Entrega 2: Desarrollo

A continuación se presenta el proyecto realizado por el equipo en el contexto de Trabajo final de Máster.

1. Objetivo

Desarrollar un **Modelo de Churn** para la empresa **Treatwell Spain SL** que permita identificar los factores que influyen en la cancelación de los salones después de utilizar los servicios de la plataforma.

2. Justificación

Desarrollar un Modelo de Churn para la empresa Treatwell Spain SL que permita identificar los factores que influyen en la cancelación de los salones después de utilizar los servicios de la plataforma. Al conocer las razones detrás de la pérdida de salones, la empresa podrá implementar medidas preventivas para retener a los salones actuales y mejorar su satisfacción y experiencia en la plataforma.

Además, la implementación de este modelo también es crucial para optimizar las estrategias de adquisición de salones de Treatwell, ya que permite enfocarse en los salones con mayor probabilidad de permanecer en la plataforma por un período prolongado. Esto puede tener un impacto significativo en la rentabilidad de la empresa, ya que retener a los salones existentes resulta más económico que adquirir nuevos.

En conclusión, la creación de un Modelo de Churn permitirá a Treatwell mejorar la satisfacción de los salones, retener a los salones existentes, atraer a nuevos salones y, así, mejorar la rentabilidad de la empresa.

3. Antecedentes

Treatwell Spain SL, es una empresa de servicios B2B especializada en la gestión de reservas y citas para salones de belleza. Como parte de su visión estratégica, la empresa busca retener a los salones actuales y mejorar su experiencia, y así, su satisfacción en la plataforma. Para ello, se ha solicitado el desarrollo de un modelo de Churn que permita identificar los factores que influyen en la cancelación de una suscripción y, de esta manera, tomar medidas preventivas.

Además, la implementación de este modelo también permitirá optimizar las estrategias de adquisición de nuevos salones. Al identificar los factores que influyen en la cancelación de una suscripción, podrán enfocarse en atraer a salones con mayor probabilidad de permanecer en la plataforma a largo plazo.

4. Scope

El alcance del proyecto se centrará en construir un modelo de churn para identificar los salones de belleza que son más propensos a cancelar su relación con Treatwell después de utilizar los servicios de la plataforma.

5. Desafíos

Uno de los principales desafíos que debemos considerar es la calidad y la disponibilidad de los datos necesarios para construir el modelo de churn. La pandemia de COVID-19 ha tenido un impacto significativo en la industria de los salones de belleza, lo que puede haber provocado cambios en el comportamiento de los clientes y en las operaciones comerciales de los salones. Esto puede influir en los datos disponibles y su representatividad.

En primer lugar, es posible que algunos salones de belleza hayan cerrado temporal o permanentemente debido a la pandemia. Esto podría llevar a una disminución en la cantidad de datos disponibles o incluso a la falta de datos de los salones que ya no están en operación. Como resultado, el conjunto de datos utilizado para construir el modelo puede ser menos completo y representativo de la situación actual.

Además, los salones de belleza que han continuado operando post-pandemia pueden haber experimentado cambios en su comportamiento, como una disminución en la demanda de servicios o una reducción/aumento en la frecuencia de uso de Treatwell. Esto puede generar un desequilibrio en los datos, lo que significa que las muestras pueden estar sesgadas hacia un tipo específico de comportamiento, dificultando la generalización del modelo a todos los salones de belleza.

Otro aspecto a considerar es que las restricciones y regulaciones impuestas debido al COVID-19 pueden haber afectado la forma en que los salones de belleza interactúan con los clientes y utilizan la plataforma de Treatwell. Por ejemplo, algunos salones pueden haber cambiado su modelo de negocio para adaptarse a las restricciones, como ofrecer servicios a domicilio en lugar de atender clientes en el salón. Estos cambios en las prácticas comerciales pueden influir en los datos disponibles y en la relación entre los salones y Treatwell.

Para abordar los desafíos mencionados, se ha desarrollado un modelo específico para el período posterior a la pandemia, teniendo en cuenta los cambios en la industria de los salones de belleza y el comportamiento de los clientes. Se ha puesto especial atención en verificar que los datos utilizados en el modelo sean imparciales, y se han realizado ajustes para garantizar su representatividad.

Es importante tener en cuenta que la naturaleza de los datos puede variar considerablemente entre los años 2021-2022 y 2022-2023, debido a los efectos post-pandemia mencionados previamente.

Por lo tanto, al utilizar este modelo para predecir la propensión de Churn de los salones de belleza, es fundamental considerar las diferencias y posibles limitaciones que puedan surgir debido a las variaciones en la naturaleza de los datos entre ambos períodos. Será necesario adaptar y actualizar continuamente el modelo para reflejar con precisión la situación actual y tomar medidas proactivas que permitan retener a los salones de belleza en la plataforma.

6. Hitos Temporales del Proyecto

Con el objetivo de resolver el problema planteado, el equipo estableció hitos temporales para el progreso del proyecto.

Dado que el proyecto involucra datos de Treatwell, se comenzó definiendo los términos para compartir la información con el equipo de trabajo. Se acordó que Bruno sería el encargado de solicitar la extracción de datos y compartirlos en formato CSV con el resto del equipo.

En conjunto, se examinaron trabajos, foros, tutoriales y otros recursos para analizar el tratamiento adecuado de los datos y el formato en el que se presentan al modelo. Posteriormente, fue fundamental llevar a cabo reuniones para comprender el entorno empresarial con el que trabajaríamos, así como los desafíos y los resultados óptimos que esperamos obtener de nuestro proyecto.

Después de familiarizarnos con el entorno de negocio, se nos proporcionaron los nombres de múltiples variables con las que Treatwell trabaja a diario. A partir de estas variables, procedimos a filtrarlas teniendo en cuenta su significado empresarial y su uso actual en la empresa. Una vez realizada la selección, el equipo analizó cada variable, identificando aquellas con un exceso de valores nulos, variables que carecían de sentido en función de su significado, entre otras, para luego consultarlas con el equipo de Data Engineering de la empresa. Después de varias consultas con el equipo de datos, logramos realizar el primer filtrado, que fue la base para el análisis exploratorio de aproximadamente 50 variables (las conclusiones de este análisis se presentan en la sección correspondiente). Posteriormente, se propusieron nuevas variables/métricas para identificar los factores más influyentes en la pérdida de clientes.

Una vez creadas las nuevas variables y eliminadas aquellas que no aportan valor al análisis, y basándonos en los requisitos de la empresa, procedimos a realizar la primera prueba del modelo con fines comparativos.

A lo largo del proceso, el equipo se ha preocupado por documentar su análisis, no sólo para las entregas relacionadas con el TFM (Trabajo de Fin de Máster), sino también para que la empresa pueda utilizar estos recursos en el futuro.

A continuación, se presenta el detalle en semanas de cada mes:

	Mayo				Junio				Julio			
	S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4
Definición términos y firma de NDA	x	x	x									
Estudio de modelos Churn	x	x										
Estudio y análisis del contexto de negocio			x									
Recepción de variables de negocio				x								
Exploración de negocio de las variables					x	x						
Consulta con equipo de data					x	x	x	x				
Presentación de análisis inicial						x						
Exploración de variables seleccionadas						x	x	x	x			
Revisión variables seleccionadas							x	x	x			
Creacion variables propuestas									x			
Preparación modelo benchmarking									x	x		
Documentación	x	x	x	x	x	x	x	x	x	x	x	

7. Gestión del equipo de trabajo

Roles y responsabilidades

Al inicio del proyecto definimos los siguientes **roles y responsabilidades** para garantizar una buena organización y desarrollo del proyecto:

- **Bruno Pedemonte:** Encargado del área de negocio, se encargará de tomar decisiones en cuanto a la ejecución del proyecto.
- **Arnau Andrews** y **Oriol Masias**: Encargados de la preparación del ambiente de modelación, serán responsables de la configuración del entorno de trabajo y de la gestión de datos.
- **Camila Silva**: Encargada de la modelación, liderará el desarrollo del modelo de Churn.

Durante estas semanas, hemos adoptado una **metodología de trabajo transversal** y colaborativa en la cual todos los miembros del equipo han tenido la oportunidad de participar activamente en todos los aspectos del proyecto. Además, hemos implementado un **sistema de revisión cruzada**, permitiendo a cada miembro del equipo examinar y evaluar el trabajo de los demás.

Comunicación y colaboración

Durante el desarrollo del proyecto hemos llevado a cabo reuniones semanales de aproximadamente dos horas utilizando **Google Meet**. Estas reuniones han sido fundamentales para llevar a cabo diferentes discusiones sobre el proyecto, compartir ideas,

revisar las tareas, plantear dudas y establecer metas y objetivos claros para cada etapa del proyecto.

Además, para mantenernos conectados de manera continua y ágil, hemos utilizado **WhatsApp** como una vía adicional de comunicación en la que hemos podido resolver consultas rápidas, coordinar tareas y mantenernos al tanto de cualquier actualización o cambio en el proyecto.

Por último, también hemos utilizado **Google Drive** como una plataforma para compartir todos los documentos relacionados con el proyecto. Desde archivos de texto y hojas de cálculo hasta documentos colab y documentos csv con los datos, Google Drive nos ha permitido tener una ubicación centralizada donde todos los miembros del equipo han podido acceder y actualizar todos los documentos de manera colaborativa.

Planificación y asignación de tareas

En este punto es importante destacar que hemos mantenido reuniones recurrentes de aproximadamente dos horas cada miércoles.

Responsable	Tarea	Fecha de entrega
Equipo	Definición de términos, firma de NDA, estudio de modelos Churn y Estudio y análisis del contexto de negocio.	Reuniones 1, 2, 3 y 4.
Bruno	Subir el CSV con la data de Venues.	Reunión 5
Equipo	Subir los notebooks con el análisis exploratorio de las variables asignadas del CSV Venues.	Reunión 6
Bruno	Subir el CSV con la data de Orders.	Reunión 6
Arnau y Cami	Creación de la variable dependiente “Churn”	Reunión 6
Equipo	Revisión de los notebooks con el análisis exploratorio de las variables de Venues.	Reunión 7
Bruno y Uri	Revisión de la variable dependiente “Churn”	Reunión 7
Equipo	Subir los notebooks con el análisis exploratorio de las variables asignadas del CSV Orders.	Reunión 7
Uri	Unificación de los notebooks con el análisis exploratorio de las variables de Venues.	Reunión 7
Equipo	Revisión de los notebooks con el análisis exploratorio de las variables de Orders.	Reunión 8
Equipo	Definición de las variables del dataset final.	Reunión 8
Bruno	Creación del documento plantilla para la segunda entrega.	Reunión 8
Uri	Unificación de los notebooks con el análisis exploratorio de las variables de Orders.	Reunión 8

Equipo	Creación del dataset definitivo.	Reunión 9
Bruno	Merge y verificación de la consistencia de los datos.	Reunión 10
Cami	Balance y verificación de la consistencia de los datos.	Reunión 10
Arnau y Uri	Creación del primer modelo random forest.	Reunión 10
Equipo	Revisión de los datos y las variables.	Reunión 11
Equipo	Revisión y mejora del modelo random forest.	Reunión 11
Equipo	Documentación de todo el trabajo realizado.	Reunión 11
Equipo	Revisión de la documentación y todo el trabajo realizado.	Reunión 12
Equipo	Entrega de todo el trabajo realizado.	Reunión 12

Estimación de costes

Estimación de costes: El coste total del proyecto depende del número de horas que dedicadas y del salario de los Data Scientists. Suponiendo que el salario medio de un Data Scientist es de 40.000 euros al año y el proyecto tiene una duración de 6 meses y que cada uno de los miembros del equipo dedica una media de 8 horas por semana, el coste total del proyecto sería de 17.664 euros.

