Fidelización Cursos Udemy

Mateo Zapata López & Ana María Montes &
Nicolas Rozo Bedoya

Universidad de los Andes.
Departamento de Ingeniería de Sistemas.
Ciencia de datos aplicada

Tabla Contenido

Definición de roles: Para el desarrollo del presente proyecto se propone abordar los datos de la plataforma digital Udemy respecto al servicio de cursos virtuales, con el fin generar una propuesta de valor, el trabajo se desarrollará mediante la metodología de ASUM-DM la cual para este segundo sprint abarca hasta la etapa 10. Los roles para el trabajo se encuentran definidos de la siguiente manera:

Líder del proyecto: Ana María Montes.
Líder de datos: Nicolas Rozo Bedoya.
Líder de modelos: Mateo Zapata López.

Entender el problema

En los últimos años se ha observado un aumento de usuarios en las plataformas de educación en línea. Esta transformación se debe a la facilidad del acceso al servicio de internet y dispositivos tecnológicos, la educación en línea consiste en cursos virtuales que son un espacio de enseñanza y aprendizaje que acontecen a través de internet. A continuación, se exponen las ventajas y desventajas de los cursos en línea.

Ventajas:
- El acceso al contenido de los cursos no está atado a una posición geográfica.
- El ritmo de aprendizaje es determinado por el estudiante.
- Hay variedad y diversidad en el tipo de contenido.
- Los precios de los cursos tienden a ser accesibles.
- El consumo de contenido es asíncrono.

Desventajas:
- Se requiere de una conexión internet para subir y acceder al contenido de los cursos.
- Existe poca personalización en las clases.
- No existe un seguimiento a los estudiantes que toman la clase.
- Requiere constancia por parte del estudiante para lograr un aprendizaje.

Se estima que el mercado para la educación en línea será de 350 millones de dólares para el 2025 (Koksal, 2020). Dentro del mercado las plataformas más conocidas son Udemy, Coursera, Udacity, Edx, Platzi, entre otros.

Udemy es una plataforma digital MOOC (Massive Online Open Course) creada en el año 2009 que tiene como objetivo ofrecer cursos en línea gratuitos y pagos, que ofrecen clases de programación, finanzas, diseño gráfico, música, entre otros. La misión de Udemy es conectar a los estudiantes con los mejores instructores y ayudar a las personas alcanzar sus metas y perseguir sus sueños. Actualmente tiene más de 50 millones de estudiantes y 57.000 instructores enseñando 150.000 cursos en 65 lenguajes, algunos de los cursos están acreditados y generan certificaciones. (Udemy, 2021).

Udemy permite crear y vender el contenido de una persona, que se le domina instructor. Los estudiantes pueden comprar el curso y el instructor recibe un porcentaje de la compra. Cualquier persona con interés por enseñar puede registrarse a Udemy sin costo alguno, grabar su curso, difundirlo, y generar ingresos extras a través de la suscripción paga a los cursos. El éxito de un curso depende de la calidad del material del curso, la demanda que exista por el tema que se enseña, y el nivel de competencia en la oferta de cursos similares; cada copia vendida del curso genera una comisión para el instructor, que es normalmente del 50% del precio de venta del curso. Dicho porcentaje de comisión puede llegar hasta un 97% si el instructor vende el curso a través de promociones propias como en su canal de YouTube, redes sociales o blogs” (Armenta, 2017).

Ideación

La herramienta Arquetipo, nos permite modelar, entender y establecer el público objetivo, para este caso, el interés está centrado en un modelo de fidelización para los cursos ofrecidos por Udemy, en la *ilustración 1* se exponen los comportamientos, acciones, detalles demográficos y puntos de dolor identificados para el arquetipo de interés.

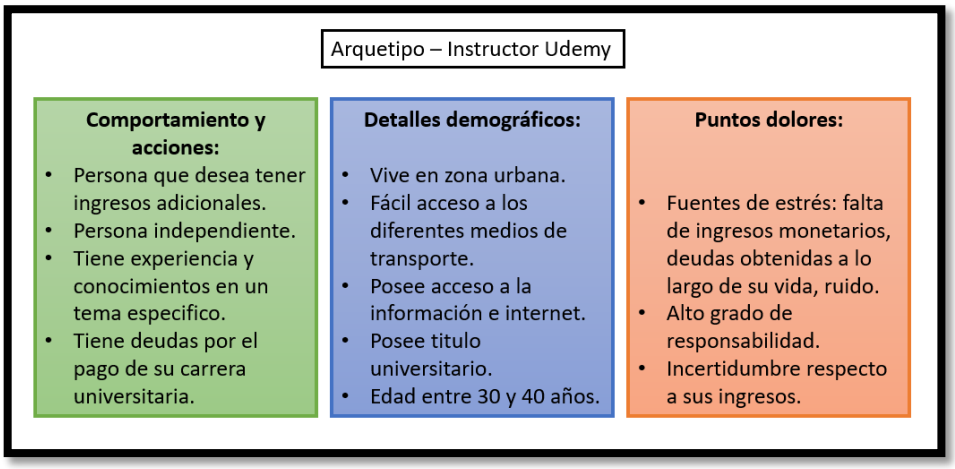

Ilustración 1 Arquetipo Instructor Udemy

La definición de este arquetipo permite empatizar y entender el punto de vista y los dolores del "instructor Udemy", con lo encontrado se generaron las siguientes estrategias:

