UNIVERSIDAD TORCUATO DI TELLA

MiM + Analytics

Predicción de churn en restaurantes para aplicaciones de delivery

Ihde, Nicolas

Tutor: Ramiro Gálvez

Resumen

Para las aplicaciones de delivery de comida online es importante contar con una base de comercios activos en la aplicación que puedan brindar sus servicios a través de ella. Para poder garantizar una oferta variada y sana, relevante para los consumidores, es fundamental velar por la cantidad y la calidad de los comercios asociados.

En una industria de márgenes bajos, las comisiones y condiciones comerciales de las Apps ejercen mucha presión en los restaurantes. Estos en muchos casos finalmente deciden dejar de operar en la aplicación por problemas operativos y económicos.

La manera de medir esta problemática para las apps es a través de la tasa de puntos de venta que se dan de baja de la plataforma por mes (churn).

Es importante poder predecir el churn de restaurantes a futuro y las causas que lo generan, de esta manera se podrán tomar medidas a tiempo para evitar que los restaurantes salgan de la aplicación.

El propósito de este trabajo es desarrollar un modelo de predicción que pueda anticipar cuando un restaurante va a cerrar su perfil en la aplicación, usando datos tanto de restaurantes como de clientes para poder detectar tendencias y así poder predecir la atrición.

Para poder abordar el problema se utiliza una técnica de aprendizaje automático supervisado llamado Light GBM, desarrollado por Microsoft, para hacer las predicciones.

La herramienta logra predecir cuales son los restaurantes que van a dejar de operar en la plataforma con un área bajo la curva ROC del 0.92, lo que indica que tiene buen poder de predicción.

Finalmente, se discute la aplicación del modelo en la industria, y cómo puede ayudar a atacar el problema de manera proactiva para disminuir la cantidad de restaurantes que salen, y eficientizar la inversión en fidelización.

Abstract

For food delivery apps, it is important to have a base of active restaurants in their offering where they can provide their services. In order to guarantee a varied and healthy offer, relevant to consumers, it is essential for these companies to ensure the quantity of the associated businesses.

In a low-margin industry, the commissions and commercial conditions from the Apps put a lot of pressure on restaurants. Where in many cases they decide to stop operating in the platform due to operational and economic problems.

The way to measure this problem is through the rate of points of sale that are unsubscribed from the platform per month (churn).

It is important for these companies to be able to predict future churn and the causes that generate it, so that the correct actions can be made in time to prevent restaurants from leaving the application.

The purpose of this work is to develop a prediction model that can anticipate when a restaurant is going to close its profile in the application.

In this model, data from both restaurants and customers will be used to detect trends and thus be able to predict attrition.

In order to address the problem, a supervised machine learning technique called Light GBM, developed by Microsoft, will be used to make the predictions.

The tool manages to predict which restaurants are going to stop operating on the platform with an area under the ROC curve of 0.92, which indicates that it has good predictive power.

Finally, the application of the model in the industry is discussed, and how it can help to proactively attack the problem to reduce the number of restaurants that leave, and make the investment in loyalty more efficient.

Índice

1 - Introducción:

1.1 - Dominio:

El comercio minorista y la industria de alimentos y bebidas fueron de las industrias más afectadas durante la pandemia de COVID. Las restricciones de los gobiernos a nivel mundial sobre la circulación significaron que los locales minoristas y los restaurantes tuvieron que explorar otras vías para extender la experiencia a los hogares de los clientes. Eso significaba ofrecer el mismo tipo de experiencia de marca y nivel de servicio, pero con la transición a los canales en línea (Shah,A.,2021).

El mercado necesitaba adaptarse y tomar decisiones rápidas a las restricciones en constante cambio para continuar operando con éxito. Incluso después de la pandemia, la entrega de pedidos y las ofertas digitales son imprescindibles para las empresas minoristas y de alimentos. El factor clave detrás de los competidores que continuaron operando durante la pandemia fue su capacidad para atender a los clientes sin tener ningún paso físico en sus puntos de venta.

En este contexto las aplicaciones de entrega de productos de conveniencia y alimentos se han vuelto cada vez más populares. Es una versión del e-commerce donde la promesa de entrega es casi inmediata. Está compuesta por empresas de tecnología con plataformas que conectan a proveedores con sus consumidores, y donde en muchos casos también se encargan de la logística gestionando una flota de repartidores.

Generalmente se centran en productos de conveniencia o consumo inmediato donde el tiempo para tenerlo disponible es un factor primordial a la hora de hacer la compra por parte del consumidor.

Es una industria con crecimientos exponenciales desde hace más de una década, siendo cada vez más común en los hábitos de consumo de la población. Se espera que el tamaño del mercado de Comercio Rápido se transforme en una industria de alrededor de $72 mil millones de dólares a nivel mundial para 2025 (Bommireddipalli,R., 2022).

La industria funciona como un nexo entre sus dos stakeholders fundamentales (consumidores y vendedores), al integrarlos para generar una excelente experiencia.

Los clientes quieren recibir sus pedidos rápidamente, a bajo costo y sin ninguna reducción en la calidad donde se busca tener una experiencia de marca similar a la que conocen en otros formatos. Esperan encontrar exactamente lo que buscan y recibirlo de la manera que lo deseen de manera conveniente, con poco esfuerzo. Los vendedores buscan una fuente de ingresos adicional mejorando la capacidad operativa de sus cocinas, apalancado en una mayor visibilidad a través de internet para llegar a más clientes (Hirschberg, C., Rajko, A., Schumacher, T., & Wrulich, M., 2016).

Beneficios de las aplicaciones de delivery (Gupta, M., 2019):

- Fácil de usar
- Pagos flexibles
- Seguimiento de las órdenes en tiempo real
- Eficaz atención al cliente

Para este tipo de aplicaciones es importante contar con una base de comercios activos en la plataforma que puedan brindar sus servicios a través de ella. En Latinoamérica se cuenta con más de 70.000 negocios adheridos (Moreno Padilla, G. A. 2021), creciendo constantemente todos los meses. Para poder garantizar una oferta variada y sana, relevante para los consumidores, es fundamental para las aplicaciones velar por la cantidad y la calidad de los comercios asociados. Con equipos especializados tanto para captar nuevos locales como para medir la performance de los actuales, las apps de delivery llevan un arduo control de la oferta presente y velan por su salud.

1.2 - Problema:

En una industria de márgenes bajos, las comisiones y condiciones comerciales de las apps ejercen mucha presión en los restaurantes (Feldman, P., Frazelle, A. E., & Swinney, R. 2019). En muchos casos finalmente deciden dejar de operar en la aplicación por problemas operativos y económicos.

Para los restaurantes, las órdenes que vienen a través de aplicaciones de delivery tienen márgenes muy bajos al incluir las comisiones que se pagan (entre 15% y 30% del valor del pedido). Por otro lado, las aplicaciones ejercen presión para hacer importantes descuentos en los precios, que son absorbidos por los restaurantes. Esto hace que a pesar de las órdenes incrementales, no siempre sea rentable para el restaurantes operar en estas aplicaciones (Dunn, E. 2018).

Es imprescindible para la industria contar con una cartera de restaurantes operativos para dar un buen servicio a sus usuarios. Es por eso que se desarrollan distintas estrategias de fidelización de restaurantes y descuentos comerciales en los casos que sean necesario para asegurar que el número de bajas no crezca. La manera de medir esta problemática para las apps es a través de la tasa de puntos de venta que se dan de baja de la plataforma por mes (churn).

Es importante para el negocio poder predecir el churn a futuro y las causas que lo generan. De esta manera se podrán tomar medidas a tiempo para evitar que los restaurantes salgan de la aplicación.
Al tratarse de una industria relativamente nueva y en crecimiento, no hay una gran cantidad de bibliografía específica sobre las aplicaciones de delivery de comida. Dentro de la documentación disponible si hay material sobre el churn de consumidores. Es decir, hay análisis hechos para entender las razones y los costos detrás de la salida de consumidores de la app (de la Llave Montiel, M. A., & López, F. 2020), como así también hay herramientas implementadas en las principales compañías para predecir el churn de consumidores para tratar de prevenirlo.
En cuanto al churn de restaurantes, no se encontró bibliografía sobre este tema para esta industria o similares, y en la práctica no es un problema relevante para el negocio hasta el momento.

1.3 - Objetivo:

La meta de este trabajo es desarrollar un modelo de predicción que pueda anticipar que un restaurante va a cerrar su perfil en la aplicación. Con esta información se podrán recomendar distintas acciones comerciales y campañas de fidelización en base al modelo para evitar el churn a futuro. De este modo se podrá ayudar a la empresa a hacer un uso más eficiente del gasto en retención de restaurantes.

Una vez entrenado el modelo para cada restaurante, resultará de vital importancia encontrar un umbral de decisión adecuado para hacer las distintas acciones comerciales con el objetivo de aumentar la retención y optimizar el gasto de marketing.

En la actualidad la organización con cuyos datos vamos a trabajar cuenta con un modelo de predicción de churn para clientes, pero no para restaurantes. De esta manera se lanzan

campañas de adquisición y retención de usuarios de acuerdo a sus comportamientos pasados.

Así como se hacen análisis prescriptivos del análisis de churn de clientes, sería importante hacerlo también para restaurantes de manera de conseguir un manejo más eficiente de los recursos utilizados para mantener los restaurantes activos.

1.4 - Propuesta de trabajo:

A través de técnicas de aprendizaje automático supervisado, se creará un modelo que permite predecir la baja en la cartera de restaurantes activos en la plataforma para un periodo de tiempo determinado.

El modelo buscará asignar una probabilidad de churn para cada restaurante en un futuro próximo y de esta manera asignar distintas estrategias para lograr una mayor fidelización para recuperar la cuenta.

2 - Metodología

2.1 - Modelos

Para problemas de predicción con estas características, comúnmente se usan distintos modelos de ensambles de árboles para predecir el churn. En un primer análisis, se comparó la performance de los dos modelos más modernos y de mejor poder predictivo para entender fundamentalmente cuál era el más acertado. En esta comparación se usaron los modelos XGBoost y Light GMB para predecir el churn.