8. Requerimientos

A partir de la conversación con nuestro cliente sobre los usos del modelo propuesto, se llega a los siguientes requerimientos:

- La empresa ha solicitado la identificación de la probabilidad de que un cliente realice "churn", es decir, que abandone la empresa. Se define la fuga de cliente como aquel que haya abandonado la empresa (según la variable disabled_date) y que no haya regresado en un periodo de 3 meses calendario.
- La empresa dispone de un sistema de llamadas para fidelizar a los clientes, y se espera que la lista de clientes a contactar se genere a partir del modelo propuesto.
- Además de obtener el detalle de los clientes propensos a churn, se busca identificar las variables que más influyen en el abandono de un cliente. La empresa utilizará estas variables con el objetivo de reducir la tasa de churn.
- Se ha solicitado trabajar con modelos de ensamblado debido a su simplicidad y facilidad de comprensión, así como para realizar comparaciones con un modelo básico previamente implementado utilizando Random Forest en la empresa.
- El alcance del trabajo comprenderá el periodo a partir de junio de 2021. Para el entrenamiento del modelo se utilizará información correspondiente al periodo de enero a junio de 2023, y finalmente se validará con datos recopilados desde junio de 2023 hasta octubre de 2023.

9. Fuentes de Información Adicional y Benchmark

Fuente de Información Adicional: En la plataforma interna de Treatwell se encuentran disponibles varias páginas de documentación en Confluence, las cuales proporcionan información detallada y específica para abordar consultas y dudas particulares. Además, se ha implementado una herramienta de visualización (Looker), que permite acceder, visualizar y descargar información relevante. Para asegurar la integridad de los datos y obtener respuestas precisas, contamos con el apoyo del equipo de Analytics Engineer, quienes están disponibles para resolver preguntas relacionadas con los datos y asegurar la consistencia de la información.

Adicionalmente, se ha desarrollado un diccionario de datos que ha sido compartido con el equipo, con el objetivo de aclarar cualquier duda en relación a los significados de las variables analizadas. Este recurso se ha diseñado para facilitar la comprensión y correcta interpretación de los datos utilizados en los análisis y reportes.

En resumen, en Treatwell contamos con una variedad de recursos y herramientas para respaldar la búsqueda de información precisa y la resolución de dudas, incluyendo la documentación en Confluence, la herramienta de visualización Looker, el equipo de Analytics Engineer y un diccionario de datos. Estos recursos están disponibles para garantizar la calidad y coherencia en nuestros análisis y decisiones basadas en datos.

Benchmark: Dentro de la empresa, en el año 2019, existió un equipo de Data Science que realizó un Venue Churn Prediction Model que consistió en lo siguiente (*con un enfoque de Data distinto al nuestro*):
“El objetivo de este proyecto era identificar cuáles son los venues que tienen más probabilidades de abandonar y señalarlos a los equipos de suministro y marketing en un esfuerzo por reducir la tasa de abandono de venues. Para hacerlo, recopilamos métricas que pensamos que serían relevantes para determinar si un lugar abandonará o no y, utilizando un modelo de aprendizaje automático, obtuvimos un puntaje de probabilidad de churn para cada venue. El modelo obtuvo una puntuación ROC AUC de prueba de 0,84. Dos características se destacan como las más importantes para determinar el churn, a saber, las reservas del mercado en los últimos 30 días y una medida de la actividad de Connect, los empleados activos diarios durante la última semana sobre los empleados activos semanales durante la última semana (dae_over_wae_l1w).”
Sin embargo, usar este modelo como Benchmark no sería aplicable ya que:

- Con el tiempo, los datos disponibles pueden cambiar o evolucionar. Si no actualizas tu modelo con los datos más recientes, es posible que no refleje las tendencias o patrones actuales, lo que afectaría la precisión de las predicciones.
- Cambios en el contexto: El contexto en el que se aplican los modelos de aprendizaje automático también cambian.
- Si los datos de entrenamiento de 2019 no son representativos de la realidad actual, es probable que el modelo no pueda generalizar correctamente y producirá resultados sesgados o erróneos.

- Es posible que los algoritmos y técnicas utilizados en el modelo de 2019 no sean los más eficientes o precisos disponibles en 2023. Usar un modelo más actualizado podría mejorar los resultados.

Por consiguiente, se empleará como punto de partida y como concepto general para comprender las exigencias comerciales y de mercado del año 2019, así como para determinar su pertinencia continua para el negocio en el año actual.

De esta manera, a través de nuestro avance, crearemos un modelo de referencia para mejorar progresivamente el desempeño del modelo y, por ende, aumentar la precisión del Modelo de Predicción de Churn.

10. Desarrollo

Recopilación y preparación de datos: Se recopilaron y prepararon los datos relevantes relacionados con los salones de belleza, incluyendo información sobre su historial de transacciones, servicios utilizados, duración de la relación con Treatwell, pagos, órdenes y cualquier otro dato relevante.

Identificación de variables predictoras: Se analizaron los datos recopilados para identificar las variables que podrían influir en la probabilidad de cancelación de los salones.

Construcción del modelo de churn: Se emplearon técnicas de aprendizaje automático (modelos de ensamblado) y análisis predictivo para desarrollar un modelo de churn que permita calcular la probabilidad de cancelación para cada salón.

Validación del modelo: Se evaluó el rendimiento del modelo utilizando técnicas de validación cruzada y métricas adecuadas. Esto permitiría determinar la capacidad del modelo para predecir correctamente qué salones son más propensos a cancelar su relación con Treatwell.

Acciones basadas en las predicciones: Una vez que el modelo esté validado, se utilizarían las predicciones de propensión para identificar los salones de belleza que tienen una alta probabilidad de cancelar y de esta manera Treatwell pueda tomar acciones proactivamente para retenerlos. Estas acciones podrían incluir programas de incentivos, mejoras en los servicios ofrecidos, comunicación directa con los salones, entre otras estrategias.

Es importante tener en cuenta que el éxito de este proyecto también dependerá de la disponibilidad de datos confiables y completos, así como de los recursos y la capacidad de implementación de Treatwell Spain SL.

En el marco del proyecto, actualmente nos encontramos en la etapa de validación del primer modelo de churn. Hemos aplicado técnicas de validación cruzada y estamos evaluando las métricas pertinentes para mejorar el rendimiento y la precisión de nuestro modelo. Este es solo el comienzo de nuestros esfuerzos, ya que desarrollaremos más

modelos en el futuro y comparemos sus métricas para ver cuál se adecua mejor al caso de negocio.

Una vez que hayamos completado la validación de los modelos y estemos satisfechos con sus rendimientos, seleccionaremos el modelo más adecuado y se lo presentaremos a Treatwell para que lo utilicen, y así, puedan realizar acciones basadas en las predicciones.

11. Hipótesis a probar

- Mientras más funcionalidades de Treatwell use la venue, menor será la probabilidad de churn.
- Mientras más tiempo se relaciona la venue con Treatwell, menor será la probabilidad de churn.
- Las venues que tienen una mayor cantidad de net_orders, tienen una menor probabilidad de churn.
- Las venues que tienen una mayor cantidad de Appointments_l30d, tienen una menor probabilidad de churn.
- Si la venue se ha ido y vuelve (reactivada), debería aumentar la lealtad de la venue, y por tanto, disminuir su probabilidad de churn.
- Si la venues está listada en el marketplace deberían ser menos propensas a hacer churn.
- Si las venues solamente ofrecen la opción de hacer prepago (prepay_only) deberían ser más propensas a hacer churn.
- Si las venues pre-pagan el plan deberían tener menor tendencia a realizar Churn.
- Las venues que no están listadas en el marketplace (Is_tw_mp_listed = False) deberían ser más propensas a hacer churn.
- Las venues que solamente ofrecen la opción de hacer prepago (marketplace_payment_method = prepay_only) deberían ser más propensas a hacer churn.
- Las venues que solamente ofrecen la opción de hacer prepago (widget_payment_method = prepay_only) deberían ser más propensas a hacer churn.
- Las venues con menor cantidad de “net orders” deberían ser más propensas a hacer churn.
- Las venues con mayor cantidad de orders canceladas deberían ser más propensas a hacer churn.
- Las venues con menor “net_aov” deberían ser más propensas a hacer churn.
- Las venues con mayor “net_take_rate” deberían ser más propensas a hacer churn.

12. Trabajo realizado con las variables

La estructura de datos de Treatwell explora con 2 grandes entidades: Las venues que representan distintos tipos de salones y ‘Orders’ que representa las citas que son

agendadas a la venue. A partir de lo anterior se procede a explicar el trabajo en detalle que se realizó en cada caso.

12.1. Variable originales venues

- ### Created

Significado de la variable
Fecha de creación de la venue en el sistema.
Formato: YYYY-MM-DD.
Conclusiones del análisis realizado
Cada venue presenta solo una fecha de creación aunque esta haya abandonado Treatwell y haya vuelto. La data presenta valores desde '2008-03-25' hasta '2023-06-01'. Durante los meses de Junio, Julio y Agosto se presenta una baja en la creación de venues. Esta baja es sostenida en los años
Conclusiones en su uso en el modelo
Aunque esta variable podría ser un buen candidato para la definición de la antigüedad de la venue en el sistema, no se utilizará en el modelo dado que existe una mejor variable 'latest_live_date' para el mismo fin.

- ### Updated

Significado de la variable
Fecha de actualización de algún elemento de la venue en el sistema.
Formato: YYYY-MM-DD.
Conclusiones del análisis realizado
Cada venue presenta desde 0 fechas de actualizaciones, es decir, esta no fue actualizada posterior a su creación. Hasta 4 fechas distintas de actualizaciones.
Conclusiones en su uso en el modelo
Dado que la data fue obtenida leyendo la tabla histórica una vez en el mes, la actualización de algún elemento se ve reflejado en la observación de ese mes. En conclusión, la variable no aporta información relevante para el modelo.

- ### Venue_Status

Significado de la variable
Character que representa el estado que presenta la venue para el sistema.
Tipo de Dato: Cualitativa Nominal.
Conclusiones del análisis realizado
Existen 4 posibles valores que puede tomar la variable:
LIVE (551,822), NaN (414,765), DISABLED (303,953), PENDING (28,704), ARCHIVED(24,053). Estos valores están asociados si la venue se ha ido, es decir, terminado su contrato, por cuanto tiempo se ha ido, etc. También es importante mencionar su alto número de valores nulos. Existen venues sin estado asociado (10,463 venues) y venues que presentan hasta 3 estados en su historia.
Conclusiones en su uso en el modelo

Los valores presentes como estados en esta variable son una respuesta a la variable respuesta que se busca predecir. Por lo anterior no puede ser una variable que el modelo tenga en consideración para la predicción de churn.

- ## Venue_Active_From

Significado de la variable
Fecha en que la venue comienza a estar activa en el sistema.
Formato: YYYY-MM-DD.
Conclusiones del análisis realizado
Existe un gran porcentaje de valores nulos en la variable llegando al 92% de su total. En concreto, 178.462 venues no tienen fecha de activación en alguno de sus 'date_days'. Por problemas en la migración de sistemas, la variable presenta valores nulos.
Conclusiones en su uso en el modelo
La variable no agrega información de negocio al modelo, en consecuencia, no se utilizará.

- ## Original_marketplace_venue_type_name

Significado de la variable
Variable de tipo character que representa el tipo de negocio de la venue. Esto representa el tipo de lugar estático asignado a un lugar en TW Marketplace.
Tipo de Dato: Cualitativa Nominal.
Conclusiones del análisis realizado
La variable presenta 29 valores distintos. 'Hair Salon' es el tipo de venue más típico con 492.160 records, mientras 'Hammam' es el menos típico con 39 records. Existen venues con más de un 'type' en su historia, esto pueden llegar a ser 3.
Conclusiones en su uso en el modelo
La variable es valiosa para identificar qué tipo de venue son más propensas a hacer churn, sin embargo, al tener tantos subtipos no se logra tener información suficiente de cada uno para entrenar un modelo. Se procede a mapear la variable de la siguiente manera, siguiendo la lógica de negocio en Treatwell logrando un conjunto de 11 tipos de venues. El mapeo se presenta en Apendice 1.
Con esta transformación el tipo de venue con menos ejemplares es 'Face Salon' con 16.735 records de información. La variable se procede a trabajar con OneHotEncoder dado que es una variable nominal.