- Identificar posibles elementos diferenciadores que se puedan ofrecer en cada curso y tengan impacto en la plataforma.
- Proponer lineamientos de los cursos para obtener mayor cantidad de interacciones con los estudiantes, respecto al contenido, duración y precio.
- Mayor vinculación a los cursos y una constante actualización del contenido.

Enfoque analítico

Udemy no cuenta con un estándar en los cursos ofrecidos, generando alta variabilidad en la estructura de los mismo. Adicionalmente, la gran cantidad y variedad de plataformas, tanto privadas como open source, evidencia que Udemy no ofrece un elemento diferenciador. Según este orden de ideas y teniendo en cuenta lo identificado tanto en el entendimiento del problema como en el enfoque analítico**,** el desarrollo del presente proyecto tendrá tres objetivos:

- Generar lineamientos sobre las características principales que debe tener la estructura de un curso respecto a la duración del contenido, tipo de contenido, categoría, nivel y tema en específico.
- Aumentar la interacción de los estudiantes con la plataforma generando mayor cantidad de visualizaciones y comentarios para que los instructores actualicen constantemente su contenido.
- Lograr que las personas que ven los videos de manera gratuita se vinculen a se conviertan en suscriptores de los cursos ofrecidos por la plataforma.

Para lograr estos objetivos se hará un énfasis en la exploración de datos, en especial en el análisis bivariado y se implementaran dos modelos de regresión lineal, para el primero modelo se tendrá una regresión sencilla con la variable respuesta (Numero de reviews) y la variable predictora (Número de subscriptores). Para el segundo modelo se tendrá una regresión múltiple con la variable respuesta (Número de subscriptores) y las variables predictoras (Precio, duración del contenido y Numero de lecturas).

Entendimiento de los datos

El *dataset* a utilizar en el proyecto fue obtenido de la plataforma ***Kaggle***, cuenta con 3.682 datos, que abarcan desde el 2011 hasta el 2017, referentes a los cursos correspondientes a cuatro grandes categorías, allí se encuentra información de 12 variables tanto numéricas como categóricas correspondientes al id, url, título del curso, si es pago o no, fecha de publicación, duración del contenido y los valores correspondientes a subscriptores y los diferentes tipos de interacciones. El formato de los datos corresponde a un archivo csv (Comma Separated Values).

En el análisis de la calidad de los datos encontramos que el dataset se encuentra completo, sin datos nulos o faltantes, lo que facilita la etapa de preparación, continuando con la limpieza encontramos que solo tiene 6 datos duplicados para los cuales se identifica que corresponden a url repetidos (cada curso tiene un url único), dado que el porcentaje es muy bajo y poco representativo no se considera realizar una imputación y solamente se eliminan del dataset.

Respecto a las transformaciones pertinentes, para la variable published_timestamp se realizó un cambio del formato y adicionalmente se creó una nueva columna donde se extrae el año; continuando con el enriquecimiento del data set se creó la columna con el cálculo de las ganancias obtenidas por subscriptores y se transformó la variable de precio en categorías(Económico, estándar, premium, royal) para completar el análisis de segmentación, con estas nuevas variables se crea un nuevo dataframe para el cual se eliminan las columnas published_timestamp, url e id ya que no aportan información relevante en el análisis que se desea realizar. Para finalizar se crearon graficos de los primeros 10 cursos que poseen más suscriptores, más número de reviews, más número de lecturas, más contenido de duración y mayores precios, también se identificó la cantidad de registros pagos y gratuitos) con el fin de identificar cuáles deben tener mayor importancia. A continuación, se presentan los principales hallazgos de la exploración de datos.

