)	(0.035)

subcategoryId136	0.058***	0.015	-0.277***	-0.114	-0.047
	(0.017)	(0.028)	(0.058)	(0.088)	(0.078)
subcategoryId138	0.007	-0.029	-0.163**	-0.041	-0.037
	(0.016)	(0.023)	(0.078)	(0.083)	(0.095)
subcategoryId140	0.002	0.011	-0.086	-0.071	-0.288***
	(0.010)	(0.016)	(0.060)	(0.048)	(0.039)
Constant	-2.357***	-2.151***	-3.907***	-4.673***	-1.159***
	(0.148)	(0.225)	(0.690)	(0.553)	(0.340)
Observations	32,936	23,460	7,105	3,950	14,495
R^2	0.741	0.621	0.552	0.745	0.632
Adjusted R^2	0.740	0.620	0.551	0.744	0.631
Residual Std. Error	0.711 (df = 32918)	0.756 (df = 23442)	1.437 (df = 7086)	1.100 (df = 3931)	0.813 (df = 14478)
F Statistic	5,526.185*** (df = 17; 32918)	2,255.225*** (df = 17; 23442)	485.356*** (df = 18; 7086)	637.519*** (df = 18; 3931)	1,551.492*** (df = 16; 14478)
Note:					*p**p***p<0.01

10.4 Instructivo de para la Ejecución de los Archivos de R y Python

Archivo	Requisitos	Genera
ReadData.R	Data recolectada cada semana en formato xls.	• notFixedTable.xlsx • fixedTable12.xlsx • finalData.xlsx • TableForFORECAST.xlsx
AnalizeData.R	• fixedTable12.xlsx • Datos recolectados de la semana 1 en formato xls.	-
AnalizeNoFixedData.R	• notFixedTable.xlsx • fixedTable12.xlsx	-
Clustering.R	• finalData.xlsx • AllData.xlsx	-
LM_FinalResults.R	• TableForFORECAST.xlsx • CleanData.xslx	Resultados de la regresión lineal de los modelos generales con y sin descuento.
LM_Clusters_FinalResults.R	• TableForFORECAST.xlsx • CleanData.xslx	Resultados de la regresión lineal de los modelos de cada uno de los clusters.
Api.py	-	Data recolectada cada semana en formato xls.