- ## Country_code

Significado de la variable
Variable categórica que representa el país donde opera la venue.
Tipo de Dato: Cualitativa Nominal.
Conclusiones del análisis realizado
La variable presenta 16 valores distintos. Los últimos 3 valores tienen una frecuencia de 31, 4 y 1 respectivamente.
Conclusiones en su uso en el modelo
La variable es valiosa para identificar de dónde son las venue que son más propensas a hacer churn. Se procede a dejar los últimos 3 países como 'otros' en el modelo y de esta manera, incorporar nuevos países con baja frecuencia al modelo que serán mapeados a

esta categoría. La variable se procede a trabajar con OneHotEncoder dado que es una variable nominal.

- Tier

Significado de la variable
Variable relacionada a la locación y su nivel. Actualmente, los niveles de ciudad están definidos por los gerentes de los países, y el equipo de estrategia los recopila y los mantiene. Existen 2 opciones: Tier 1 y 2.
Tipo de Dato: Cualitativa Ordinal.
Conclusiones del análisis realizado
La variable presenta un alto porcentaje de valores nulos (49%) y su locación es también presentada en la variable country_code.
Conclusiones en su uso en el modelo
La variable no será utilizada en el modelo al no presentar valor para el análisis.

- Saas_product

Significado de la variable
Variable relacionada con el tipo de producto (SaaS) de Treadwell que la venue usa.
Tipo de Dato: Cualitativa Nominal.
Conclusiones del análisis realizado
La variable presenta dos valores distintos: tw_connect y tw_pro. Siendo el primero el más frecuente. Pueden existir más tipos de SaaS, sin embargo, se seleccionaron los que entregan ganancias a la empresa y en los cuales la empresa está interesado en modelar.
Conclusiones en su uso en el modelo
La variable representa dos tipos de software y se considera relevante para el modelo. se procede a trabajar con OneHotEncoder.

- First_live_date

Significado de la variable
Fecha que indicar la primera vez que se activa un plan de facturación de pago o la primera fecha publicada del venue (el que ocurra primero).
Formato: YYYY-MM-DD
Conclusiones del análisis realizado
Para garantizar la integridad de los datos, se eliminaron todos los registros que presentaron valores NaN en la columna "first_live_date". Este ajuste es necesario, dado que dichos valores nulos parecen ser resultado de un error en los datos, representando aproximadamente un 22.63% del dataset (sobre 4.606.923 observaciones). La serie temporal abarca desde el año 2008 hasta el año 2023.
Conclusiones en su uso en el modelo
La variable por sí sola no nos entrega información relevante para el modelo. Al existir la variable Latest_live_date en el dataset, se descarta la variable del modelo. Sin embargo cuando latert_live_date era nula y first_live_date no lo fue, se utilizó para imputar nulos.

- Latest_live_date

Significado de la variable

La fecha más reciente en que comenzó un plan de facturación pagado o la fecha publicada más reciente (el que ocurra primero). Si sigue un plan posterior dentro de los 5 días de un plan anterior, se muestra la fecha de vigencia del plan anterior.
Formato: YYYY-MM-DD
Conclusiones del análisis realizado
Para garantizar la integridad de los datos, se eliminaron los registros que presentaron valores NaN en la columna "latest_live_date" y "first_live_date". Este ajuste es necesario, dado que dichos valores nulos parecen ser resultado de un error en los datos, representando aproximadamente un 23.053565% del dataset (sobre 4.606.923 observaciones). La serie temporal abarca desde el año 2008 hasta el año 2023. Adicionalmente, se ha observado que un 11.4% de las Venues ha sido deshabilitada en algún momento y luego ha vuelto a trabajar con Treatwell, lo que indica la posibilidad de cambios o reactivaciones de venues. Esta información proporciona un contexto adicional sobre la actividad de las venues en el dataset.
Conclusiones en su uso en el modelo
La variable por sí sola no nos entrega información relevante para el modelo., sin embargo, a partir de esta variable y disabled_date podemos calcular el tiempo de existencia de la Venue.

- ## Disabled_date

Significado de la variable
Es el primer día completo en que el venue no tiene acceso al SaaS (un salón que esté deshabilitado el 30 de junio de 2022 a las 14:00 p. m. mostrará el 1 de julio de 2022 como fecha_deshabilitada).
Formato: YYYY-MM-DD
Conclusiones del análisis realizado
Variable más importante del modelo ya que a partir de ella podemos calcular el flag de Churn, que es la variable a predecir. La serie temporal abarca desde el año 2009 hasta el año 2023. Adicionalmente, la variable "disabled_date" muestra un porcentaje de null de 73.58%, lo cual indica que aproximadamente el 73% de las observaciones (venues-date) no han sido deshabilitadas durante el periodo estudiado de un total de 4.606.923 observaciones.
Conclusiones en su uso en el modelo
Variable usada para el cálculo del flag de Churn.

- ## Reactivated_status

Significado de la variable
Se considera que una venue está reactivada cuando la diferencia entre disabled_date y latest_live_date es inferior a 6 meses (reactivated). Si el salón se desconecta y vuelve a estar activo después de 6 meses, se considera que la venue está reactivada new cuando la diferencia entre disabled_date y latest_live_date es mayor a 6 meses (reactivated_new). Si no, la variable toma el valor None.
Tipo de Dato: Cualitativa Nominal.
Conclusiones del análisis realizado
La variable "reactivated" tiene tres categorías distintas con sus correspondientes recuentos: "none": 98.60%, "reactivated": 0.84% y "reactivated new": 0.55%. Esto significa que la categoría "none" es la más común, con un porcentaje del 98.60%. La categoría "reactivated" tiene un porcentaje del 0.84%, y la categoría "reactivated new" tiene un porcentaje del 0.55%.
Conclusiones en su uso en el modelo
Variable no relevante debido a la definición de empresa (6 meses). Se calculará una nueva definición de reactivated status en base a las variables latest_live_date y disabled_date llamada Reactivada.

- ## Is_twconnect_migrated

Significado de la variable
El venue se ha migrado de Treatwell Connect a Treatwell Pro.
Tipo de Dato: Cualitativa Dicotómica.
Conclusiones del análisis realizado
La variable presenta un 99.47% de False y un 0.53% de True, lo que nos indica que no es relevante debido a la poca variabilidad de la variable (<1% es True).
Conclusiones en su uso en el modelo
La variable no será utilizada en el modelo debido a que se trata de una variable irrelevante para la predicción del churn y no se relaciona con el actuar de la venue.

- is_heavy_saas

Significado de la variable
Variable definida por la empresa para saber si un venue ocupa el SaaS o no. Entre las condiciones están (mensualmente):

- 40 citas de ecosistema por estilista activo
- Al menos 10 citas directas por estilista activo

Tipo de Dato: Cualitativa Dicotómica.
Conclusiones del análisis realizado
La variable presenta un 85.61% de False y un 14.38% de True, sin embargo, al ser una nueva variable generada por la empresa (se desarrolló en el año 2022) se empieza a medir a partir de las venues con latest_live_date desde 2015. Es decir, perdemos 7 años de información para la modelación (ya que tenemos latest_live_date a partir del año 2008), y por definición, el sistema los anota como False generando un sesgo en la data (no sabemos cuales venues son False reales o por sistema). Por lo tanto, se descarta. Sería interesante usarla en un futuro cuando la métrica esté presente en todo el dataset. De todas maneras, sería relevante si hacemos una transformación para considerar, por ejemplo, antigüedad de is_heavy_saas.
Conclusiones en su uso en el modelo
La variable se descarta del modelo ya que la data no está presente de manera consistente en el dataset al ser una variable nueva, generando un sesgo en el modelo.

- First_Heavy_Saas_Date

Significado de la variable
Variable relacionada con la fecha en la que la venue consiguió ser “First Heavy Saas” por primera vez.
Formato: YYYY-MM-DD.
Conclusiones del análisis realizado
La variable en cuestión no es relevante para el modelo, ya que lo importante es determinar si la venue pertenece a la categoría “heavy_Saas”, no la fecha en que comenzó a serlo. Cabe destacar que esta variable presenta un alto porcentaje de valores nulos (98%).
Conclusiones en su uso en el modelo
La variable no será utilizada en el modelo debido a que se trata de una variable irrelevante para la predicción del churn.

- Active_Employees

Significado de la variable
Representa la cantidad de empleados de cada venue.
Tipo de Dato: Cuantitativa Discreta.
Conclusiones del análisis realizado
Se trata de una variable bastante dispersa con un valor mínimo de 1 empleado, un valor máximo de 68 empleados y un valor medio de 2-3 empleados por venue. Además, la variable presenta un alto porcentaje de valores nulos (80%). Hemos consultado con el equipo de data de Treatwell y nos han comentado que este alto porcentaje se debe a problemas de “Data Quality” (en este caso, sólo se ha mapeado la variable al día presente perdiendo la historia de la variable).
Conclusiones en su uso en el modelo

La variable no será utilizada en el modelo debido a la cantidad de valores nulos que presenta.

- ## Direct_Appointments_l30d

Significado de la variable
Representa la cantidad de “appointments” (citas) que la venue ha añadido a través del Saas durante los últimos 30 días.
Tipo de Dato: Cuantitativa Discreta.
Conclusiones del análisis realizado
Se trata de una variable muy dispersa con un valor mínimo de 0 “direct appointments”, un valor máximo de 21.488 “direct appointments” y un valor medio de 56 “direct appointments” por venue. Además, la variable presenta un moderado porcentaje de valores nulos (22%). Por último, mencionar que la variable presenta una distribución sesgada hacia la derecha, con una gran cantidad de valores pequeños y pocos valores extremadamente grandes. Esto se refleja en su cola larga hacia la derecha, lo que indica la presencia de valores atípicos muy grandes.
Conclusiones en su uso en el modelo
La variable es valiosa para identificar qué tipo de venue son más propensas a hacer churn. Sin embargo, hemos decidido descartar esta variable y utilizarla para crear dos nuevas variables que nos pueden aportar más información.

- Appointments_l30d: Se trata de la suma de las variables “direct_appointments_l30d” y “online_appointments_l30d”. Nos aportará la cantidad total de “appointments” que ha tenido cada venue durante los últimos 30 días.
- Online_Appointments_Rate: Se trata de la división entre las variables “online_appointments_l30d” y “appointments_l30d”. Nos aporta el ratio de “appointments” que vienen por parte de los clientes a través del Saas durante los últimos 30 días.

- ## Online_Appointments_l30d

Significado de la variable
Representa la cantidad de “appointments” (citas) que los clientes han pedido a través del Saas durante los últimos 30 días.
Tipo de Dato: Cuantitativa Discreta.
Conclusiones del análisis realizado
Se trata de una variable muy dispersa con un valor mínimo de 0 “online appointments”, un valor máximo de 4.039 “online appointments” y un valor medio de 15-16 “online appointments” por venue. Además, la variable presenta un moderado porcentaje de valores nulos (22%). Por último, mencionar que la variable presenta una distribución sesgada hacia la derecha, con una gran cantidad de valores pequeños y pocos valores extremadamente grandes. Esto se refleja en su cola larga hacia la derecha, lo que indica la presencia de valores atípicos muy grandes.
Conclusiones en su uso en el modelo
La variable es valiosa para identificar qué tipo de venue son más propensas a hacer churn. Sin embargo, hemos decidido descartar esta variable y utilizarla para crear dos nuevas variables que nos pueden aportar más información.

- Appointments_l30d: Se trata de la suma de las variables “direct_appointments_l30d” y “online_appointments_l30d”. Nos aportará la cantidad total de “appointments” que ha tenido cada venue durante los últimos 30 días.
- Online_Appointments_Rate: Se trata de la división entre las variables “online_appointments_l30d” y “appointments_l30d”. Nos aporta el ratio de “appointments” que vienen por parte de los clientes a través del Saas durante los últimos 30 días.