El primer modelo analizado es XGBoost, que es una versión mejorada del algoritmo de aumento de gradiente, muy eficiente en la reducción de sesgo y la varianza de un modelo. La idea detrás de su concepto es crear muchos árboles de decisión de manera secuencial, donde se impulsan a los árboles débiles con bajo poder de predicción. Se utiliza un modelo más regularizado para reducir y controlar el sobreajuste y así mejorar su rendimiento. Además, tiene la capacidad de disminuir el consumo de tiempo junto con el uso de los recursos óptimos de memoria, ejecución paralela y manejo de los valores faltantes, mientras se genera la construcción del árbol. La implementación de XGBoost como árbol de algoritmos considera las características en el conjunto de datos como un nodo condicional, donde se divide en varias ramas y se parte hasta la hoja nodo que representa la detección seleccionada del problema (AL-Shatnwai, A. M., & Altibbi, M. F., 2020).

El segundo algoritmo en estudio es Light GBM[1] (Gradient Boosting Machine) desarrollado por Microsoft. El sistema está basado en el ensamble de árboles de decisión de débil poder de clasificación que en su conjunto hacen un clasificador robusto.

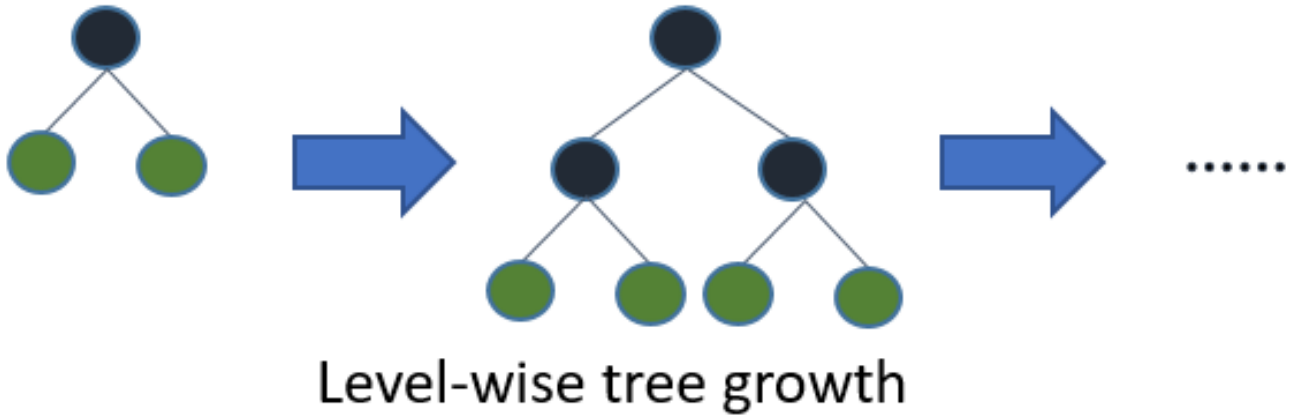
Está diseñado para ser distribuido y eficiente con las siguientes ventajas:

- Mayor velocidad de entrenamiento y mayor eficiencia.
- Menor uso de memoria.
- Mejor precisión.
- Soporte de aprendizaje paralelo y GPU.
- Capaz de manejar datos a gran escala.

[1] LightGBM utiliza algoritmos basados en histogramas, que agrupan valores de características continuas (atributos) en contenedores discretos. Esto acelera el entrenamiento y reduce el uso de la memoria (Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, Tie-Yan Liu., 2017).

La mayoría de los algoritmos de aprendizaje de árboles de decisión hacen crecer los árboles por nivel (profundidad), como en la siguiente imagen:

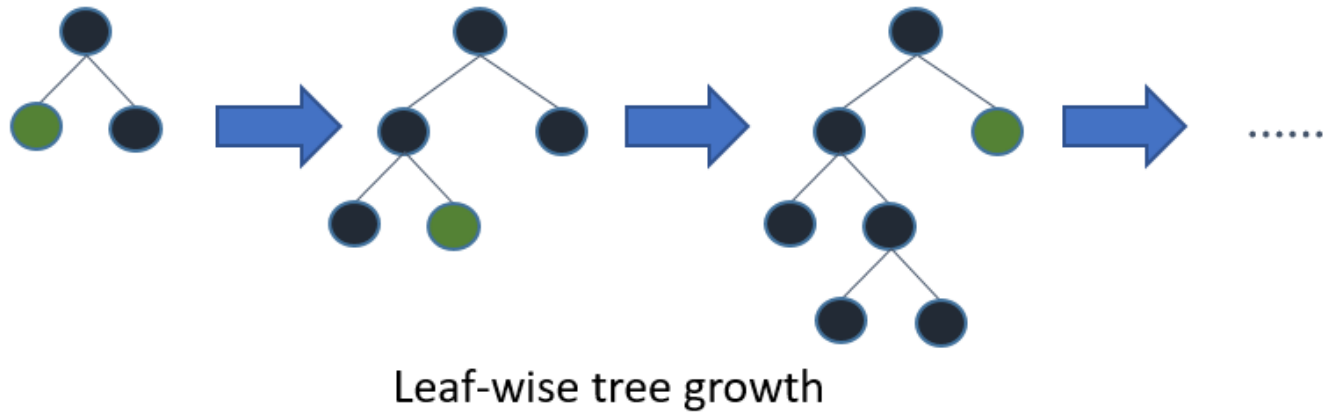
Figura 1: árboles de decisión hacen crecer el árbol por nivel (profundidad)



Esto genera un diagrama que representa el crecimiento del árbol por niveles, en el que el mejor nodo posible se divide un nivel hacia abajo. La estrategia da como resultado un árbol simétrico, donde cada nodo en un nivel tiene nodos secundarios que dan como resultado una capa adicional de profundidad (Shi, H, 2007).

LightGBM introduce la creación de árboles por hojas, donde se elegirá la hoja con el máximo delta en la función de pérdida para crecer. Manteniendo la cantidad de hojas fijo, los algoritmos basados en hojas tienden a lograr una función de pérdida más baja que los algoritmos basados en niveles:

Figura 2: LightGBM hace crecer el árbol por hojas. Al cultivar la misma hoja, el algoritmo de hojas puede reducir más pérdidas que el algoritmo de niveles.



Leaf-wise tree growth

Un diagrama que representa el crecimiento del árbol por hojas en el que solo el nodo con el mayor delta en la función de pérdida se divide y no se molesta con el resto de los nodos en el mismo nivel. Esto da como resultado un árbol asimétrico donde la subsiguiente división ocurre solo en un lado del árbol.

Finalmente, se eligió Light GBM como modelo por su optimización de uso de la memoria, su velocidad de procesamiento, y que maneja las variables categóricas de una manera más eficiente que XGBoost.

2.2 - Hiperparámetros

Light GBM provee más de 100 parámetros para ajustar, de los cuales se usaron los siguientes para el modelo:

Tabla 1: Listado de hiperparametros para Light GBM y su función

Hiperparámetro	**función**
Max_depth	limitar la profundidad máxima para el modelo de árbol. <= 0 significa que no hay límite.
num_leaves	número máximo de hojas por árbol
learning_rate	tasa de contracción
n_estimators	número de iteraciones
reg_lambda	Regularización L2
subsample	LightGBM seleccionará aleatoriamente un subconjunto de variables. Se puede utilizar para acelerar el entrenamiento y para tratar el exceso de ajuste
colsample_bytree	LightGBM seleccionará aleatoriamente un subconjunto de variables en cada iteración (árbol). Se puede utilizar para acelerar el entrenamiento y para tratar el exceso de ajuste
boosting_type	tipo de boosting
scale_pos_weight	Peso de la clase positiva para contrarrestar que la base está desbalanceada

(Microsoft Corporation, 2022)

2.3 - Entrenamiento, validación y testeo

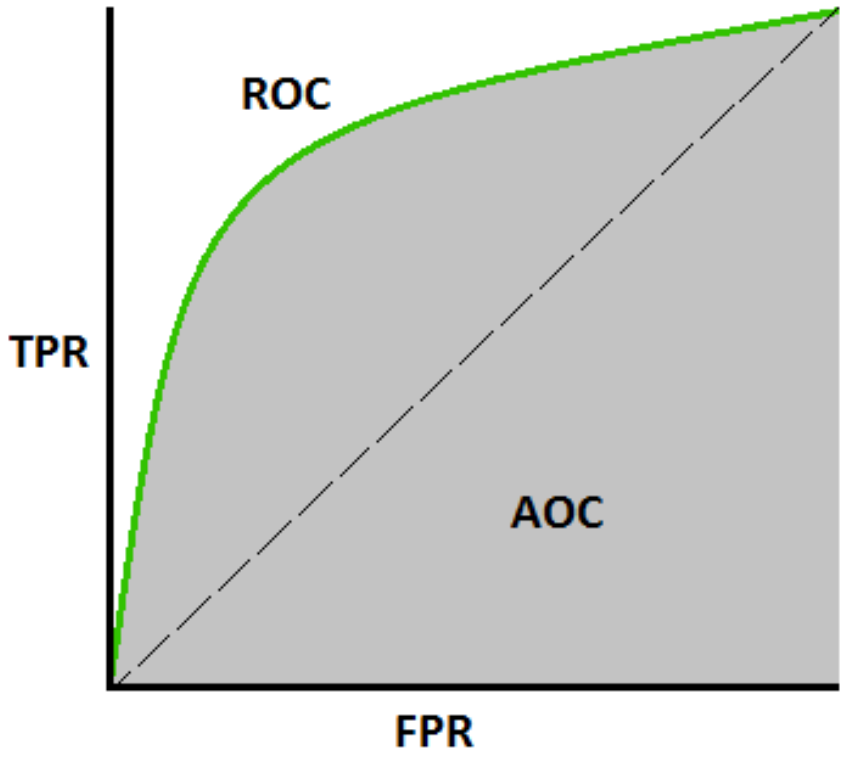
Al estar usando datos temporales, el armado de los conjuntos de entrenamiento, validación y testeo se separan de tal manera que el modelo aprenda y valide los resultados para ayudar a predecir para el mes que le sigue tratando de emular como se va a utilizar el modelo una vez que esté implementado. Teniendo en cuenta la temporalidad de los datos.