Análisis univariado

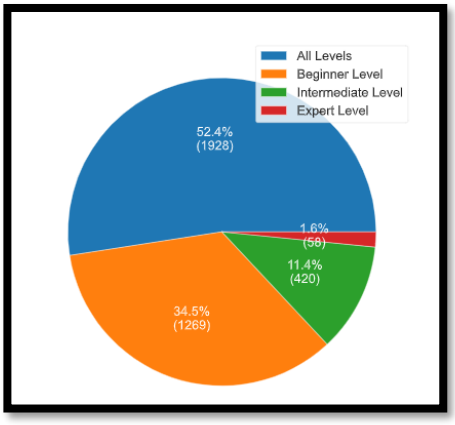

Ilustración 2 Niveles cursos

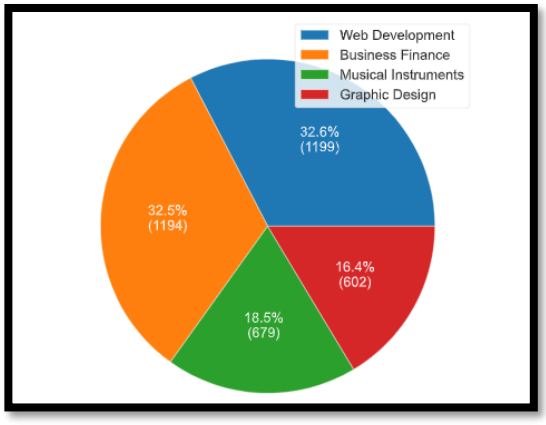

Ilustración 3 Categorías cursos

La *ilustración 2* corresponde a los 4 niveles que se pueden tener por curso, se observa como la mayoría de los estudiantes se encuentran en todos los niveles, seguido del nivel principiante con un 34,5 además se destaca el bajo porcentaje de nivel experto menos de 2%, esto nos permite confirmar que los usuarios de Udemy utilizan la plataforma para iniciarse en un tema y después continúan especializándose en otras. En la *ilustración 3* vemos la participación de cada categoría, donde encontramos que la más solicitada es Web Development pero en general se tienen porcentajes muy parejos por lo cual se considera explorar con mayor profundidad esta variable.

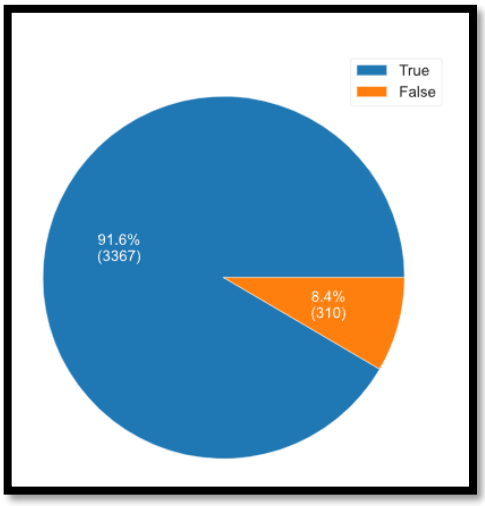

Ilustración 4 Gratuito o pago

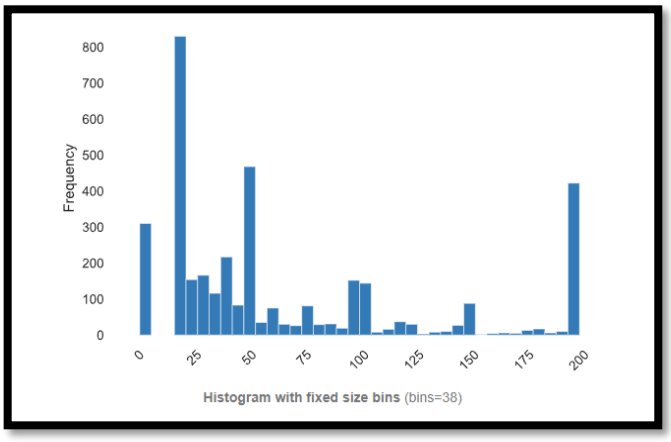

Ilustración 5 Distribución precios

En la *ilustración 4* observamos que casi el 92% de los cursos que contiene la plataforma son pagos y en la *ilustración 5* tenemos el histograma de los precios, notando picos altos en los valores de 9, 50 y 20$ dólares, estos valores se incluirán en rangos (Económico, estándar, premium, experto) para categorizarlos y realizar un análisis más valioso.