● Is_Tw_Mp_Listed

Significado de la variable
Representa si la venue está listada en el marketplace o no.
Tipo de Dato: Cualitativa Dicotómica.
Conclusiones del análisis realizado
La variable presenta un 5.92% de valores “True” y un 94.08% de valores “False”. Además, la variable no presenta valores nulos.
Conclusiones en su uso en el modelo
La variable es valiosa para identificar qué tipo de venue son más propensas a hacer churn. Según nuestra hipótesis, las venues que no están listadas en el marketplace (Is_tw_mp_listed = False) deberían ser más propensas a hacer churn.

● Tw_Mp_Listed_First_Time_Date

Significado de la variable
Variable relacionada con la fecha en la que la venue consiguió ser “Tw_Mp_Listed” (pertenecer al marketplace) por primera vez.
Formato: YYYY-MM-DD.
Conclusiones del análisis realizado
La variable en cuestión no es relevante para el modelo, ya que lo importante es determinar si la venue pertenece a la categoría “tw_mp_listed”, no la fecha en que comenzó a serlo. Además, cabe destacar que esta variable presenta un moderado porcentaje de valores nulos (14.6%).
Conclusiones en su uso en el modelo
La variable no será utilizada en el modelo debido a que se trata de una variable irrelevante para la predicción del churn.

● Marketplace_Payment_Method

Significado de la variable
Representa los métodos de pago disponibles en la venue.
Tipo de Dato: Cualitativa Nominal.
Conclusiones del análisis realizado
Existen 4 posibles valores que puede tomar la variable:
- All_Methods (84.81%)
- Prepay_Only (8.98%)
- Pay_at_Venue_Only (1.59%)
- Unknown (0.18%)

Además, la variable presenta un bajo porcentaje de valores nulos (4.45%). En este caso, al tener la categoría “Unknown”, hemos decidido unificar los valores nulos dentro de esta categoría.

Conclusiones en su uso en el modelo

La variable es valiosa para identificar qué tipo de venue son más propensas a hacer churn. Según nuestra hipótesis, las venues que solamente ofrecen la opción de hacer prepago (marketplace_payment_method = prepay_only) deberían ser más propensas a hacer churn.

- Widget_Payment_Method

Significado de la variable

Representa los métodos de pago disponibles en la venue en el widget.

Tipo de Dato: Cualitativa Nominal.

Conclusiones del análisis realizado

Existen 4 posibles valores que puede tomar la variable:

- All_Methods (81.77%)
- Pay_at_Venue_Only (6.85%)
- Prepay_Only (6.75%)
- Unknown (0.18%)

Además, la variable presenta un bajo porcentaje de valores nulos (4.45%). En este caso, al tener la categoría “Unknown”, hemos decidido unificar los valores nulos dentro de esta categoría.

Conclusiones en su uso en el modelo

La variable es valiosa para identificar qué tipo de venue son más propensas a hacer churn. Según nuestra hipótesis, las venues que solamente ofrecen la opción de hacer prepago (widget_payment_method = prepay_only) deberían ser más propensas a hacer churn.

- Prepay

Significado de la variable

Variable que indica si la venue prepago su plan o no.

Tipo de Dato: Cualitativa Dicotómica.

Conclusiones del análisis realizado

La variable presenta una distribución en la cual el 93.78% de los casos tiene un valor de 0, mientras que el 6.2% de los casos tiene un valor distinto de cero. No se registran valores nulos en esta variable.

Conclusiones en su uso en el modelo

La variable será utilizada en el modelo debido a que se trata de una variable relevante para la predicción del churn y para el negocio, ya que existe la hipótesis que si prepagas el plan deberías tener menor tendencia a realizar Churn.

- Plan_name

Significado de la variable

El nombre del plan de pago (billing plan) al que está asociado a la venue.

Tipo de Dato: Cualitativa Nominal.

Conclusiones del análisis realizado

Variable Relevante para el negocio que presenta múltiples estados (33 estados diferentes). Se realiza un mapping de la variable para reducir la dimensionalidad que puede ser visualizado en el apéndice 2. La variable Partner SaaS se elimina ya que no es un plan de pago. Debido a la migración de Data, solo existe un 60% de las venues que presentan plan de pago. Por lo tanto, se realizarán 2 modelos, uno con todas las variables eliminando los NA de plan_name (alrededor de 1.5M de observaciones) y otra con la todas las variables menos plan_name. Esto ya que es una variable relevante para el negocio.

Conclusiones en su uso en el modelo

La variable será utilizada en el modelo debido a que se trata de una variable relevante para la predicción del churn y el negocio.

- Is_plan_trial

Significado de la variable

Indica si el cliente tiene un plan de prueba, lo cual significa que su antigüedad es menor a 3 meses. Tiene un formato binario en el que 1 indica que el cliente está en período de prueba y 0 indica lo contrario.

Tipo de Dato: Cualitativa Dicotómica.

Conclusiones del análisis realizado:

Se encontró que esta variable está altamente desbalanceada, con solo un 2.4% de las venues con un plan de prueba. La fecha más antigua en la que se registra 'is_plan_trial' como 1 es en diciembre de 2021, mientras que la más reciente es en junio de 2023. Debido a esto, la variable presenta un alto porcentaje de valores nulos, alcanzando el 41.51%. Además, se observó que ninguno de los valores de 'is_plan_trial' igual a 1 está asociado con churn .

Conclusiones en su uso en el modelo

Basado en el análisis realizado, se concluye que la variable 'is_plan_trial' no es relevante para el modelo, ya que no aporta información significativa. Además, esta variable está altamente relacionada con la variable de antigüedad ('antigüedad'). Cabe destacar que ninguno de los venues con un plan de prueba presenta churn, lo que refuerza su falta de influencia en el modelo. Además, debido al alto porcentaje de valores nulos, su utilidad se ve aún más limitada.

- Plan_active_from

Significado de la variable

Indica desde cuando el venue inicia un plan en Treatwell.

Formato: YYYY-MM-DD.

Conclusiones del análisis realizado:

Se encontró que esta variable está altamente correlacionada con la variable first_live_date, y disabled_date.

Conclusiones en su uso en el modelo

La variable no se incluye en el modelo porque no aporta información adicional, ya existe en nuestro dataset la variable antigüedad, que capta la duración en meses en los cuales la venue estuvo activa

- Plan_active_to

Significado de la variable

Indica cuando el venue termina su plan en Treatwell.
Formato: YYYY-MM-DD.
Conclusiones del análisis realizado:
Se encontró que esta variable está altamente correlacionada con la variable latest_live_date,
Conclusiones en su uso en el modelo
La variable no se incluye en el modelo porque no aporta información adicional, ya existe en nuestro dataset la variable antigüedad, que capta la duración en meses en los cuales la venue estuvo activa

- Discount

Significado de la variable
Indica si la venue tiene descuento en su plan, lo cual significa que paga una cuota menor por cada mes o no. Tiene un formato binario en el que 1 indica si la venue tiene descuento y 0 indica lo contrario.
Tipo de Dato: Cualitativa Dicotómica.
Conclusiones del análisis realizado:
La variable 'discount' está altamente desbalanceada, con solo un 0.74% de las venues con un descuento aplicado en su cuota mensual. Sin embargo, cabe destacar que de los clientes que tienen un descuento aplicado solo un 5.34% presentan churn, mientras que de los clientes que no tienen el descuento aplicado este porcentaje se eleva a 21.7%. En este sentido, se puede concluir que una cuota mensual con descuento conlleva un churn menor.
Conclusiones en su uso en el modelo
Esta variable se decidió descartar en el uso del modelo por la falta de información alrededor de 'discount'. En primer lugar, convendría saber porqué se aplica el 'discount' en algunas venues mientras que en otras no, las condiciones y el % de descuento aplicadas en cada una de estas. Solo de esta manera podremos asesorar el impacto y las consecuencias que tiene con el churn rate.

- Plan_discount_end_date

Significado de la variable
Indica, por las venues que tienen descuento, la fecha en el que finaliza.
Formato: YYYY-MM-DD.
Conclusiones del análisis realizado:
La variable 'discount' tiene la fecha más antigua en el 2019-05-29 y una fecha futura del 2024-06-06. Es una variable interesante de analizar creando el período de discount_plan por cada unique_venue_id y analizando estadísticamente su impacto con churn.
Conclusiones en su uso en el modelo
Esta variable se decidió descartar por el elevado número de valores nulos (99.27%).

- Bill_every

Significado de la variable:
Nos indica, cada cuando se tasa, en meses, a las venues de Treatwell.
Tipo de Dato: Cuantitativa Discreta.
Conclusiones del análisis realizado:

Bill_every tiene una media de 1 meses, lo cual nos indica, que la gran mayoría de venues pagan de manera mensual.
Conclusiones en su uso en el modelo
Esta variable se decidió descartar por su falta de relevancia para predecir churn

- Prepaid_until_to

Significado de la variable
Nos indica el número de meses que la venue tiene prepagados.
Tipo de Dato: Cuantitativa Discreta.
Conclusiones del análisis realizado El análisis no se pudo realizar debido a la gran presencia de nulls (94%).
Conclusiones en el modelo:
Debido a la gran presencia de nulls, tampoco se tendrá en cuenta para el modelo

- is_zero-commission_plan

Significado de la variable
Nos indica si la venue tiene un plan de comisiones del 0%.
Tipo de Dato: Cualitativa Dicotómica.
Conclusiones del análisis realizado:
Para los salones con "is_zero_commission_plan" igual a False, hay 628,888 salones que no han churned (churn = 0), lo que representa aproximadamente el 38.9% del total de salones en esta categoría. Además, hay 986,529 salones que sí han churned (churn = 1), lo que representa aproximadamente el 61.1% del total de salones en esta categoría. Para los salones con "is_zero_commission_plan" igual a True, hay 272,899 salones que no han churned, lo que representa aproximadamente el 95.3% del total de salones en esta categoría. Por otro lado, hay 13,355 salones que sí han churned, lo que representa aproximadamente el 4.7% del total de salones en esta categoría. Hay una proporción más alta de salones con "is_zero_commission_plan" igual a False que han churned en comparación con los salones con "is_zero_commission_plan" igual a True. Esto sugiere una posible relación entre tener una comisión cero (False) y una mayor probabilidad de churn.
Conclusiones en su uso en el modelo
Esta variable se decidió descartar por su alta correlación con la variable is_tw_mp_listed.

- Plan_fee_eur

Significado de la variable
Nos indica el valor del pago de la cuota que las venues deben hacer a Treatwell en euros.
Tipo de Dato: Cuantitativa Nominal.
Conclusiones del análisis realizado:
Los clientes de Treatwell tienen planes de pagos que van de los 0 a los 828 euros y pagan 25.24 euros de media. El coeficiente de correlación biserial puntual es de -0.349. Esta cifra indica una correlación negativa moderada entre la variable continua "plan_fee_eur" y la variable binaria "churn". Esta correlación negativa sugiere que a medida que aumenta el valor de "plan_fee_eur", la cuota que deben pagar las venues de Treatwell, existe una tendencia a disminuir la probabilidad de churn.
Conclusiones en su uso en el modelo

Esta variable se decidió descartar porque presenta una correlación muy elevada con la cantidad de net_orders que presentan los clientes por cada venue, variable que tiene una mayor capacidad predictiva sobre churn

- is_purchasable

Significado de la variable
Indica si al menos un servicio está disponible para su compra en el marketplace. Puede tener dos posibles valores: "Verdadero" si el servicio se puede comprar y "Falso" si no es posible adquirirlo.
Tipo de Dato: Cualitativa Dicotómica.
Conclusiones del análisis realizado:
La proporción de establecimientos con is_purchasable = True es del 26.06%. Para is_purchasable=False, el porcentaje de churn es de 39.6% y el porcentaje de no churn es de 60.4%. Para is_purchasable=True, el porcentaje de churn es de 8.4% y el porcentaje de no churn es de 91.6%.Esto nos indica que las venues que no tienen habilitado la posibilidad de compra son mucho más propensos a churn que aquellas venues que si que lo tienen habilitado.
Conclusiones en su uso en el modelo: Se creó una nueva variable llamada "is_purchasable_duration" qué tiene en cuenta la duración durante la cual un servicio con "is_purchasable=True" estuvo activo. La variable no solo tiene en cuenta si el servicio está a la venta , sino también cuánto tiempo estuvo disponible para su compra.