- Se entrena el modelo con los datos de noviembre 2021
- Se validan con los datos de diciembre 2021
- Se testean con la información de enero 2022

2.4 - Métricas de evaluación:

La métrica a utilizar para evaluar el modelo es el área bajo la curva ROC (ROC AUC). Esta es una medida de rendimiento para los problemas de clasificación en varias configuraciones de umbral, especialmente útil para problemas de clasificación con clases desbalanceadas. Indica cuánto es capaz el modelo de distinguir entre clases. La curva ROC se traza con TPR contra FPR, donde TPR está en el eje y y FPR está en el eje x (Marzban, C. 2004).

Figura 3: TPR / FPR



Un modelo excelente tiene un área cerca de 1, lo que significa que tiene una buena medida de separabilidad. Esta es una situación ideal, donde es perfectamente capaz de distinguir entre clase positiva y clase negativa.

Cuando dos distribuciones se superponen, introducimos errores de tipo 1 y tipo 2. Dependiendo del umbral, podemos minimizarlos o maximizarlos.

Y cuando AUC es 0.5, significa que el modelo no tiene capacidad de separación de clases. Esta es la peor situación. El modelo no tiene capacidad de discriminación para distinguir entre clase positiva y clase negativa.

Un modelo pobre tiene un AUC cercano a 0, lo que significa que tiene la peor medida de separabilidad. De hecho, significa que está reciprocando el resultado. Está prediciendo 0 como 1 y 1 como 0. Significa que el modelo predice una clase negativa como una clase positiva y viceversa (James G, Witten D, Hastie T, Tibshirani R. 2017).

3 - Datos

3.1 - Conjuntos de datos

Los datos provienen de diferentes conjuntos de datos dentro de la compañía que se recopilan de manera diaria. Entre ellos se encuentran:

- Órdenes: base de datos que recopila información sobre las órdenes.
    - Order ID: Código único para cada órden
    - El estado de la orden: si está fue confirmada o rechazada
    - Órdenes Confirmadas: Cantidad de órdenes confirmadas procesadas durante el período informado
    - GFV: Valor Bruto del pedido, neto de Descuentos: Monto total pagado por los usuarios, sin costo de envío, por pedidos confirmados. Valor de la comida pagado por el cliente después de aplicar los descuentos. EUR
    - GMV: Valor Bruto de Mercado neto de Descuentos: Valor total del boleto pagado por el cliente después de aplicar los descuentos, incluido el monto del envío, para pedidos confirmados. GMV = GFV + costo de envío. EUR
    - Tasa de comisión: Ingresos por comisión / GFV.
    - Consumidores activos: Cliente asociado a cada órden
    - Frecuencia: Pedidos Confirmados / Usuarios Activos. Frecuencia de pedidos confirmados en el segmento de "usuarios activos".

- Logística
    - Porcentaje retraso del proveedor >= 10 minutos: Porcentaje de pedidos donde el tiempo de preparación efectivo es mayor al estimado por 10 minutos o más, sobre el total de pedidos confirmados.
    - Tiempo de preparación real promedio: Tiempo desde que el restaurante recibe el pedido hasta que lo tiene listo para entregar al repartidor, medido en minutos.

- Restaurantes: Datos sobre los perfiles de restaurantes
    - Restaurant ID: Código único para cada órden
    - País: Donde opera el restaurante
    - Ciudad: Donde opera el restaurante
    - Área: Donde opera el restaurante

- Nombre del restaurante: Nombre o marca bajo el que opera el restaurant
- tipo de cocina: Que tipo de cocina se prepara. (ej: hamburguesas, pizza, etc)
- tipo de delivery: Si cuenta con delivery propio o usa la logistica de la aplicación
- Días Online: cantidad de días al mes con el perfil activo
- Horario de atención real: Horario real que estuvo con el perfil activo
- Restaurantes en línea: Restaurantes disponibles para ordenar dentro del período del informe.
- Restaurantes activos: Restaurantes que recibieron al menos 1 pedido de un cliente durante el período informado.
- Restaurantes zombis: Restaurantes en línea con cero pedidos confirmados en un período de tiempo.
- Restaurantes abandonados (churn): cantidad de restaurantes que estuvieron en línea en el período de informe anterior, sin embargo, no se conectaron ni una sola vez durante el período de informe actual.
- Tiempo Efectivo: Tiempo de apertura efectivo sobre el horario de apertura del cronograma de un restaurante. El tiempo de apertura efectivo es la diferencia entre el tiempo de apertura del cronograma y el tiempo de cierre de un restaurante debido a fechas especiales y disparadores. El restaurante debe estar en línea en la fecha del informe.
- Tasa de cancelación: Porcentaje del total de pedidos no confirmados sobre el total de pedidos (todos los estados). Pedidos no confirmados = Rechazados + Cancelados + Pendientes
- Tasa de cancelación del restaurante: porcentaje del total de pedidos no confirmados donde el motivo de la cancelación es responsabilidad del restaurante.
- Sesiones: Cantidad de sesiones recibidas en el perfil del restaurante.
- CVR3: Porcentaje de sesiones creadas por los usuarios en la plataforma principal que realizaron un pedido sobre todas las sesiones que visitaron los detalles de la tienda. CVR3 = Sesiones con Transacción Detalles de Tienda Visitada / Sesiones Visitadas Detalles de Tienda.

Se recopila la información de todas las fuentes para preparar una única base de datos con la información donde cada fila es un restaurante en un mes dado. La información se agrega de manera mensual para poder simplificar la base y poder hacer predicciones de manera mensual.

La métrica “Restaurantes abandonados (churn)” cuenta a mes cerrado cuando un restaurante no se conecto nunca en el periodo analizado. Es decir, dejó de operar en la aplicación en el periodo anterior, y en el mes en curso ya no se conectó ni una vez. Por ejemplo, si un restaurante prende por última vez su perfil en Julio, va a figurar como churn en Agosto.
Para poder predecir con antelación el churn, se corrió esta métrica un período para atrás. De esta manera indica efectivamente cuando un restaurante va a dejar de prender su perfil. Volviendo al ejemplo anterior, con esta metodología va a aparecer como churn en Julio, el mes donde dejó de operar.

3.2 - Ingeniería de atributos

Para poder incrementar el poder de predicción del modelo se incluyeron variables de tendencia. Para cada variable numérica, se agrega una variable más con la variación de la métrica contra el mes anterior para el mismo restaurante. De esta manera para cada entrada en la base de datos (més - restaurante), se va a poder dimensionar si crece o cae contra el periódo anterior.

Tabla 2 : Listado de variables de tendencia incluidos en la base de datos

Variables de tendencia[2]
Orders_vs_pp
Sessions_vs_pp
agg_fact_sessions_by_partner_daily_cvr_1_vs_pp
fact_orders_percentage_commercial_fail_rate_vs_pp
fact_orders_customers_with_confirmed_orders_vs_pp
agg_fact_partner_times_actual_open_time_rate_vs_pp
agg_fact_partner_times_open_times_hour_vs_pp
dim_historical_partners_qty_online_days_vs_pp
fact_orders_kpi_gfv_gross_vs_pp
fact_orders_kpi_gmv_vs_pp
fact_orders_kpi_fee_revenue_vs_pp
fact_orders_kpi_total_comm_revenue_vs_pp
fact_orders_total_commission_order_1_vs_pp
dim_partner_avg_rating_vs_pp
logistic_orders_kpi_percentage_vendor_late_10_vs_pp
fact_logistic_orders__deliveries_kpi_avg_assumed_actual_preparation_time_vs_pp
fact_logistic_orders__deliveries_avg_delivery_distance_vs_pp
logistic_orders_kpi_avg_actual_delivery_time_vs_pp

Estas variables intentan captar el comportamiento de los restaurantes a lo largo del tiempo para cada variable predictiva.

Por otro lado, a la hora de entrenar el modelo, se retiraron las variables “Partner ID” y “Partner name” de la base de datos. para potenciar el poder predictivo del modelo, y hacer la base más liviana antes de realizar One Hot encoding. Estas variables tenían una baja correlación en el análisis descriptivo, y además intuitivamente no deberían poder explicar el churn.

[2] “vs pp” = versus el período anterior

3.3 - Exploración de los datos

En esta sección se analiza la información recopilada y se buscan tendencias y patrones que ayuden a entender mejor los datos, para preparar una mejor estrategia a la hora de armar el modelo.

La base cuenta con 815.013 observaciones con 47 variables, correspondiente a 111.503 restaurantes únicos desde enero de 2021 hasta febrero de 2022.