Análisis bivariado y multivariado

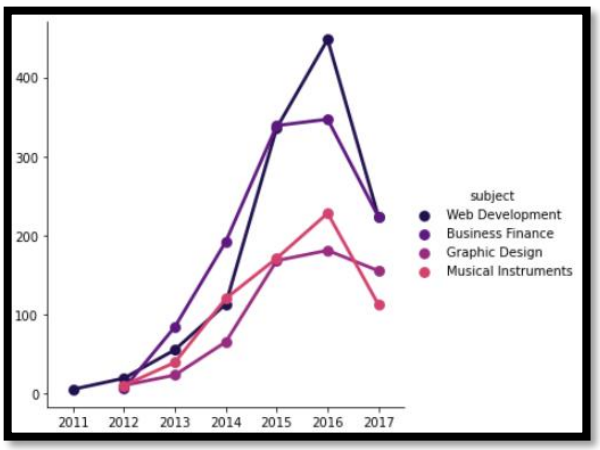

Ilustración 6 Año – Categorías

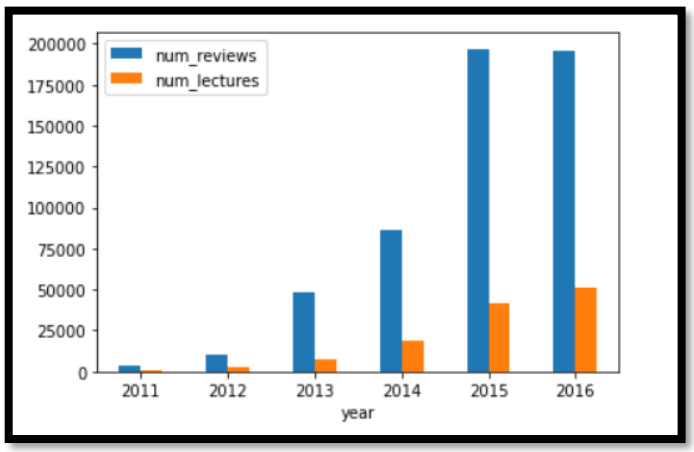

Ilustración 7 Años- Reviews - Lectures

Para la *ilustración 6* el grafico bivariado permite observar con mayor claridad la distribución de las categorías, ya es notable la diferencia entre cada una y los picos que ha tenido la serie en los 7 años de análisis, destacando sobre todo las dos primeras categorías. Dado que los histogramas relacionados con las variables de interacción con el usuario no eran muy representativos se construyó la *ilustración 7* que permite observar la comparación del comportamiento del número de reviews respecto al número de lecturas para los años de análisis, la gráfica permite observar la diferencia significativa entre ambas variables mostrando que ambas presentan un crecimiento exponencial anual pero no presentan una relación.

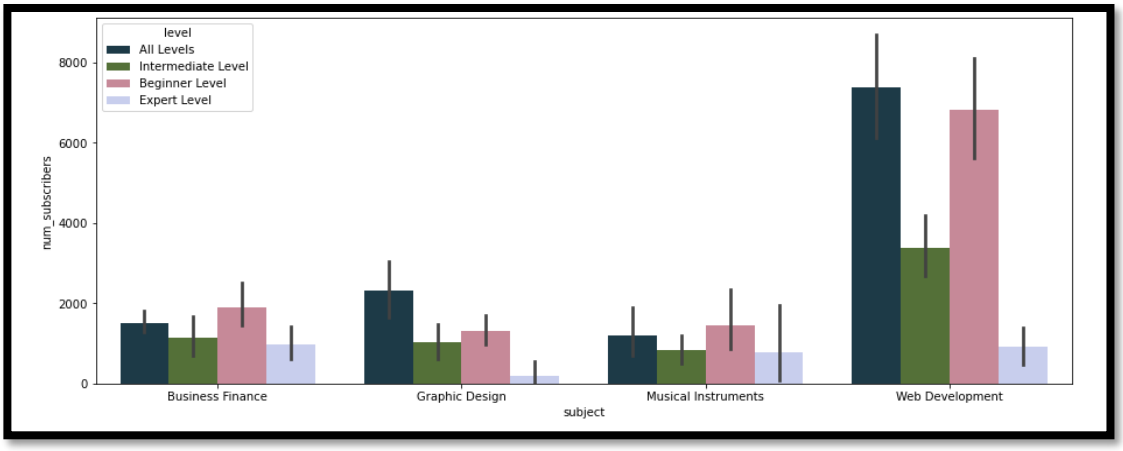

Ilustración 8 Categoría - Niveles – Subscriptores

La *ilustración 8* muestra de manera detallada, el comportamiento de las categorías, el nivel de cada una respecto al número de subscriptores, en esta grafica es aún más marcada la diferencia entre las categorías, sobre todo entre Web Development y Business Finance, que para la ilustración 3 parecían ser casi iguales, la grafica permite destacar el alto nivel de subscriptores para la categoría Web y sobre todo la diferencia del nivel principiante confirmando lo analizado en la ilustración 2.