- is_widget_enabled

Significado de la variable
Indica si la venue tiene o no un widget habilitado. Es una variable booleana que tiene dos opciones posibles, True, si el widget está habilitado y False cuando no lo está.
Tipo de Dato: Cualitativa Dicotómica.
Conclusiones del análisis realizado:
La proporción de establecimientos con is_widget_enabled= True es del 26.84%. Es más probable que los clientes con is_widget_enabled = False abandonen el servicio en comparación con aquellos con is_widget_enabled establecido = True. El porcentaje de churn de clientes con is_widget_enabled establecido en False es del 39,8 %, que es superior al porcentaje de abandono del 9,5 % para clientes con is_widget_enabled establecido en True.
Conclusiones en su uso en el modelo
Se creó una nueva variable llamada “is_widget_enabled_duration" que tiene en cuenta la duración durante la cual un servicio con "is_widget_enabled_duration=True" estuvo activo. La variable no solo tiene en cuenta si el servicio tiene un widget habilitado, sino también cuánto tiempo estuvo activo, en meses.

12.2. Variable originales Orders

- Year

Significado de la variable

Año de la fecha en la que las órdenes son creadas.
Formato fecha YYYY
Conclusiones del análisis realizado
La variable presenta los siguientes valores (años): 2021, 2022 y 2023.
Conclusiones en su uso en el modelo
La variable debe unirse con la variable 'Month' para hacer el merge con la información de venues.

- Month

Significado de la variable
Mes de la fecha en la que las órdenes son creadas.
Formato fecha MM
Conclusiones del análisis realizado
La variable presenta los valores (meses) del 1 al 12. Del análisis se observa que el mes de mayo presenta mayor concentración de órdenes.
Conclusiones en su uso en el modelo
La variable debe unirse con la variable 'Year' para hacer el merge con la información de venues.

- Content_Channel

Significado de la variable
Canal donde se realizó la orden: Widget, Book with Google o Marketplace.
Tipo de Dato: Cualitativa Nominal.
Conclusiones del análisis realizado
La variable presenta 3 posibles valores. El canal Marketplace es donde se generan más órdenes mientras que el canal 'Book with Google' es el canal que presenta menos órdenes. En un mes la venue puede tener órdenes creadas por todos los canales que haya contratado.
Conclusiones en su uso en el modelo
La variable será transformada para identificar cuantos canales tiene asociada la venue en un mes. Para lo anterior se usará la variable asociada a venues. Se busca probar la tesis de que con mayor uso de canales, mayor será la lealtad hacia treatwell y será menos probable que exista fuga.

- Gross_Orders

Significado de la variable
Representa la cantidad total de órdenes de una venue en un mes-año.
Tipo de Dato: Cuantitativa Discreta.
Conclusiones del análisis realizado
Se trata de una variable muy dispersa con un valor mínimo de 1 orden al mes, un valor máximo de 3.389 orders al mes y un valor medio de 54-55 orders al mes. Además, la variable no presenta valores nulos. Por último, mencionar que la variable presenta una distribución sesgada hacia la derecha, con una gran cantidad de valores pequeños y pocos valores extremadamente grandes. Esto se refleja en su cola larga hacia la derecha, lo que indica la presencia de valores atípicos muy grandes.
Conclusiones en su uso en el modelo

La variable es valiosa para identificar qué tipo de venue son más propensas a hacer churn. Sin embargo, hemos decidido descartar esta variable y utilizarla para crear dos nuevas variables que nos pueden aportar más información.

- Net_Orders: Se trata de la diferencia entre las variables “gross_orders” y “cancelled_orders”. Nos aportará la cantidad neta de órdenes realizadas por cada venue durante un mes-año.
- Orders_Cancellation_Rate: Se trata de la división entre las variables “cancelled_orders” y “gross_orders”. Nos aporta el ratio de órdenes que han sido canceladas durante un mes-año.

● Gross_Item_Eur_Amount_Ttv

Significado de la variable

Representa el valor total en euros de las transacciones de las órdenes de un mes-año de cada venue. TTV significa “Total Transaction Value”.

Tipo de Dato: Cuantitativa Continua.

Conclusiones del análisis realizado

Se trata de una variable muy dispersa con un valor mínimo de 0 euros al mes, un valor máximo de 124.495 euros al mes y un valor medio de 2.309 euros al mes. Además, la variable no presenta valores nulos. Por último, mencionar que la variable presenta una distribución sesgada hacia la derecha, con una gran cantidad de valores pequeños y pocos valores extremadamente grandes. Esto se refleja en su cola larga hacia la derecha, lo que indica la presencia de valores atípicos muy grandes.

Conclusiones en su uso en el modelo

La variable es valiosa para identificar qué tipo de venue son más propensas a hacer churn. Sin embargo, hemos decidido descartar esta variable y utilizarla para crear una nueva variable que nos pueda aportar más información.

- Net_TTV: Se trata de la diferencia entre las variables “gross_item_eur_amount_ttv” y “cancelled_item_eur_amount_ttv”. Nos aportará la cantidad neta en euros de las transacciones de las órdenes de un mes de cada venue.

● Gross_Aov

Significado de la variable

Representa el valor promedio de las órdenes de cada venue por mes. AOV significa “Average Order Value”.

Tipo de Dato: Cuantitativa Continua.

Conclusiones del análisis realizado

Se trata de una variable muy dispersa con un valor mínimo de 0 euros, un valor máximo de 22.057 euros y un valor medio de 404-405 euros. Además, la variable no presenta valores nulos. Por último, mencionar que la variable presenta una distribución sesgada hacia la derecha, con una gran cantidad de valores pequeños y pocos valores extremadamente grandes. Esto se refleja en su cola larga hacia la derecha, lo que indica la presencia de valores atípicos muy grandes.

Conclusiones en su uso en el modelo

No vamos a utilizar esta variable ya que se trata de una transformación de Treatwell que podemos calcular nosotros mismos a partir de las variables “net_ttv” y “net_orders”. La utilizaremos para crear una nueva variable que nos pueda aportar más información.

- Net_AOV: Se trata de la división entre las variables “net_ttv” y “net_orders”. Nos aportará el valor promedio neta en euros de las órdenes de un mes-año de cada venue.

● Gross_Revenue_Eur

Significado de la variable
Representa la ganancia total en euros que ha tenido la venue por las órdenes del respectivo mes-año.
Tipo de Dato: Cuantitativa Continua.
Conclusiones del análisis realizado
Se trata de una variable muy dispersa con un valor mínimo de 0 euros al mes, un valor máximo de 11.519 euros al mes y un valor medio de 206-207 euros al mes. Además, la variable no presenta valores nulos. Por último, mencionar que la variable presenta una distribución sesgada hacia la derecha, con una gran cantidad de valores pequeños y pocos valores extremadamente grandes. Esto se refleja en su cola larga hacia la derecha, lo que indica la presencia de valores atípicos muy grandes.
Conclusiones en su uso en el modelo
La variable es valiosa para identificar qué tipo de venue son más propensas a hacer churn. Sin embargo, hemos decidido descartar esta variable y utilizarla para crear una nueva variable que nos pueda aportar más información.

- Net_Revenue_Eur: Se trata de la diferencia entre las variables “gross_revenue_eur” y “cancelled_revenue_eur”. Nos aportará la ganancia neta en euros que ha tenido la venue por las orders del respectivo mes.

● Gross_Take_Rate

Significado de la variable
Representa la comisión ganada por Treatwell (en porcentaje) durante el respectivo mes-año.
Tipo de Dato: Cuantitativa Continua.
Conclusiones del análisis realizado
Se trata de una variable relativamente dispersa con un valor mínimo de 0, un valor máximo de 6,25 y un valor medio de 1. Además, la variable presenta un 1.24% de valores nulos. Por último, mencionar que la variable presenta una distribución sesgada hacia la derecha, con una mayor cantidad de valores pequeños y pocos valores extremadamente grandes. Esto se refleja en su cola larga hacia la derecha, lo que indica la presencia de valores atípicos grandes.
Conclusiones en su uso en el modelo
No vamos a utilizar esta variable ya que se trata de una transformación de Treatwell que podemos calcular nosotros mismos a partir de las variables “net_revenue_eur” y “net_ttv”.

● Cancelled_Item_Eur_Amount_Ttv

Significado de la variable
Representa el valor total en euros de las transacciones canceladas de las órdenes de un mes-año de cada venue. TTV significa “Total Transaction Value”.
Tipo de Dato: Cuantitativa Continua.

Conclusiones del análisis realizado
La variable representa un conjunto de datos con 1.764.164 observaciones. La media de los valores es de aproximadamente 135.307, con una desviación estándar de 271.0465. El rango de valores va desde 0 hasta 15.578.5, y los percentiles revelan que el 25% de los valores son menores o iguales a 35, el 50% son menores o iguales a 65, y el 75% son menores o iguales a 132. Además, el 65% de los valores son nulos y se han reemplazado con ceros ya que indica que la venue en ese mes-año no tuve órdenes canceladas.
Conclusiones en su uso en el modelo
La variable es valiosa para identificar qué tipo de venue son más propensas a hacer churn. Sin embargo, hemos decidido descartar esta variable y utilizarla para crear una nueva variable que nos pueda aportar más información.

- Net_TTV: Se trata de la diferencia entre las variables “gross_item_eur_amount_ttv” y “cancelled_item_eur_amount_ttv”. Nos aportará la cantidad neta en euros de las transacciones de las órdenes de un mes de cada venue.

● Cancelled_Aov

Significado de la variable
Representa el valor promedio de las órdenes canceladas de cada venue por mes. AOV significa “Average Order Value”.
Tipo de Dato: Cuantitativa Continua.
Conclusiones del análisis realizado
La variable representa un conjunto de datos que contiene 1.764.164 observaciones. La media de los valores es de aproximadamente 49.01946, con una desviación estándar de 39.81634. El valor mínimo es 0 y el valor máximo es 11.100. Los percentiles revelan que el 25% de los valores son menores o iguales a 28, el 50% son menores o iguales a 40, y el 75% son menores o iguales a 59. Estos datos indican que la mayoría de los valores se encuentran en un rango relativamente bajo, con una dispersión moderada. Además, el 65% de los valores son nulos y se han reemplazado con ceros ya que indica que la venue en ese mes-año no tuve órdenes canceladas.
Conclusiones en su uso en el modelo
No vamos a utilizar esta variable ya que se trata de una transformación de Treatwell que podemos calcular nosotros mismos a partir de las variables “net_ttv” y “net_orders”. La utilizaremos para crear una nueva variable que nos pueda aportar más información.

- Net_AOV: Se trata de la división entre las variables “net_ttv” y “net_orders”. Nos aportará el valor promedio neta en euros de las órdenes de un mes-año de cada venue.

● Cancelled_Revenue_Eur

Significado de la variable
Representa el ingreso total cancelado en euros que ha tenido la venue por las órdenes del respectivo mes-año.
Tipo de Dato: Cuantitativa Continua.
Conclusiones del análisis realizado
La variable representa un conjunto de datos que contiene 1.764.164 observaciones. La media de los valores es aproximadamente 14.87785, con una desviación estándar de 38.24708. El valor mínimo es 0 y el valor máximo es 3.885. Los percentiles muestran que el

25% de los valores son iguales o menores a 0, el 50% son iguales o menores a 0.96, y el 75% son iguales o menores a 16. Esto indica que la mayoría de los valores se encuentran en el rango inferior, con una dispersión relativamente amplia. Además, el 65% de los valores son nulos y se han reemplazado con ceros ya que indica que la venue en ese mes-año no tuve órdenes canceladas.

Conclusiones en su uso en el modelo

La variable es valiosa para identificar qué tipo de venue son más propensas a hacer churn. Sin embargo, hemos decidido descartar esta variable y utilizarla para crear una nueva variable que nos pueda aportar más información.

- Net_Revenue_Eur: Se trata de la diferencia entre las variables “gross_revenue_eur” y “cancelled_revenue_eur”. Nos aportará la ganancia neta en euros que ha tenido la venue por las orders del respectivo mes.