Tabla 3: Variables presentes en el conjunto de datos

Variable	Non-null values	Data type
date_lookup_date_month	815,013	object
dim_country_country_name	815,013	object
dim_partner__city_city_name	815,013	object
dim_area_area_name	815,013	object
dim_partner__concept_concept_name	7,466	object
dim_partner_partner_id	815,013	int64
dim_partner_partner_name	815,013	object
dim_partner_main_cousine_category_name	796,825	object
dim_partner_is_logistic_market_place	815,013	object
fact_orders_total_confirmed_orders	815,013	int64
agg_fact_sessions_by_partner_daily_qty_sessions	815,013	int64
agg_fact_sessions_by_partner_daily_cvr_1	796,882	float64
fact_orders_percentage_commercial_fail_rate	639,008	float64
fact_orders_customers_with_confirmed_orders	815,013	int64
agg_fact_partner_times_actual_open_time_rate	815,013	float64
agg_fact_partner_times_open_times_hour	815,013	float64
Daily_Online_Hours	815,013	float64
dim_historical_partners_qty_online_days	815,013	int64
fact_orders_kpi_gfv_gross	815,013	float64
fact_orders_kpi_gmv	815,013	float64
fact_orders_kpi_fee_revenue	815,013	float64
fact_orders_kpi_total_comm_revenue	815,013	float64
fact_orders_total_commission_order_1	615,592	float64
dim_partner_avg_rating	583,588	float64
logistic_orders_kpi_percentage_vendor_late_10	539,014	float64
fact_logistic_orders__deliveries_kpi_avg_assumed_actual_preparation_time	519,164	float64
fact_logistic_orders__deliveries_avg_delivery_distance	563,304	float64
logistic_orders_kpi_avg_actual_delivery_time	563,304	float64
fact_partners_monthly_is_churned	815,013	object
Orders_vs_pp	558,268	float64
Sessions_vs_pp	687,292	float64
agg_fact_sessions_by_partner_daily_cvr_1_vs_pp	575,101	float64
fact_orders_percentage_commercial_fail_rate_vs_pp	380,071	float64
fact_orders_customers_with_confirmed_orders_vs_pp	558,268	float64
agg_fact_partner_times_actual_open_time_rate_vs_pp	655,939	float64
agg_fact_partner_times_open_times_hour_vs_pp	656,030	float64
dim_historical_partners_qty_online_days_vs_pp	698,749	float64
fact_orders_kpi_gfv_gross_vs_pp	558,266	float64
fact_orders_kpi_gmv_vs_pp	558,013	float64
fact_orders_kpi_fee_revenue_vs_pp	469,107	float64
fact_orders_kpi_total_comm_revenue_vs_pp	557,632	float64
fact_orders_total_commission_order_1_vs_pp	557,634	float64
dim_partner_avg_rating_vs_pp	517,432	float64
logistic_orders_kpi_percentage_vendor_late_10_vs_pp	382,118	float64
fact_logistic_orders__deliveries_kpi_avg_assumed_actual_preparation_time_vs_pp	469,754	float64
fact_logistic_orders__deliveries_avg_delivery_distance_vs_pp	507,056	float64
logistic_orders_kpi_avg_actual_delivery_time_vs_pp	491,415	float64

Tabla 4: Variables numéricas y sus principales estadísticos descriptivos

Variables numéricas	*count*	*mean*	*std*	*min*	*25%*	*50%*	*75%*	*max*
dim_partner_partner_id	815,013	195,121	71,548	15	147,714	203,478	249,124	328,694
fact_orders_total_confirmed_orders	815,013	74	140	0	1	22	88	5,883
agg_fact_sessions_by_partner_daily_qty_sessions	815,013	699	910	0	90	409	965	37,352
agg_fact_sessions_by_partner_daily_cvr_1	796,882	0	0	0	0	0	0	1
fact_orders_percentage_commercial_fail_rate	639,008	0	0	0	0	0	0	1
fact_orders_customers_with_confirmed_orders	815,013	61	111	0	1	20	76	3,718
agg_fact_partner_times_actual_open_time_rate	815,013	1	0	0	0	1	1	1
agg_fact_partner_times_open_times_hour	815,013	139	107	0	47	123	215	741
Daily_Online_Hours	815,013	5	5	0	2	5	8	24
dim_historical_partners_qty_online_days	815,013	27	7	1	28	30	31	31
fact_orders_kpi_gfv_gross	815,013	274,215	1,063,079	0	11	10,590	129,630	62,977,979
fact_orders_kpi_gmv	815,013	287,663	1,112,357	0	10	11,005	135,701	66,730,280
fact_orders_kpi_fee_revenue	815,013	21,326	97,490	0	0	136	6,420	6,779,091
fact_orders_kpi_total_comm_revenue	815,013	64,828	239,291	-0	2	2,402	28,697	15,548,000
fact_orders_total_commission_order_1	615,592	24	6	0	20	25	27	84
dim_partner_avg_rating	583,588	4	1	1	4	4	5	5
logistic_orders_kpi_percentage_vendor_late_10	539,014	0	0	0	0	0	0	1
fact_logistic_orders__deliveries_kpi_avg_assumed_actual_preparation_time	519,164	17	6	3	13	17	21	59
fact_logistic_orders__deliveries_avg_delivery_distance	563,304	2	28	0	1	2	2	13,697
logistic_orders_kpi_avg_actual_delivery_time	563,304	28	11	0	23	29	34	689
fact_partners_monthly_is_churned	815,013	0	0	0	0	0	0	1
Orders_vs_pp	558,268	1	5	-1	-0	-0	0	1,090
Sessions_vs_pp	687,292	3	36	-1	-0	-0	0	3,635
agg_fact_sessions_by_partner_daily_cvr_1_vs_pp	575,101	0	1	-1	-0	0	0	167
fact_orders_percentage_commercial_fail_rate_vs_pp	380,071	0	3	-1	-1	-0	0	401
fact_orders_customers_with_confirmed_orders_vs_pp	558,268	0	5	-1	-0	-0	0	930
agg_fact_partner_times_actual_open_time_rate_vs_pp	655,939	-115	67,838	-1	-0	-0	0	18,766
agg_fact_partner_times_open_times_hour_vs_pp	656,030	-613	364,253	-1	-0	-0	0	20,819
dim_historical_partners_qty_online_days_vs_pp	698,749	1	3	-1	-0	0	0	30
fact_orders_kpi_gfv_gross_vs_pp	558,266	2	277	-1	-0	-0	0	140,999
fact_orders_kpi_gmv_vs_pp	558,013	1	36	-1	-0	-0	0	24,170
fact_orders_kpi_fee_revenue_vs_pp	469,107	1	8	-1	-0	-0	0	1,149
fact_orders_kpi_total_comm_revenue_vs_pp	557,632	1	200	-1	-0	-0	0	110,239
fact_orders_total_commission_order_1_vs_pp	557,634	-0	1	-1	0	0	0	270
dim_partner_avg_rating_vs_pp	517,432	0	0	0	0	0	0	0
logistic_orders_kpi_percentage_vendor_late_10_vs_pp	382,118	0	1	-1	-1	-0	0	42
fact_logistic_orders__deliveries_kpi_avg_assumed_actual_preparation_time_vs_pp	469,754	-0	0	-1	-0	-0	0	7
fact_logistic_orders__deliveries_avg_delivery_distance_vs_pp	507,056	0	9	-1	-0	-0	0	3
logistic_orders_kpi_avg_actual_delivery_time_vs_pp	491,415	-0	0	-1	-0	-0	0	68

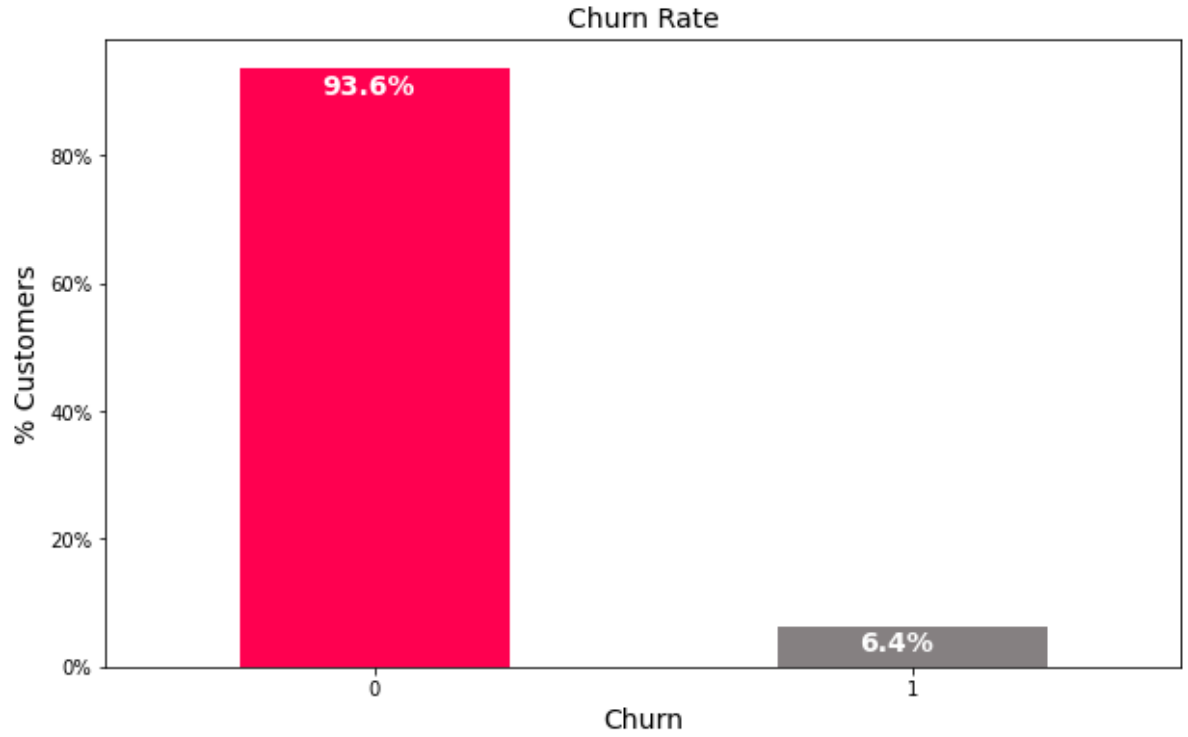
Tabla 5: Variables categóricas y cardinalidad

Variable categórica	Count Unique	Count
date_lookup_date_month	14	815,013
dim_country_country_name	15	815,013
dim_partner__city_city_name	531	815,013
dim_area_area_name	3,198	815,013
dim_partner__concept_concept_name	16	7,466
dim_partner_partner_name	111,478	815,013
dim_partner_main_cousine_category_name	85	796,825
dim_partner_is_logistic_market_place	2	815,013