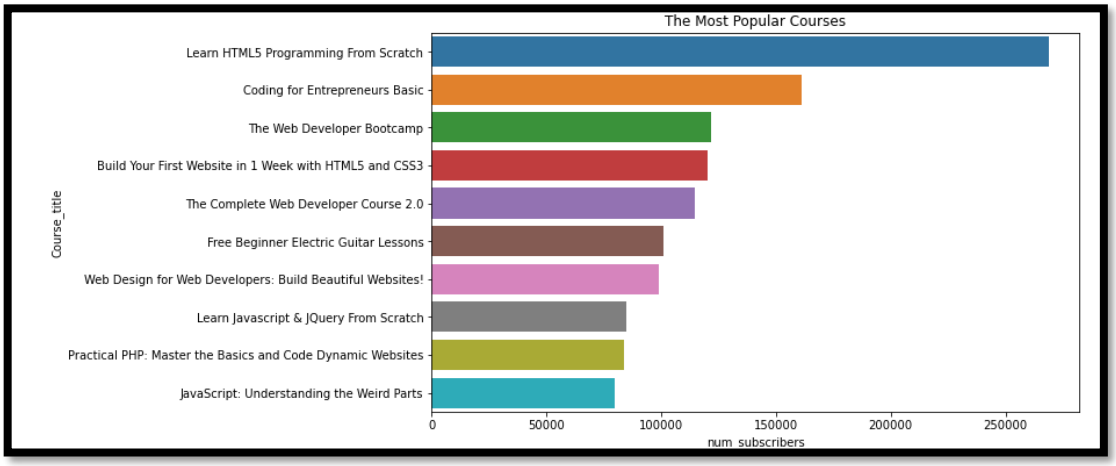

Ilustración 9 Top 10 de cursos por cantidad de subscriptores

La *ilustración 9* muestra los 10 cursos más populares con el criterio de número de subscriptores, la idea es tener estos cursos identificados ya que los instructores de Udemy tienen que actualizar su contenido, la idea es enfocarnos en los cursos del top 10. Con el fin de entender los ingresos que están generando los mismos. Se realizo un conteo sobre los que son pagos y no son pagos, respecto a su categoría obteniendo la distribución de la tabla 1.

	is_paid	subject	0
0	False	Web Development	5
1	True	Web Development	4
2	False	Musical Instruments	1

Tabla 1 Cursos del Ranking pagos y gratuitos

Adicional en las gráfica creadas por el perfilamiento de datos se encuentra que, la mayoría de los cursos tienen una duración entre 0-5 horas, que usualmente hay entre 1 y 50 lecturas por curso, que los cursos suelen tener pocos reviews, la mayoría de los cursos están en el mismo rango de suscriptores, hay cursos para los cuales es muy marcado que se encuentran en el trending topic y para finalizar suponiendo que los precios se encuentran en dólares, el rango esta entre 0 y 250 dólares, mostrando como el precio más común $25. A continuación, la *ilustración 10* presenta la distribución de los precios respecto a su categoría.

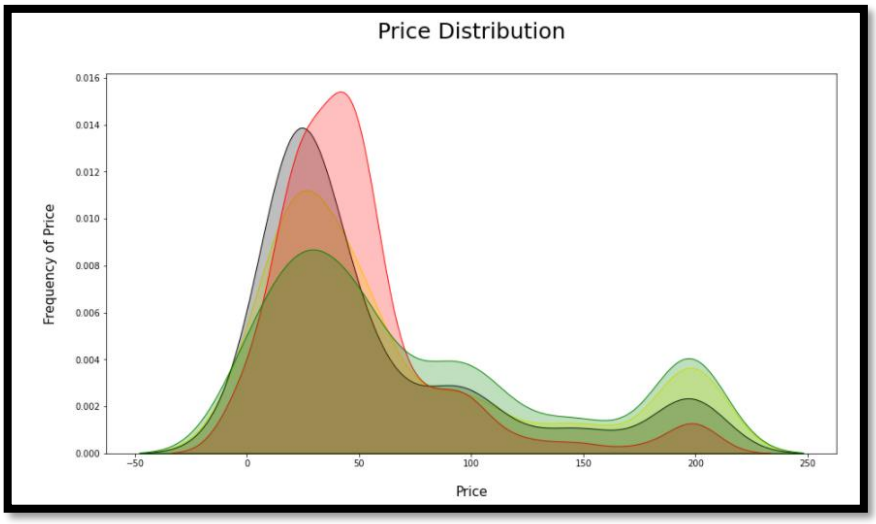

Ilustración 10 Distribución de precios por categoría

Donde el color amarillo corresponde a Business Finance, el negro a Graphic Design, rojo a Musical Instruments y el verde a Web development. Esta grafica permite confirmar muchos de los análisis realizados respecto al nivel principiante dado que los cursos de mayor valor correspondientes a nivel experto son los menos comunes y se ven nuevamente los picos observados en el perfilamiento, destacando que la categoría más popular correspondiente a Web development en color verde no es la que tiene los precios más altos, todo lo contrario es la que cuenta con más cursos gratuitos [133 cursos gratuitos] (Resultado de análisis exploratorio contenido en el código).

Ilustración 11 Boxplot contenido, categorías, pago o gratuito

Finalmente, la *ilustración 11* correspondiente a los boxplot de las categorías respecto a la duración del contenido y si el curso es pago o no, resalta como el contenido gratuito tiene una menor duración de contenido.

Preparación de los datos

Con el fin de realizar los modelos predictivos, se divide el dataset en dos conjuntos, tomando el 80% para la etapa de entrenamiento del modelo y el 20% para las pruebas. Con el fin de verificar que la partición se haya realizado de manera adecuada se grafican cuatro histogramas con la distribución de los datos para training y test confirmando que sean iguales, en la *ilustración 12 se* muestra un resumen