● Cancelled_Take_Rate

Significado de la variable

Representa la comisión cancelada por Treatwell (en porcentaje) durante el respectivo mes-año.

Tipo de Dato: Cuantitativa Continua.

Conclusiones del análisis realizado

La variable representa un conjunto de datos que contiene 1.763.025 observaciones. La media de los valores es de aproximadamente 0.1358152, con una desviación estándar de 0.1600124. El valor mínimo es 0 y el valor máximo es 0.5. Los percentiles muestran que el 25% de los valores son iguales o menores a 0, el 50% son iguales o menores a 0.02, y el 75% son iguales o menores a 0.3499935. Estos datos sugieren que la mayoría de los valores se concentran en el rango inferior, con una dispersión relativamente baja. Además, el 65% de los valores son nulos y se han reemplazado con ceros ya que indica que la venue en ese mes-año no tuve órdenes canceladas.

Conclusiones en su uso en el modelo

No vamos a utilizar esta variable ya que se trata de una transformación de Treatwell que podemos calcular nosotros mismos a partir de las variables “net_revenue_eur” y “net_ttv”.

● payment_method_name

Significado de la variable

La variable "payment_method_namet" en Treatwell es una variable dicotómica con dos posibles valores, que se refiere a las categorías que representan si el pago de los clientes se hizo de manera online “Pay_online” o en el establecimiento, “Pay_at_venue

Conclusiones del análisis realizado:

La proporción de establecimientos con Pay_online es del 58.09% mientras que el porcentaje de venues con Pay_at_venue es del 41.91%

Conclusiones en su uso en el modelo

Estas categorías del "payment_method_name" permiten clasificar y comprender mejor el comportamiento y las preferencias de los clientes que reservan en línea en Treatwell. Esta información puede ser utilizada para adaptar las estrategias de marketing y retención de venues de Treatwell, y por lo tanto, va a ser utilizada en el modelo

- Online_order_segment

Significado de la variable

La variable "online_order_segment" en Treatwell es una variable categórica ordinal que se refiere a las categorías que representan el tipo de cliente que realiza reservas en línea en los salones de Treatwell. A continuación, se explican estas categorías:

- Explorer: se refiere a clientes que están explorando nuevos lugares y cambian de salón en cada reserva (cliente que genera ingresos a Treatwell).
- Loyal: se refiere a clientes que son leales y repiten sus reservas en el mismo salón (cliente que no genera ingresos a Treatwell).
- Newbie: se refiere a clientes nuevos que comienzan a utilizar los servicios de Treatwell. Estos clientes son nuevos en la plataforma y están realizando sus primeras reservas en salones (cliente que genera ingresos a Treatwell).
- Referral: se refiere a clientes que llegan a través de referencias de otros clientes. Estos clientes fueron recomendados por alguien más que ya ha utilizado los servicios de Treatwell (cliente que genera ingresos a Treatwell).

Tipo de Dato: Cualitativa Ordinal.

Conclusiones del análisis realizado

Los salones con valores más altos de “loyal” y “referral” tienen menos probabilidad de churn, mientras que aquellos con valores más bajos en estas categorías tienen una mayor probabilidad de churn.

Conclusiones en su uso en el modelo

Estas categorías del "online_order_segment" permiten clasificar y comprender mejor el comportamiento y las preferencias de los clientes que reservan en línea en Treatwell. Esta información puede ser utilizada para adaptar las estrategias de marketing y retención de venues de Treatwell, y por lo tanto, va a ser utilizada en el modelo .

12.3. Variables generadas

Se proponen nuevas variables como mejores alternativas para lograr resultados con alto nivel de precisión.

- Antigüedad

Creación de la variable

Representa el número de meses desde latest_live_date hasta la fecha en que la data fue obtenida (date_day). Es decir, la antigüedad del venue.

Tipo de Dato: Cuantitativa Discreta.

Conclusiones del análisis realizado

Se trata de una variable que presenta una media de 82 meses con una alta desviacion estándar (48 meses). Su valor mínimo es 0 para venues que recién partieron su contrato con la empresa y el valor máximo observado en la data es de 183 meses.

Conclusiones en su uso en el modelo

Con esta variable podremos concluir si tener una relación de larga duración con Treatwell garantiza una menor probabilidad de churn o no.

- ## Reactivada

Creación de la variable
Representa si la venue se ha ido y ha vuelto en mas de 30 días, es decir, la diferencia en días de la variable latest_live_date y first_live_date es mayor a 30 días. Esto nace dado que existen variaciones de días (menos de un mes) que podrían presentar falsos positivos.
Tipo de Dato: Cualitativa Dicotomica.
Conclusiones del análisis realizado
Se trata de una variable que presenta 3130848 records con valor 0, es decir el 93% de la data tratada, mientras que 222011 records tienen valor 1 que representa aproximadamente el 6%. Es importante considerar que existen récords sin el valor latest_live_date, en estos casos se utilizo first_live_date, por lo que esta variable tiene valor 0.
Conclusiones en su uso en el modelo
Con esta variable podremos concluir si volver luego de un término de contrato con Treatwell genera mayor compromiso y, por tanto, una menor probabilidad de churn.

- ## Appointments_l30d

Significado de la variable
Representa la cantidad total de “appointments” que ha tenido cada venue durante los últimos 30 días. Hemos creado esta variable a partir de la suma entre las variables “direct_appointments_l30d” y “online_appointments_l30d”.
Tipo de Dato: Cuantitativa Discreta.
Conclusiones del análisis realizado
Se trata de una variable muy dispersa con un valor mínimo de 0 “appointments”, un valor máximo de 21.507 “appointments” y un valor medio de 71-72 “appointments” por venue. Además, la variable presenta un moderado porcentaje de valores nulos (22%). Por último, mencionar que la variable presenta una distribución sesgada hacia la derecha, con una gran cantidad de valores pequeños y pocos valores extremadamente grandes. Esto se refleja en su cola larga hacia la derecha, lo que indica la presencia de valores atípicos muy grandes.
Conclusiones en su uso en el modelo
La variable es valiosa para identificar qué tipo de venue son más propensas a hacer churn. Según nuestra hipótesis, las venues con menor cantidad de “appointments” deberían ser más propensas a hacer churn.

- ## Online_Appointments_Rate

Significado de la variable
Representa el ratio de “appointments” que vienen por parte de los clientes a través del Saas durante los últimos 30 días. Hemos creado esta variable a partir de la división de las variables “online_appointments_l30d” y “appointments_l30d”.
Tipo de Dato: Cuantitativa Continua.
Conclusiones del análisis realizado
Se trata de una variable muy dispersa con un valor mínimo de 0% de “appointments online”, un valor máximo de 100% de “appointments online” y un valor medio de 9%-10% de “appointments online” por venue. Además, la variable presenta un moderado porcentaje de valores nulos (22%). Por último, mencionar que la variable presenta una distribución sesgada hacia la derecha, con una gran cantidad de valores pequeños y pocos valores

extremadamente grandes. Esto se refleja en su cola larga hacia la derecha, lo que indica la presencia de valores atípicos muy grandes.
Conclusiones en su uso en el modelo
La variable es valiosa para identificar qué tipo de venue son más propensas a hacer churn. Según nuestra hipótesis, las venues con menor cantidad de “appointments online” deberían ser más propensas a hacer churn.

- Net_Orders

Significado de la variable
Representa la cantidad limpia de órdenes de una venue en un mes-año. Hemos creado esta variable a partir de la diferencia entre las variables “gross_orders” y “cancelled_orders”.
Tipo de Dato: Cuantitativa Discreta.
Conclusiones del análisis realizado
Se trata de una variable muy dispersa con un valor mínimo de 0 orders al mes, un valor máximo de 3.021 orders al mes y un valor medio de 45 orders al mes. Además, la variable no presenta valores nulos. Por último, mencionar que la variable presenta una distribución sesgada hacia la derecha, con una gran cantidad de valores pequeños y pocos valores extremadamente grandes. Esto se refleja en su cola larga hacia la derecha, lo que indica la presencia de valores atípicos muy grandes.
Conclusiones en su uso en el modelo
La variable es valiosa para identificar qué tipo de venue son más propensas a hacer churn. Según nuestra hipótesis, las venues con menor cantidad de “net orders” deberían ser más propensas a hacer churn.

- Orders_Cancellation_Rate

Significado de la variable
Representa el ratio de órdenes canceladas por venue en un mes-año. Hemos creado esta variable a partir de la división de las variables “cancelled_orders” y “gross_orders”.
Tipo de Dato: Cuantitativa Continua.
Conclusiones del análisis realizado
Se trata de una variable muy dispersa con un valor mínimo del 0% de cancelaciones, un valor máximo del 100% de cancelaciones y un valor medio del 20% de cancelaciones. Además, la variable no presenta valores nulos. Por último, mencionar que la variable presenta una distribución un poco sesgada hacia la derecha, con una mayor cantidad de valores pequeños y pocos valores grandes.
Conclusiones en su uso en el modelo
La variable es valiosa para identificar qué tipo de venue son más propensas a hacer churn. Según nuestra hipótesis, las venues con mayor cantidad de orders canceladas deberían ser más propensas a hacer churn.

- Net_Aov

Significado de la variable
Representa el valor promedio de las órdenes de cada venue por mes (AOV = Average Order Value). Hemos creado esta variable a partir de la división de las variables “net_ttv” y “net_orders”.

Tipo de Dato: Cuantitativa Continua.
Conclusiones del análisis realizado
Se trata de una variable muy dispersa con un valor mínimo de 0 euros, un valor máximo de 2.214 euros y un valor medio de 44 euros. Además, la variable no presenta valores nulos. Por último, mencionar que la variable presenta una distribución sesgada hacia la derecha, con una gran cantidad de valores pequeños y pocos valores extremadamente grandes. Esto se refleja en su cola larga hacia la derecha, lo que indica la presencia de valores atípicos muy grandes.
Conclusiones en su uso en el modelo
La variable es valiosa para identificar qué tipo de venue son más propensas a hacer churn. Según nuestra hipótesis, las venues con menor “net_aov” deberían ser más propensas a hacer churn.

- Net_Take_Rate

Significado de la variable
Representa la comisión ganada por Treatwell en porcentaje durante el respectivo mes. Hemos creado esta variable a partir de la división de las variables “net_revenue_eur” y “net_ttv”.
Tipo de Dato: Cuantitativa Continua.
Conclusiones del análisis realizado
Se trata de una variable relativamente dispersa con un valor mínimo del 0%, un valor máximo del 1% y un valor medio del 11%-12%. Además, la variable no presenta valores nulos. Por último, mencionar que la variable presenta una distribución sesgada hacia la derecha, con una mayor cantidad de valores pequeños y pocos valores extremadamente grandes. Esto se refleja en su cola larga hacia la derecha, lo que indica la presencia de valores atípicos grandes.
Conclusiones en su uso en el modelo
La variable es valiosa para identificar qué tipo de venue son más propensas a hacer churn. Según nuestra hipótesis, las venues con mayor “net_take_rate” deberían ser más propensas a hacer churn.

- Is_purchasable_duration

Significado de la variable
Es una variable que tiene en cuenta la duración durante la cual un servicio con "is_purchasable=True" estuvo activo. La variable no solo tiene en cuenta si el servicio está a la venta , sino también cuánto tiempo estuvo disponible para su compra.
Tipo de Dato: Cuantitativa Discreta.
Conclusiones del análisis realizado:
La proporción de establecimientos con is_purchasable = True es del 26.06%. El coeficiente de correlación punto biserial con churn de is_purchasable_duration es -0,194, lo indica una correlación punto biserial negativa débil, pero sin embargo, significativa, que nos indica que aumenta la duración de is_purchsable_duration, disminuye la probabilidad de abandono.
Conclusiones en su uso en el modelo
Es una variable relevante para incorporar en el modelo, principalmente, porque nos aporta información sobre la duración, en meses, en la que la venue tuvo habilitada la opción de

compra de al menos un servicio en el marketplace y tiene una relación significativa con churn

- ## Is_purchasable_duration_bin

Es una variable categórica que tiene en cuenta tres rangos en cuanto a la duración del servicio "is_purchasable". Existen las siguientes categorías:

- Not Enabled cuando el servicio no está habilitado
- is_purchasable_less_than_one_year menos de un año
- is_purchasable_more_than_one year cuando está activo por más de un año.