Tabla 6: Cantidad de restaurantes por país

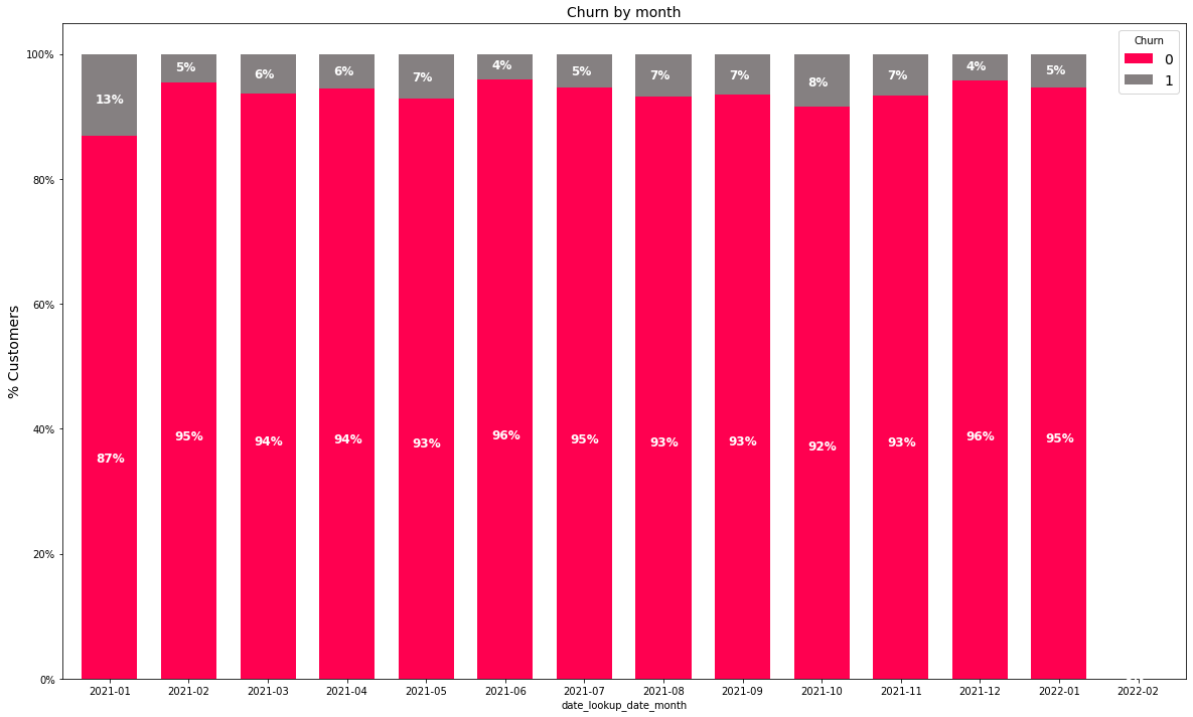
Country name	Number of Restaurants
Argentina	36,598
Bolivia	6,941
Chile	18,655
Costa Rica	3,457
Ecuador	7,687
El Salvador	2,327
Guatemala	2,639
Honduras	1,315
Nicaragua	1,466
Panamá	3,116
Paraguay	4,505
Perú	8,440
República Dominicana	7,106
Uruguay	2,408
Venezuela	4,843
Total	**111,503**

Figura 4: Tasa de churn promedio - últimos 12 meses - Latam



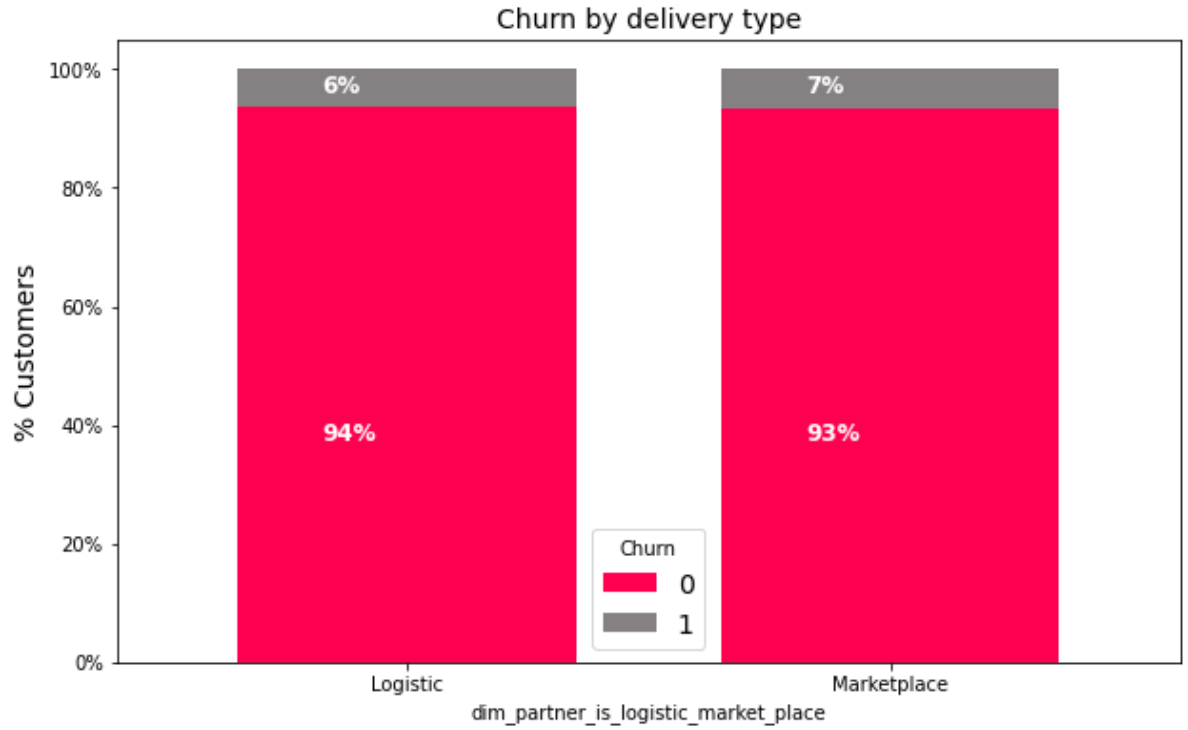
Analizando el churn, vemos que es una métrica desbalanceada donde más del 93% de las muestras no se dan de baja en la aplicación y siguen operando en el mes siguiente. Solamente un 6% de los restaurantes activos en un mes dejan de operar al mes siguiente.

Figura 5: Tasa de churn mensual - Latam



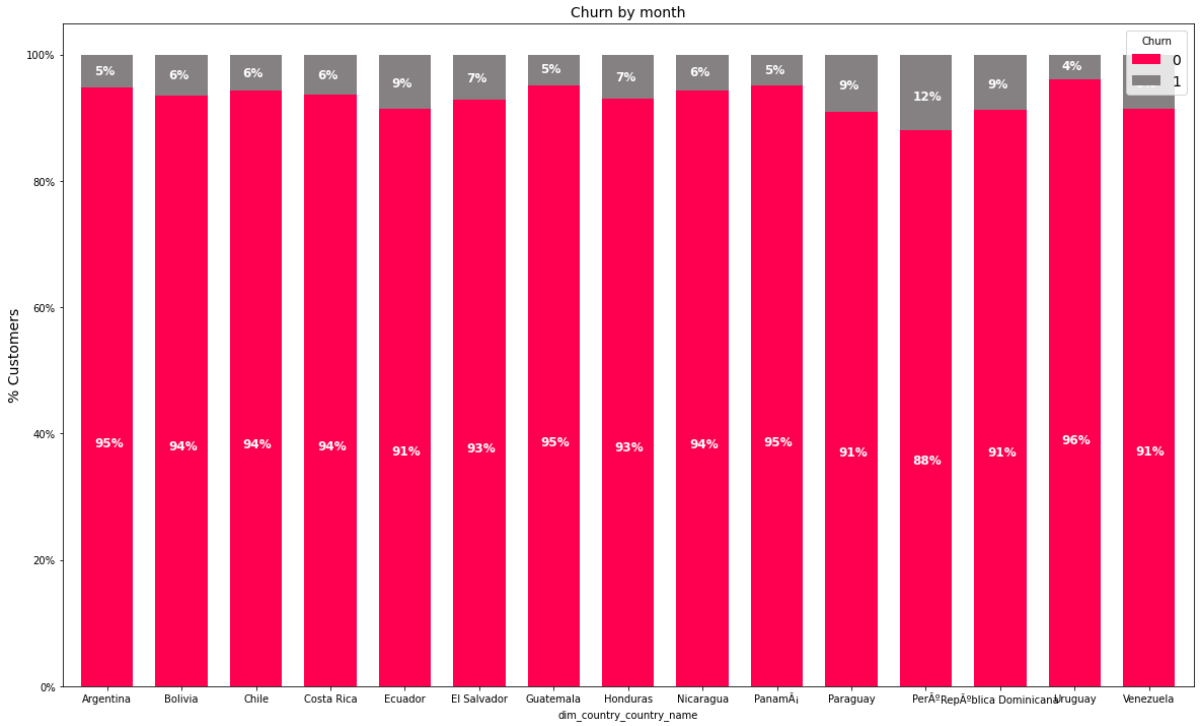
Si bien no se ve una estacionalidad marcada, es importante destacar que noviembre es un mes con un churn alto, y luego diciembre y enero tienen una tasa por debajo del promedio. Esto es importante porque son los meses que se van a utilizar como Conjuntos de entrenamiento, validación y testeo.

Figura 6: Tasa de churn promedio por tipo de delivery - últimos 12 meses - Latam



En este gráfico no vemos una diferencia significativa entre los restaurantes que cuentan con logística propia, que usan la aplicación únicamente como marketplace, versus los restaurantes que aprovechan la flota logística proporcionada por la aplicación. Intuitivamente se esperaría que los restaurantes que usan la logística de la app tengan una menor cantidad de bajas, dado que dependen más de la plataforma para hacer sus entregas. Si bien trabajan con dos modelos operativos muy diferentes, no vemos una diferencia en su tasa de churn.