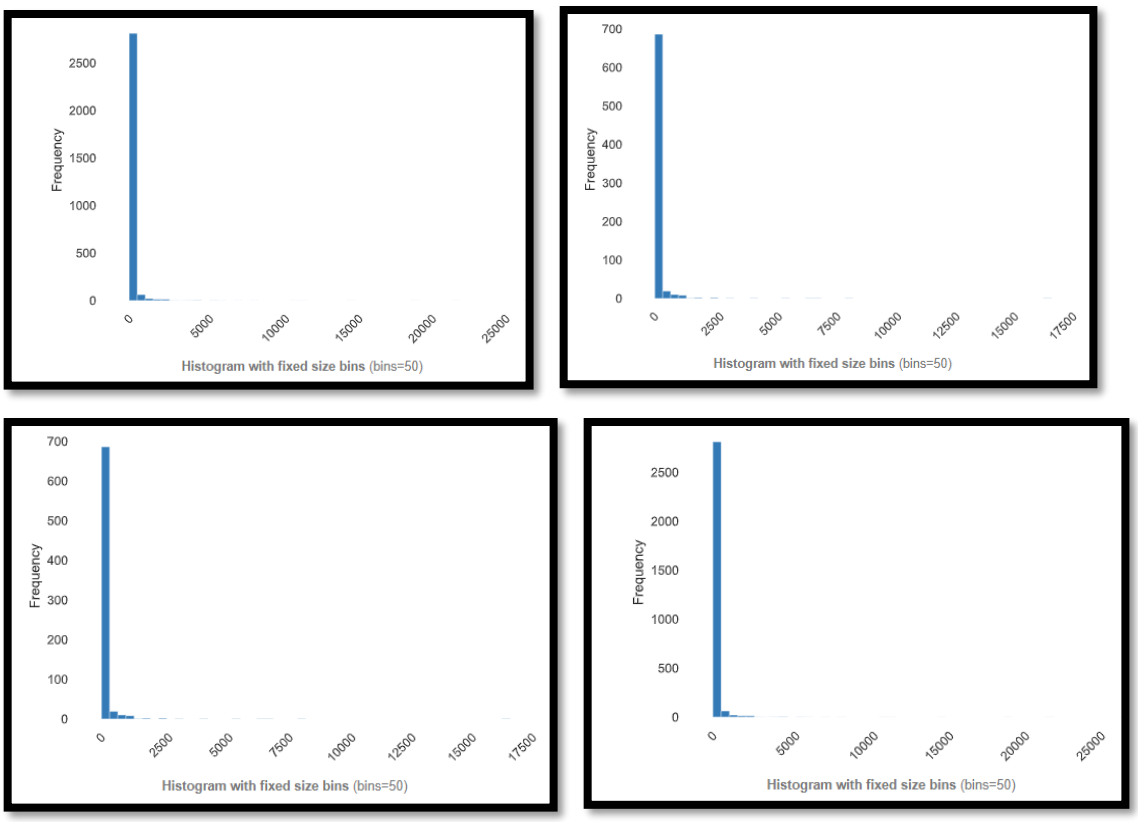

Ilustración 12 Distribución datos de entrenamiento y de prueba

Teniendo en cuenta que la distribución del conjunto de datos corresponde a una distribución exponencial, se aplica la prueba de Kolmogorov-Smirnov en ambos conjuntos de prueba con el fin de concluir si son iguales (El desarrollo se encuentra en el código). El resultado obtenido es que se acepta la hipótesis nula (Ho=Las distribuciones de las muestras son similares).

Construcción del modelo

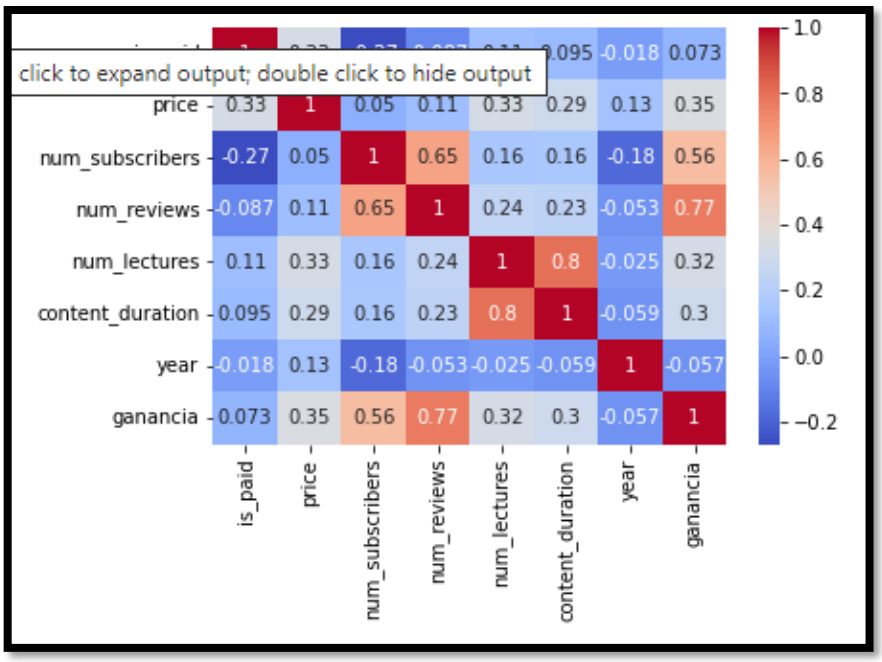

Ilustración 13 Diagrama de correlación

El diagrama de correlación presentado en *ilustración 13* permite observar que no existe una fuerte correlación entre las variables, pero si una moderada entre número de suscriptores, reviews y lectures. Con esta relación identificada se plantean los siguientes modelos de predicción:

Modelo 1-> Regresión lineal simple

Para el modelo 1, se tuvo en cuenta la mayor correlación observada referente al número de reviews (comentarios y/o reseñas) que los estudiantes dejan en el curso (variable de respuesta) y el número de suscriptores (personas que se vinculan a los cursos) (variable predictora) como se observó en la *ilustración 13,* Con este primer modelo se desea predecir el número de reviews a partir de un número de suscriptores dado, se debe tener en cuenta que un review no puede existir si la persona no se ha suscrito al curso, es por esto que se debe seguir el orden lógico del proceso.

La regresión lineal es un modelo que trata de explicar la relación que existe entre una variable dependiente (variable respuesta) y un conjunto de variables o una sola variable (variables explicativas), este modelo busca una función que se aproxime a los datos, estableciendo valores predictores y valores reales.

Matemáticamente la regresión lineal se da por la siguiente formula:

$$Y = \propto + \beta X + \varepsilon$$

Y= variable respuesta.
X=variable explicativa.
α=variable respuesta en el origen.
β=pendiente de la recta, indica como cambia Y al incrementar X.
ε=factores de error.

Modelo 2-> Regresión lineal múltiple

Para el segundo modelo se tomaron todas las variables que presentan una correlación considerable con el fin de mejorar los resultados, con esto en mente se pretende predecir el número de subscriptores (variable respuesta) respecto a la variable precio, número de lecturas y número de reviews del curso (variables predictoras).
Se puede observar que para el segundo modelo se plantea una regresión lineal múltiple, la cual es muy parecida al anterior modelo solo que acá se establece un conjunto de variables explicativas convirtiendo nuestro modelo es predecir un valor mediante múltiples variables.

Evaluación del modelo

Modelo 1

Con estos lineamientos se encuentra que para el primer modelo se obtiene un *Error medio cuadrático (varianza) de 732.43,* se aclara que este indicador es una función de riesgo que corresponde a la diferencia entre el estimador y lo que se estima, se busca que sea lo menor posible, También se obtuvo un *Error absoluto medio (desviación absoluta promedio) de 162.80* aclarando que esta es una medida que estima la diferencia entre dos variables continuas, la cual cuantifica la precisión de una técnica de predicción comparando los valores predichos con los valores observados, y finalmente se obtuvo un *R cuadrado o coeficiente de determinación de 0.437* lo que indica el ajuste del modelo a la variable que estamos intentando predecir, este indicador varía entre 0 y 1, siendo 0 el valor mínimo y 1 el valor máximo. El coeficiente de determinación ajustado indica la medida que soluciona los problemas que presenta el coeficiente de determinación, esta medida define el porcentaje explicado por la varianza de la regresión de acuerdo a la varianza experimentada por las variables aplicadas para el modelo 1 tenemos un valor 0.437, el coeficiente de regresión indica el grado de relación de dos variables siempre y cuando ambas sean cuantitativas y continuas para el modelo 1 es de 0.062.

```
MSE: 732.43
MAE: 162.80

coefficient of determination: 0.43717845003705924
adjusted coefficient of determination: 0.43702513444186364
regression coefficients: x1    0.062693
```

Modelo 2

Para el modelo 2 se implementó una regresión múltiple teniendo como variables predictoras el precio, la duración del contenido y el número de lecturas, obtenido una varianza de 9524.54 indicando la diferencia entre el estimador y lo estimado, se obtuvo también una desviación absoluta promedio de 3717.47 lo que cuantifica la precisión de un valor predicho con los observados, otro indicador importante es el R cuadrado o coeficiente de determinación que para el modelo 2 fue 0.468, con un ajuste de 0.467 lo que penaliza la inclusión de aquellas variables que no resultan trascendentales para la variable real.

```
MSE: 9524.54
MAE: 3717.47
```

```
coefficient of determination: 0.11368055017553624
adjusted coefficient of determination: 0.11295584089522193
regression coefficients: price              462.330167
content_duration    158.088925
num_lectures         19.780124
```

Modelo 1 vs Modelo 2:

Podemos concluir que le modelo 1 tiene mejores métricas e indicadores que el modelo 2, esto se debe a que las variables utilizadas tienen mayor correlación, aseverando que las predicciones que haga el modelo 1 serán más reales y con mayores acercamientos de los valores predichos a los observados que las del modelo 2.

Para mejorar los modelos es pertinente evaluar las variables que poseen mayor correlación y utilizar variables cuantitativas y continuas, así se elimina el sesgo y se pueden utilizar la regresión lineal, dado el caso de no tener una buena correlación es importante implementar transformaciones y/o funciones que mejoren la correlación.