La variable no solo tiene en cuenta si el servicio está a la venta , sino también cuánto tiempo estuvo disponible para su compra.
Tipo de Dato: Caulitativa Ordinal.
Conclusiones del análisis realizado:
La proporción de establecimientos con is_purchasable = Not enabled es del 73.94%.
Conclusiones en su uso en el modelo
Es una variable relevante para incorporar en el modelo, principalmente, porque nos aporta información sobre la duración, en meses, en la que la venue tuvo habilitada la opción de compra de al menos un servicio en el marketplace y tiene una relación significativa con churn. De esta manera, se concluye que las categoría “purchasable_not_enabled” y “purchasable_more_than_one_yearr” son las que tienen una mayor capacidad predictiva con churn.

- ## Is_widget_duration_bin

Significado de la variable
Variable categórica que tiene en cuenta tres rangos en cuanto a la duración del servicio “is_widget_enabled”. Existen las siguientes tres categorías:

- Not enabled, cuando el servicio no está habilitado
- widget_less_than_one_year, habilitado por menos de un año
- widget_more_than_one year habilitado por más de un año

La variable no solo tiene en cuenta si el servicio está a la venta , sino también cuánto tiempo estuvo disponible para su compra.
Tipo de Dato: Caulitativa Ordinal.
Conclusiones del análisis realizado:
La proporción de establecimientos con is_widget_duration_bin= Not Enabled es del 73.16%.
Conclusiones en su uso en el modelo
Es una variable relevante para incorporar en el modelo, principalmente, porque nos aporta información sobre la duración, en meses, en la que la venue tuvo habilitada el widgets. De esta manera, se concluye que la categoría “widget_not_enabled” y “widget_more_than_one_yearr” son las que tienen una mayor capacidad predictiva con churn.

- ## Churn

Significado de la variable
La variable "churn" es nuestra variable a predecir. Es una variable dicotómica con valores posibles “1” = Churn y 0 = “Not Churned. Se obtiene al calcular la diferencia entre la fecha

de la última actividad de una venue y la fecha en que se dio de baja. Si la diferencia es mayor a tres meses, se considera que el salón ha churned y se asigna el valor 1 a la variable "churn". De lo contrario, se asigna el valor 0, indicando que el salón sigue activo.
Tipo de Dato: Cualitativa Dicotómica.
Conclusiones del análisis realizado:
Las variables explicativas incluidas en el modelo tienen diferentes tipos de relaciones con la variable independiente, churn, lo cual nos proporciona una gran ventaja para predecir y comprender los factores que influyen en la cancelación de salones en Treatwell. Estas relaciones nos permiten identificar patrones y tendencias que pueden ayudarnos a tomar decisiones informadas para retener a los salones existentes y mejorar la satisfacción del cliente.
Conclusiones de su uso en el modelo:
El modelo de predicción de churn en Treatwell se traduce en la capacidad de retener salones existentes, mejorar la satisfacción del salón y optimizar las estrategias de captación de clientes nuevos con el análisis proporcionado, con tal de maximizar la rentabilidad de la empresa. Al implementar y utilizar este modelo de manera efectiva, Treatwell tiene una ventaja competitiva para entender qué factores son los más importantes y de qué manera para que un cliente abandone el servicio Treatwell, con tal de aplicar estrategias para revertir en lo máximo posible que el cliente sea “churn”.

13. Modelo

Se procede a presentar el trabajo realizado en el modelo. Cabe destacar que el trabajo realizado ha sido con fines comparativos y que ayudará en el desarrollo de futuros modelos.

13.1. Métricas a trabajar

Al considerar los requerimientos y el posterior uso del modelo propuesto, se ha planteado trabajar en la optimización de dos métricas principales:

Recall: Esta métrica evalúa la proporción de casos positivos que han sido correctamente identificados dentro del conjunto de casos positivos en los datos. Esta métrica surge debido a un desequilibrio esperado en los datos, y su objetivo principal es maximizar la detección o identificación de casos positivos, es decir, establecimientos que tienen una alta probabilidad de churn, incluso a expensas de tener un mayor número de falsos positivos. Con esta métrica, buscamos optimizar los hiperparámetros de los modelos propuestos.

Accuracy_top_prob: Esta métrica es una propuesta del equipo al considerar que Treatwell seleccionará las 100 venues con la mayor probabilidad de churn. El modelo proporcionará 100 observaciones con la mayor probabilidad de churn, las cuales serán contrastadas con la realidad para obtener un porcentaje de aciertos. Podremos determinar que dentro de los 100 registros más probables, un X% realmente hicieron churn.

13.2. Desarrollo del modelo

Como se expuso en un principio, el primer modelo a trabajar con fines comparativos es **Random Forest**.

Consideraciones:

- **Robustez:** Random Forest es robusto y puede manejar datos con ruido y valores atípicos sin sobreajustarse a los datos de entrenamiento.
- **Precisión:** Random Forest es altamente preciso y puede manejar tanto problemas de clasificación como de regresión con variables categóricas y continuas.
- **Velocidad:** A pesar de ser un algoritmo complejo, Random Forest es rápido y puede trabajar eficientemente con conjuntos de datos grandes. Además, se puede paralelizar fácilmente para acelerar el entrenamiento.
- **Importancia de características:** Random Forest proporciona una medida de la importancia de las variables, lo que ayuda en la selección de características y en la comprensión de los datos.
- **Independencia de los árboles:** Los árboles de decisión en un Random Forest se construyen de forma independiente, lo que permite capturar diferentes aspectos y relaciones en los datos, incluso en variables correlacionadas.
- **Robustez ante outliers:** Random Forest es menos sensible a los outliers debido a su estructura basada en múltiples árboles. Los outliers solo afectan a los árboles específicos en los que se encuentran, pero el modelo en conjunto puede compensarlos.
- **No se requiere escalado entre 0 y 1:** A diferencia de algunos modelos, Random Forest no necesita que los datos están escalados en un rango específico, ya que los árboles de decisión se basan en umbrales y reglas de división.
- **No se requieren relaciones lineales:** Random Forest puede capturar relaciones no lineales y complejas, no asume relaciones lineales entre las variables predictoras y la variable objetivo.

Antes de crear el modelo, hemos **ordenado los datos** por fecha y venue (“date_day” y “unique_venue_id”). También hemos **creado la variable “month**” ya que durante el análisis exploratorio de datos (EDA) hemos observado que existe **estacionalidad**. Además, hemos dividido los datos en dos conjuntos (**train y test**) utilizando la variable “date_day” debido a la estacionalidad presente en los datos. Por último, hemos creado los **índices “date_day” y “unique_venue_id”** para mantener el orden en ambos conjuntos.

Hemos **creado y entrenado el modelo** utilizando los datos del conjunto train. Además, también hemos creado dos **DataFrames** distintos tanto para el conjunto de train como de test:

- **Probs:** Contiene las probabilidades de que el registro sea churn (0-1).
- **Preds:** Contiene la predicción de que el registro sea churn (0 / 1).

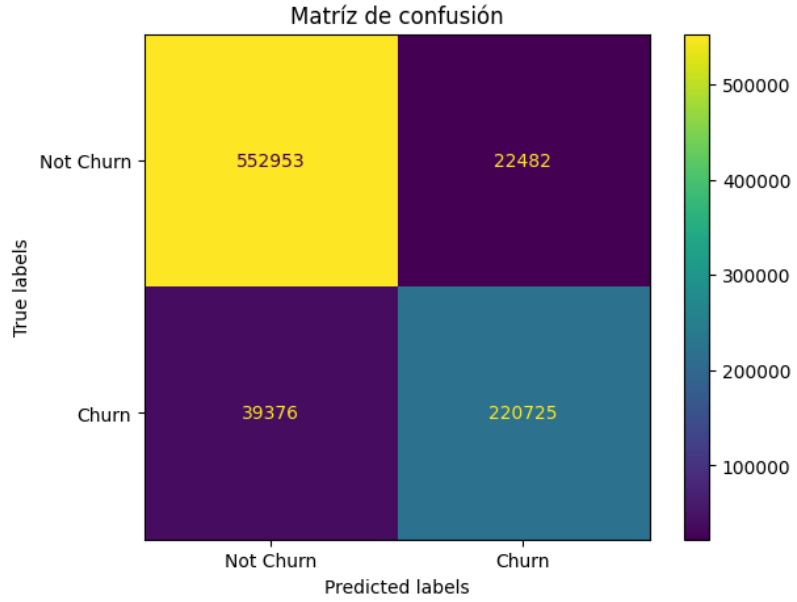
Hemos creado una **métrica propia llamada “`accuracy_top_prob`”** que mide el porcentaje de aciertos que hemos obtenido en los 100 registros con una **probabilidad de hacer churn más alta**. Hemos aplicado este métrica al conjunto train y test y hemos obtenido los siguientes resultados:

- **Test:** El porcentaje de aciertos en los 100 registros que tienen una probabilidad de hacer churn más alta es del 100%.

- **Train:** El porcentaje de aciertos en los registros que tienen una probabilidad de hacer churn más alta es del 99%.

Como se puede observar, el conjunto de entrenamiento ha mostrado un rendimiento superior al conjunto de prueba. Esto es esperado, ya que el modelo ha sido entrenado con los datos del conjunto de entrenamiento y está más familiarizado con ellos.

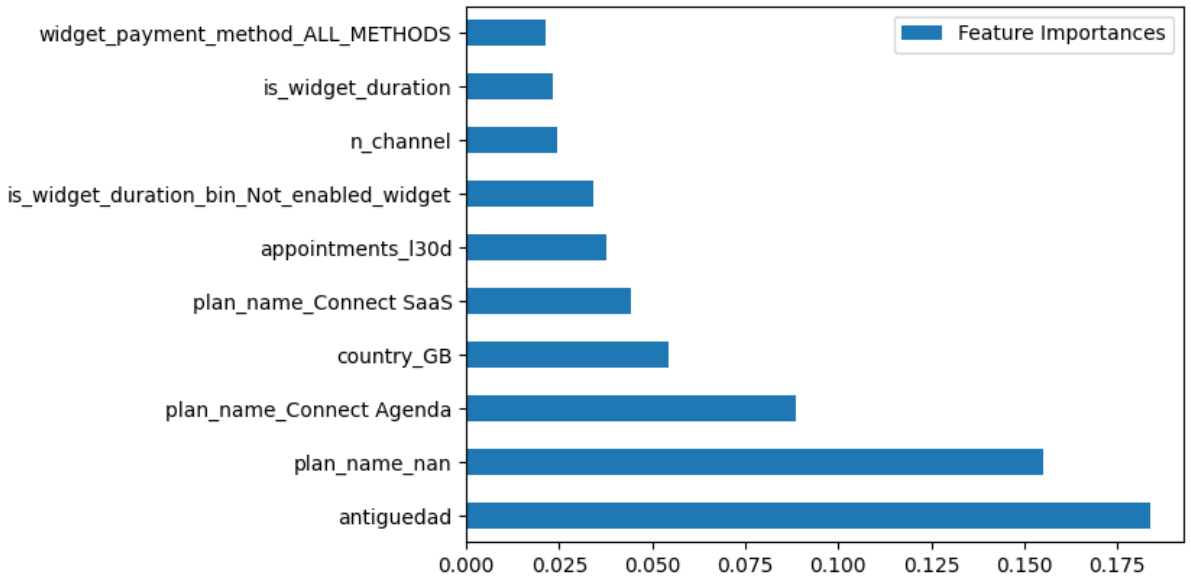
Hemos creado la **matríz de confusión** y el **classification report** de las predicciones en el **conjunto de test**. En este sentido, hemos obtenido una **alta precisión para ambas clases**, lo que indica que el modelo tiene una buena capacidad para predecir correctamente los casos positivos y negativos en ambas clases. No obstante, el **recall es ligeramente más bajo** para la clase "churn", lo que indica que el modelo puede tener más dificultades para identificar correctamente los casos de churn. En resumen, aunque el modelo muestra buenos resultados en términos de precisión y recall para ambas clases, es importante considerar que el recall para la clase minoritaria puede ser de mayor relevancia en un problema desbalanceado de churn.




```
print(classification_report(y_test, prediction_churn_random_forest))
```

	precision	recall	f1-score	support
0.0	0.93	0.96	0.95	575435
1.0	0.91	0.85	0.88	260101
accuracy			0.93	835536
macro avg	0.92	0.90	0.91	835536
weighted avg	0.93	0.93	0.93	835536

Hemos creado un gráfico para visualizar la **importancia que tiene cada variable** dentro del modelo. En este gráfico, hemos visto que las variables con mayor importancia para el modelo son: “antiguedad”, “plan_name_nan”, “plan_name_connect_agenda”, “country_GB”, “plan_name_connect_saas” y “appointments_l30d”.