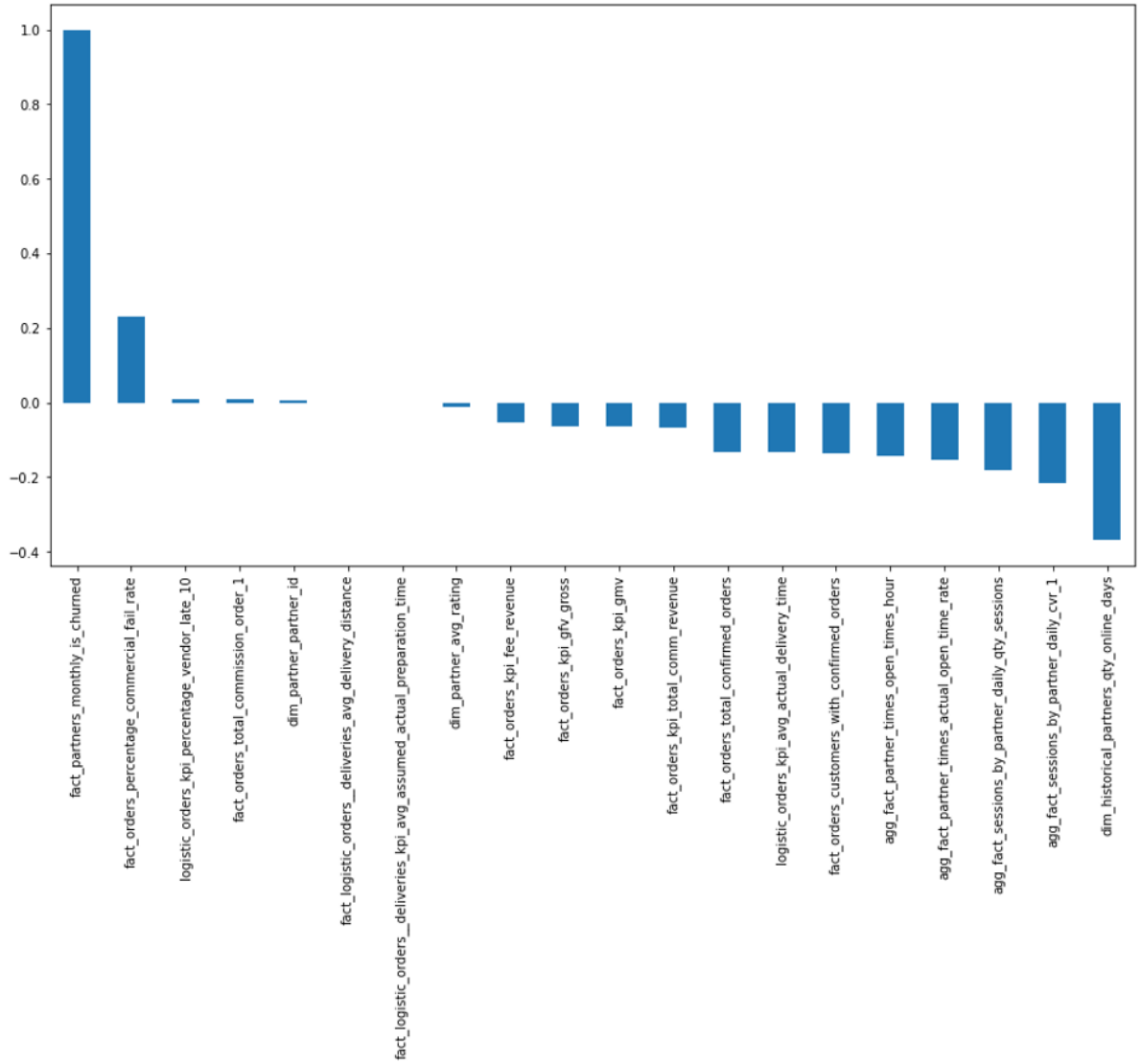
Figura 7:Tasa de churn promedio por tipo de delivery - últimos 12 meses - Latam



En este gráfico podemos ver que Perú, Venezuela y Ecuador son los países con una mayor tasa de bajas, mientras que Argentina, Guatemala, Panamá y Uruguay son los países con más estabilidad de restaurantes.

Perú y Ecuador son países donde el mercado de delivery es muy reñido, y es normal que un restaurante trabaje en más de una aplicación a la vez y no dependa tanto de una en particular. Por el caso contrario, Argentina, Panamá Guatemala y Uruguay son países donde la empresa es líder, y los restaurantes dependen en gran medida de la app para hacer generar ventas.

Figura 8:Correlación entre churn y el resto de las variables numéricas en la base de datos:



Acá se puede ver cuales son las métricas que más se mueven cuando los restaurantes hacen churn.

Dentro de las métricas con correlación negativa, vemos que los días online, la conversión, las sesiones, y las horas online son las que tienen la mayor correlación con el churn. Por el contrario, el porcentaje de cancelaciones tiene una correlación positiva.

Intuitivamente esto tiene sentido, dado que a peores métricas operativas, menor es la motivación de seguir operando en la aplicación.

Por otro lado, es sorprendente que la cantidad de órdenes o la facturación no sean de las variables con mayor correlación, dado que son la métrica más importantes en cualquier industria.

Figura 9: densidad días online para restaurantes que hicieron churn

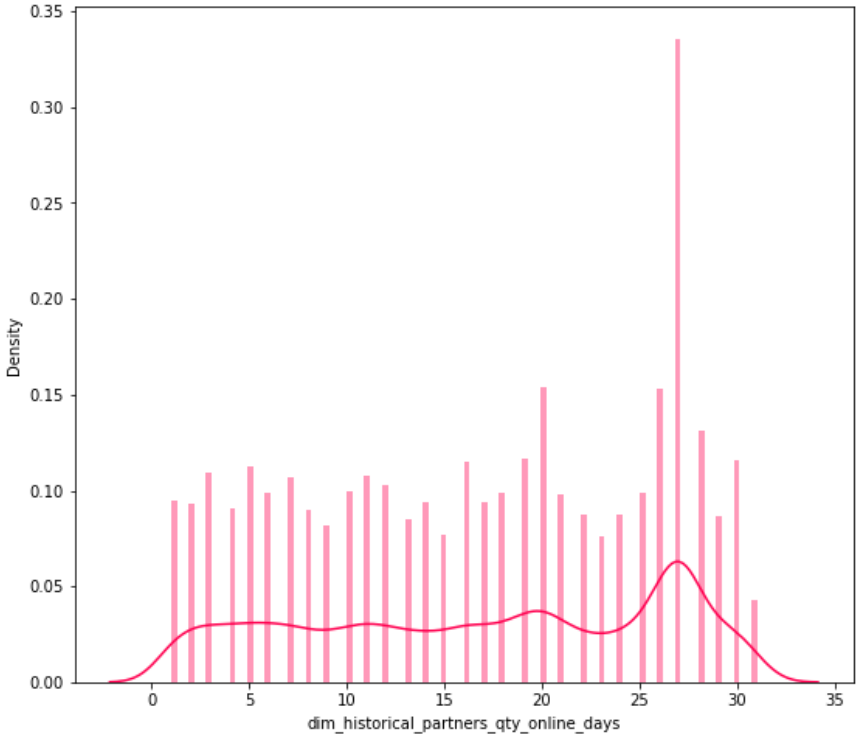


Figura 10: densidad días online para restaurantes que no hicieron churn

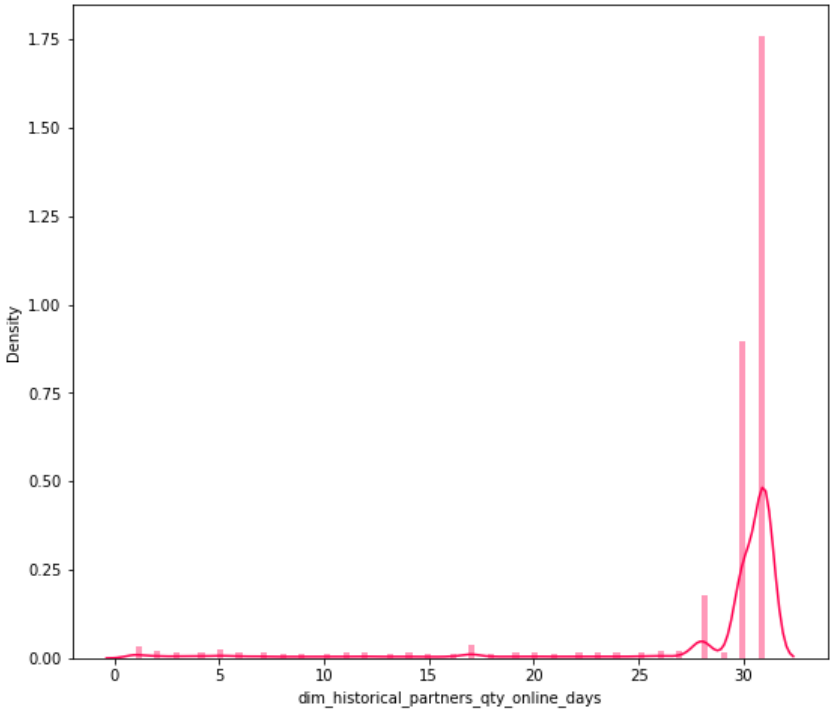


Tabla 7: Comparación análisis descriptivo días Online en restaurantes que hicieron churn vs donde no lo hicieron

Online days	*Churn*	*No churn*
Count	***47,691***	***698,749***
Mean	***16.84***	***28.04***
std	***8.97***	***6.77***
Min	***1***	***1***
25%	***9***	***30***
50%	***18***	***31***
75%	***26***	***31***
max	***31***	***31***

En estos gráficos podemos destacar la gran diferencia de días online entre los restaurantes que siguieron operando y los que abandonaron la aplicación, donde estos últimos tuvieron un promedio 40% más bajo que los que se mantuvieron operando en la aplicación.