Aplicativo

Para el despliegue de la solución se desarrollo un servicio REST con la librería Flask (micro framework, bajo el patrón Modelo Vista Controlador) y Jinja (un motor de templates para Python, configurado con Flask), creando dos métodos, cada uno para calcular un modelo (regresión lineal simple y regresión lineal múltiple), los modelos fueron entrenados con el dataset Udemy de kaggle y las predicciones se realizan por medio de la ecuación que predice el modelo con sus respectivos coeficientes.

Así cuando se ingrese una nueva variable el modelo hará las predicciones correspondientes cuando se oprima el botón calcular.

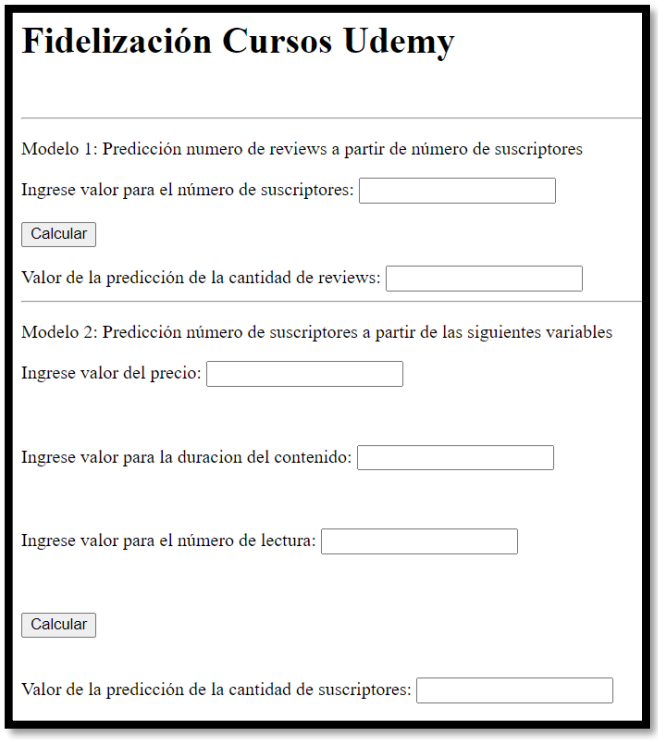

Ilustración 7. Despliegue de la solución.

Recordemos que para el primero modelo la variable respuesta es el numero de reviews y para el modelo 2 es el número de suscriptores.

Conclusiones

Respecto a las variables categóricas, la correspondiente a los niveles, muestra claramente como todos los niveles y nivel principiante son los más representativos, de lo cual se obtiene que el público objetivo son personas que van a iniciar su proceso de aprendizaje, adicionalmente gracias a la exploración de la variable de categorías se conoce que si un instructor quiere aumentar sus ganancias y número de subscriptores debe enfocarse en los cursos de la categoría Web development en especial aquellos que tienen mayor cantidad de subscritores como lo son learn HTML 5, Coding for y Bootcamp y se estima que este contenido debe tener una duración aproximada de 5 horas. Se concluye que la relación entre variables es importante a la hora de generar un modelo de regresión lineal, no necesariamente entre más variables predictoras tengamos va ser mejor el modelo ya que la correlación influencia mucho en las predicciones.

Para este análisis seria aún más valioso contar con datos del 2017 al 2021 con el fin de analizar lo ocurrido durante la pandemia y datos referentes a la edad de las personas que acceden al curso, grado de escolaridad, sexo y conocer quienes terminan todo el curso con certificación. La mayor dificultad se encontró en la baja correlación de las variables lo que género que los modelos no tuvieran un muy buen ajuste y por dicha razón el esfuerzo se centró en la exploración de los datos y análisis multivariados. Finalmente, el estudio realizado podría utilizarse para una campaña de fidelización enfocada en mejorar la estructura de los cursos y a su vez la calidad y pertinencia servicio ofrecido, logrando que más personas vean el contenido y a mediano plazo se subscriban a la plataforma.

Enlace repositorio

El proyecto se trabajó en el repositorio de Github (enlace) que contiene el notebook y el aplicativo web. Adicionalmente se utilizó la funcionalidad de proyecto para hacer un seguimiento de cada una de las tareas realizadas, especificando sus características, responsable y etapa en la que se encuentra.

Referencias

- Armenta, M. H. (7 de 09 de 2017). Udemy, la plataforma que te paga por dar cursos en línea.Forbes . Obtenido de https://www.forbes.com.mx/udemy-la-plataforma-que-te-paga-por-dar-cursos-en-linea/
- Koksal, I. (2 de 05 de 2020). The rise of online learning . Forbes. Obtenido de https://www.forbes.com/sites/ilkerkoksal/2020/05/02/the-rise-of-online-learning/?sh=2994688472f3
- Udemy. (2021). Udemy. Obtenido de https://about.udemy.com/es/
- Documentación Flask (en línea).
- Documentacion Jinja (en línea).
- Wes McKinney. Python for Data Analysis, 2nd Edition. O'Reilly Media, Inc., 2017.