Hemos realizado una **busqueda en grilla** para intentar encontrar los mejores **hiperparámetros** para maximizar el **recall** del modelo. En este sentido, hemos obtenido que los mejores parámetros son los siguientes:

- max_depth: 5
- min_samples_leaf: 10
- min_samples_split: 5
- n_estimators: 6

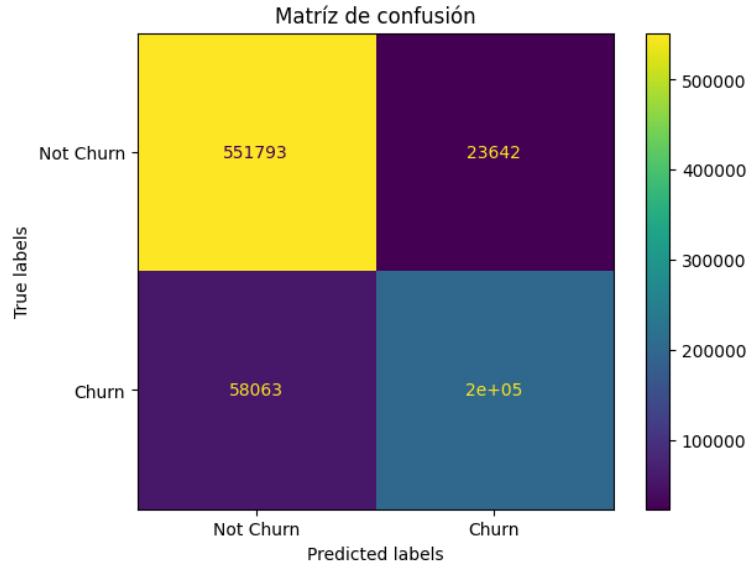
Hemos creado un nuevo modelo con las siguientes variaciones:

- **Hiperparámetros:** Hemos añadido los nuevos hiperparámetros encontrados para maximizar el recall.
- **Variable “plan_name”:** A pesar de que la empresa solicitó la inclusión de la variable "plan_name" en el modelo, hemos tomado la decisión de eliminarla debido a la alta cantidad de valores nulos que presenta. Además, durante la ejecución del modelo, hemos observado que otorga una gran importancia a los valores nulos de esta variable, lo cual resta relevancia y sentido de negocio al modelo. En consecuencia, uno de los próximos pasos consistirá en desarrollar otro modelo que pueda incluir esta variable, eliminando sus valores nulos. De esta manera, obtendremos dos modelos en paralelo con el objetivo de obtener las dos respuestas de negocio requeridas.
- **Variable “country”:** Hemos decidido eliminar la variable "country" luego de reconocer que no aporta valor como predictor en el conjunto de datos. Resulta más beneficioso aplicar el modelo de churn general, en lugar de considerar "country" como una variable predictora en sí misma. Considerar esta variable como predictora habría sido viable si todos los países hubieran mostrado una cantidad equilibrada de venues y comportamientos similares. Como se presentan valores desbalanceados para países, la variable "country_GB" adquiere mayor relevancia en el modelo, debido a que contiene un número superior de registros.

Hemos evaluado el modelo utilizando nuestra métrica propia (**accuracy_top_prob**):

- **Test:** El porcentaje de aciertos en los 100 registros que tienen una probabilidad de hacer churn más alta es del 100%.
- **Train:** El porcentaje de aciertos en los registros que tienen una probabilidad de hacer churn más alta es del 91%.

Hemos generado la **matriz de confusión** y el **classification report** de las predicciones en el **conjunto de test**. En este análisis, hemos observado que la precisión y el recall obtenidos son **inferiores** a los del primer modelo. La razón de esto radica en la eliminación de dos variables que poseían una alta importancia para un **modelo desbalanceado**. Por lo tanto, uno de los pasos siguientes consistirá en intentar optimizar los hiperparámetros de este segundo modelo con el fin de mejorar el recall.



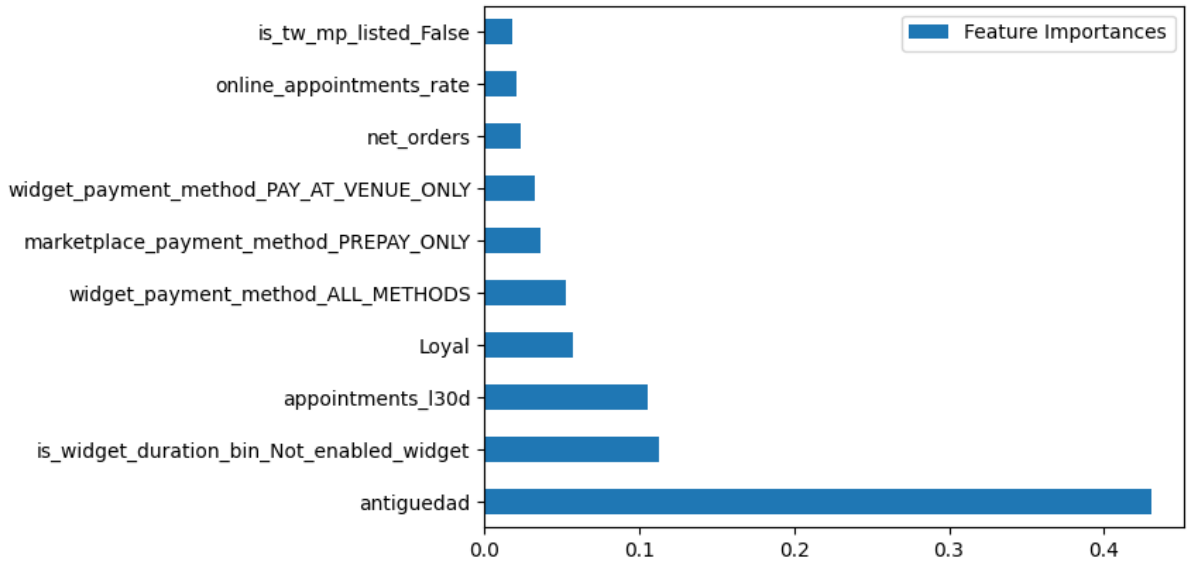

```
print(classification_report(y_test, prediction_churn_random_forest))
```

```
              precision    recall  f1-score   support

         0.0       0.90      0.96      0.93    575435
         1.0       0.90      0.78      0.83    260101

    accuracy                           0.90    835536
   macro avg       0.90      0.87      0.88    835536
weighted avg       0.90      0.90      0.90    835536
```

En último lugar hemos creado un gráfico para visualizar la **importancia que tiene cada variable** dentro del modelo. En este gráfico, hemos visto que las variables con mayor importancia para el modelo son: “antiguedad”, “is_widget_duration_bin_not_enabled”, “appointments_l30d”, “loyal” y “widget_payment_method_all_methods.



Resumen

En resumen, se ha trabajado con el modelo **Random Forest** con el objetivo de realizar **predicciones de churn**.

El modelo se ha entrenado con los datos de entrenamiento y se ha evaluado utilizando la métrica **Recall** y la metrica propia llamada **"accuracy_top_prob"**, que mide el porcentaje de aciertos en los registros con la probabilidad más alta de churn. Además se han analizado la matriz de confusión y el classification report en el conjunto de prueba, observando una alta precisión para ambas clases y un recall ligeramente más bajo para la clase "Churn".

A continuación se ha realizado una búsqueda en grilla para encontrar los mejores hiperparámetros y se ha creado un nuevo modelo con las variaciones mencionadas, eliminando la variable "plan_name" y la variable "country".

El nuevo modelo ha sido evaluado utilizando la métrica **Recall** y **"accuracy_top_prob"**, obteniendo un alto porcentaje de aciertos tanto en el conjunto de entrenamiento como en el de prueba. Sin embargo, al analizar la matriz de confusión y el classification report, se ha observado que la precisión y el recall son inferiores a los del primer modelo debido a la eliminación de variables importantes en un contexto de desbalance de clases.

Por este motivo, uno de los siguientes pasos consistirá en intentar optimizar los hiperparámetros de este segundo modelo con el fin de mejorar el recall.

Anexo

Github

A continuación, se comparte el link para acceder a todos los notebook trabajados por el equipo:
https://github.com/brunopedemonte/TFM-UB-Grupo-5.git

Apéndice 1

```
mapping = {'nan': 'Beauty Salon',
        'Hair Salon':'Hair Salon',
        'Beauty Salon':'Beauty Salon',
        'Massage & Therapy Centre':'Massage Salon',
        'Fitness Centre':'Body Salon',
        'Nail Salon':'Nail Salon',
        'Wellness Centre':'Beauty Salon',
        'Treatment Room - Beauty':'Beauty Salon',
        'Day Spa':'Spa Salon',
        'Skin Clinic':'Skin Clinic',
        'Mobile Beauty':'Beauty Salon',
        'Home-based Venue':'Beauty Salon',
        'Medical Spa':'Spa Salon',
        'Tanning Salon':'Body Salon',
        'Dental Clinic':'Dental Clinic',
        'Barbershop':'Barbershop',
        'Hotel Spa':'Spa Salon',
        'Waxing Salon':'Hair Removal Salon',
        'Treatment Room - Wellness':'Massage Salon',
        'Mobile Massage':'Massage Salon',
        'Brow Bar':'Face Salon',
        'Makeup Studio':'Face Salon',
        'Weight Loss Clinic':'Body Salon',
        'Yoga Studio':'Body Salon',
        'Pilates Studio':'Body Salon',
        'Destination Spa':'Spa Salon',
        'Chiropractic Clinic':'Massage Salon',
        'Treatment Room - Spa':'Spa Salon',
        'Chiropody Clinic':'Massage Salon',
        'Hammam':'Spa Salon'
        }
```

Apéndice 2

```
mapping = {'Entry':'Entry',
        'Plus':'Plus',
        'Partner Saas':'Partner Saas',
        'Premium': 'Premium',
        'Starter': 'Starter',
        'Advanced': 'Advanced',
        'Legacy':'Legacy',
        'Réservation En Ligne':'Starter',
        'Gr Funkmartini Basic':'Starter',
        'Gr Funkmartini Gold':'Advanced',
        'Vetrina + Agenda':'Advanced',
        'Gestion + Réservation En Ligne':'Premium',
        'Global':'Premium',
        'Trial':'Starter',
        'Réservation En Ligne + Site Internet':'Advanced',
        'Fr Uala Marketplace':'Advanced',
        'Gestion': 'Premium',
        'Site Internet Only ':'Starter',
        'Uk Churned Venue':'Churned',
        'Gestion + Site Internet ':'Premium',
        'Es Churned Venue':'Churned',
        'Fr Churned Venue':'Churned',
        'Costo Servizio Di Prenotazione':'Starter',
        'Gestionale':'Premium',
        'Es Free':'Starter',
        'De Churned Venue':'Churned',
        'Dead':'Churned',
        'It Churned Venue':'Churned',
        'Pt Basic':'Starter',
        'It Free':'Starter',
        'Saas Kadus':'Partner Saas',
        'Fr Uala Free':'Starter',
        'Comisiones':'Starter'
        }
```