Figura 11: densidad % tiempo abierto sobre el pactado por contrato para restaurantes que hicieron churn

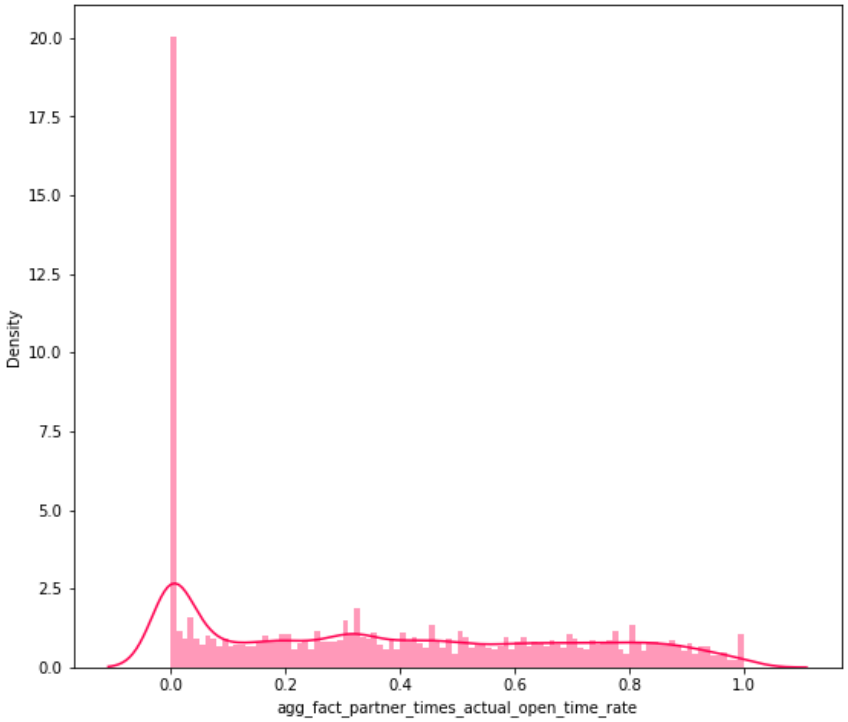


Figura 12: densidad % tiempo abierto sobre el pactado por contrato para restaurantes que no hicieron churn

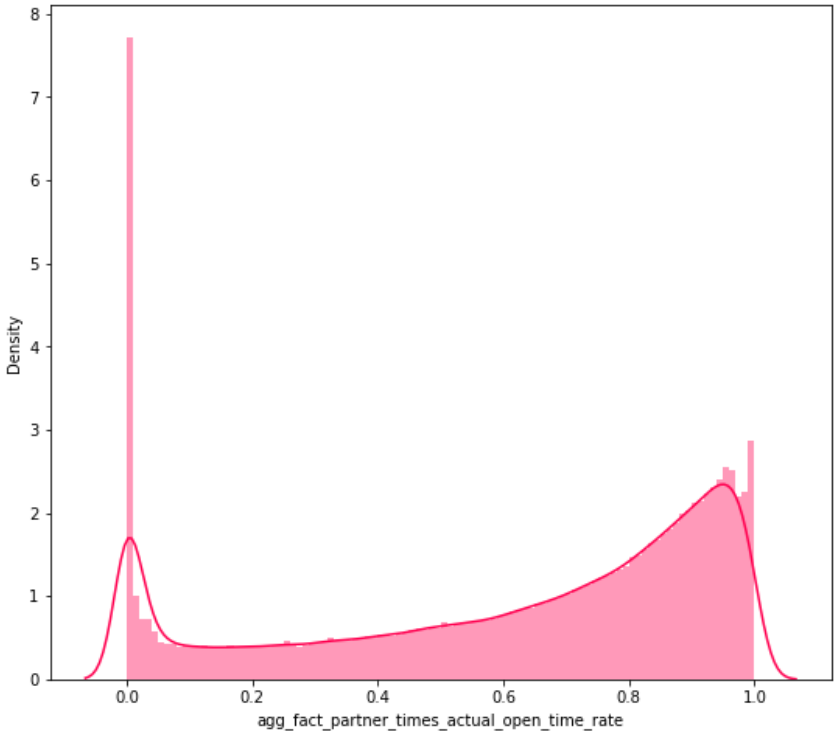


Tabla 8: Comparación análisis descriptivo % tiempo abierto en restaurantes que hicieron churn vs donde no lo hicieron

Open time rate	*Churn*	*No churn*
Count	*47,691*	*698,749*
Mean	*0.3767*	*0.6196*
std	*0.3121*	*0.3229*
Min	*0*	*0*
25%	*0.0553*	*0.3885*
50%	*0.3333*	*0.7267*
75%	*0.6452*	*0.8930*
max	*1*	*1*

De acuerdo a estos datos también podemos confirmar el grán impacto que tiene el %de tiempo abierto en la aplicación sobre el acordado por contrato, donde los partners que se bajan están un 39% de tiempo menos conectados.

Figura 13: densidad tiempo abierto por día para restaurantes que hicieron churn

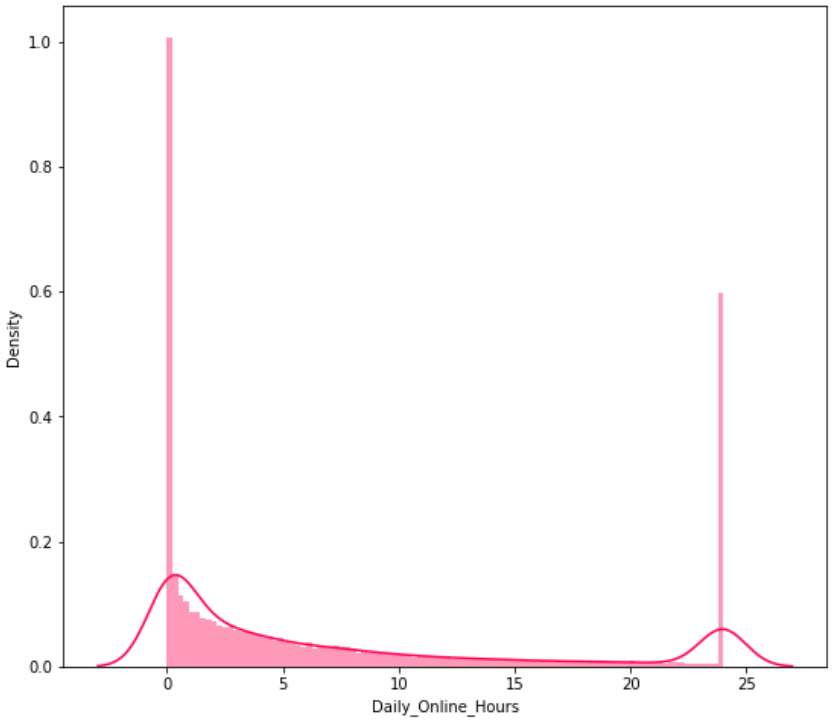


Figura 14: densidad tiempo abierto por día para restaurantes que no hicieron churn

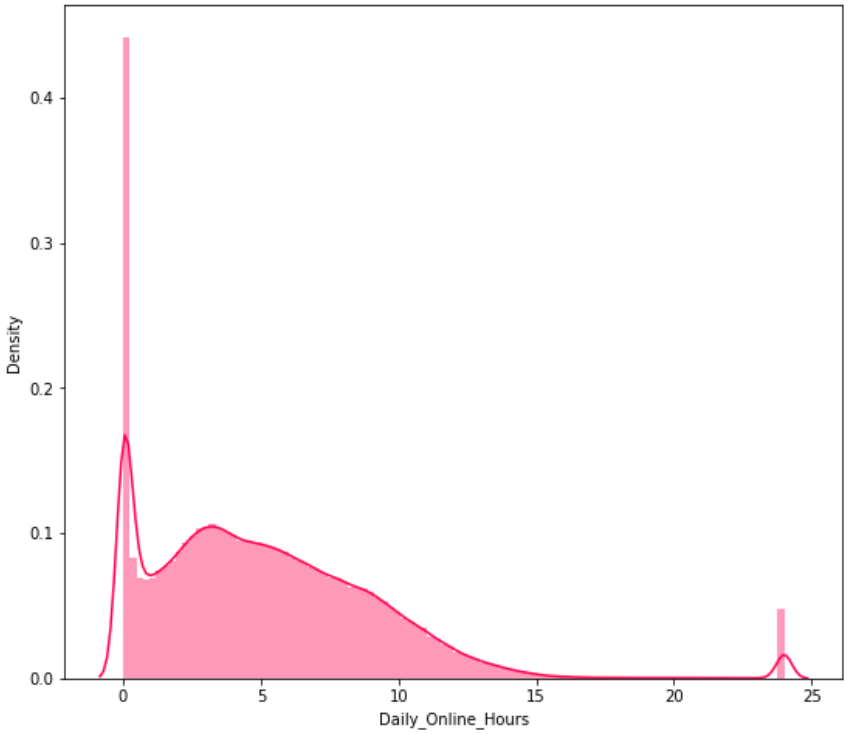


Tabla 9: Comparación análisis descriptivo horas online por día en restaurantes que hicieron churn vs donde no lo hicieron

Online Hours	*Churn*	*No churn*
Count	*47,691*	*698,749*
Mean	*7.28*	*5.21*
std	*8.56*	*4.14*
Min	*0.00*	*0.00*
25%	*0.29*	*2.13*
50%	*3.38*	*4.64*
75%	*11.65*	*7.64*
max	*24.00*	*24.00*

Analizando las horas absolutas que estuvieron con el perfil prendido, vemos que los restaurantes que hicieron churn están en promedio más tiempo con sus perfiles operativos, pero tienen un mayor desvío estándar y están distribuidos de manera más inequitativa que los restaurantes que no hacen churn.

Figura 15: densidad sesiones mensuales para restaurantes que hicieron churn

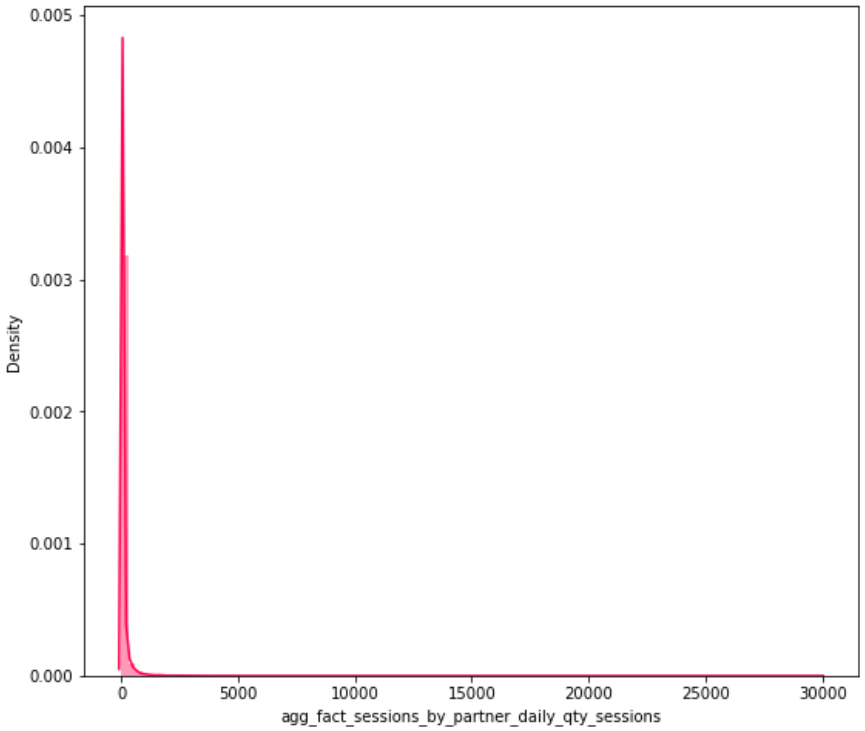


Figura 16: densidad sesiones mensuales para restaurantes que no hicieron churn

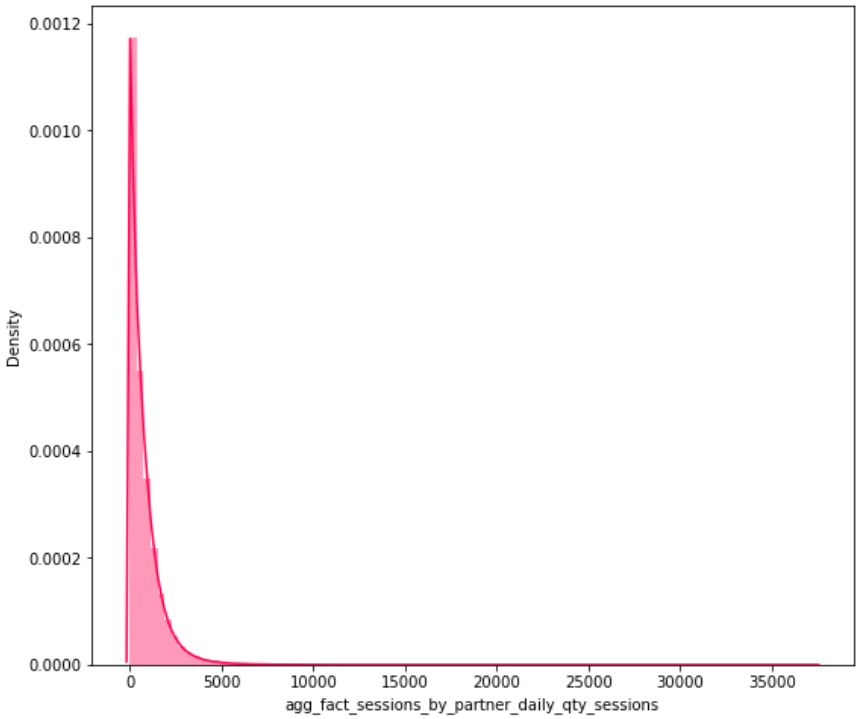


Tabla 10: Comparación análisis descriptivo sesiones mensuales en restaurantes que hicieron churn vs donde no lo hicieron

Sesiones por mes	*Churn*	*No churn*
Count	*47,691*	*698,749*
Mean	*68.41*	*753.69*
std	*269.44*	*934.19*
Min	*0*	*0*
25%	*4*	*134*
50%	*15*	*469*
75%	*49*	*1,033*
max	*29,941*	*37,352*

Para la cantidad de sesiones mensuales por restaurante vemos resultados muy dispares entre los restaurantes que permanecen vs los que salen de la aplicación. Donde los que salen hacen significativamente menos sesiones que los que logran permanecer operando en el siguiente mes.

Figura 17: densidad tasa de conversión de las sesiones para restaurantes que hicieron churn

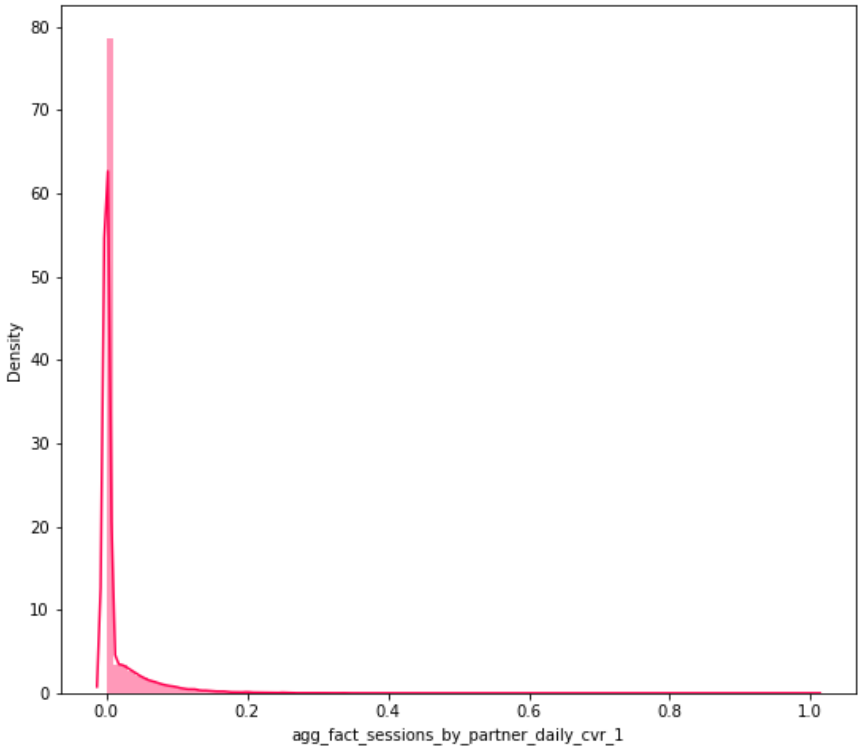


Figura 18: densidad tasa de conversión de las sesiones para restaurantes que no hicieron churn

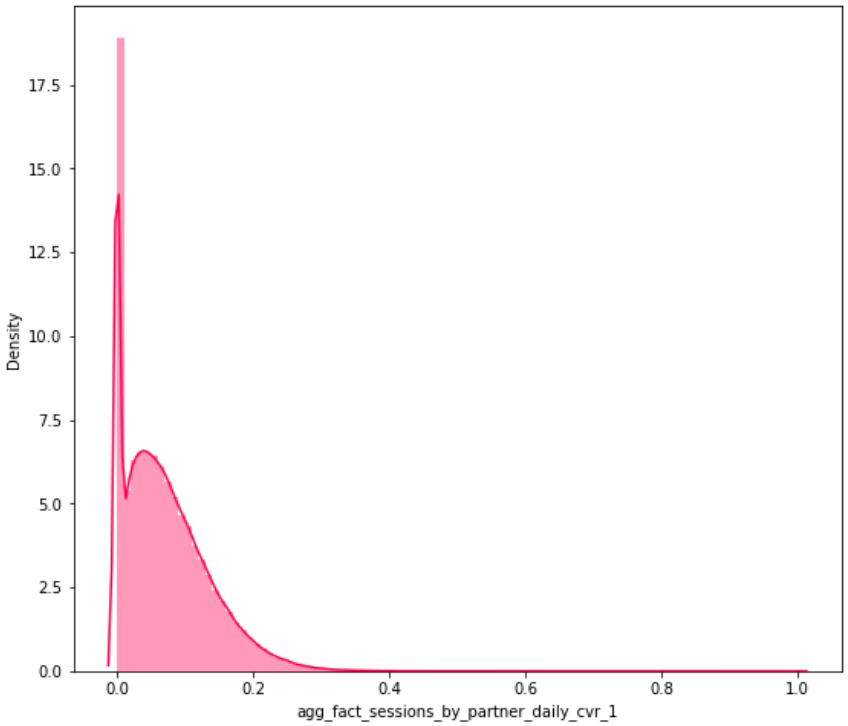


Tabla 11: Comparación análisis descriptivo conversión de las sesiones en restaurantes que hicieron churn vs donde no lo hicieron

Conversion rate	*Churn*	*No churn*
Count	*47,691*	*698,749*
Mean	*0.0133*	*0.0713*
std	*0.0375*	*0.0624*
Min	*0*	*0*
25%	*0*	*0.0211*
50%	*0*	*0.0598*
75%	*0*	*0.1073*
max	*1*	*1*

Para el caso de la conversión, vemos que los restaurantes que continuan operando en el futuro tienen un menú más atractivo para los usuarios, que logra atraerlos a comprar una vez que ingresan.

Figura 19: densidad tasa de cancelaciones para restaurantes que hicieron churn

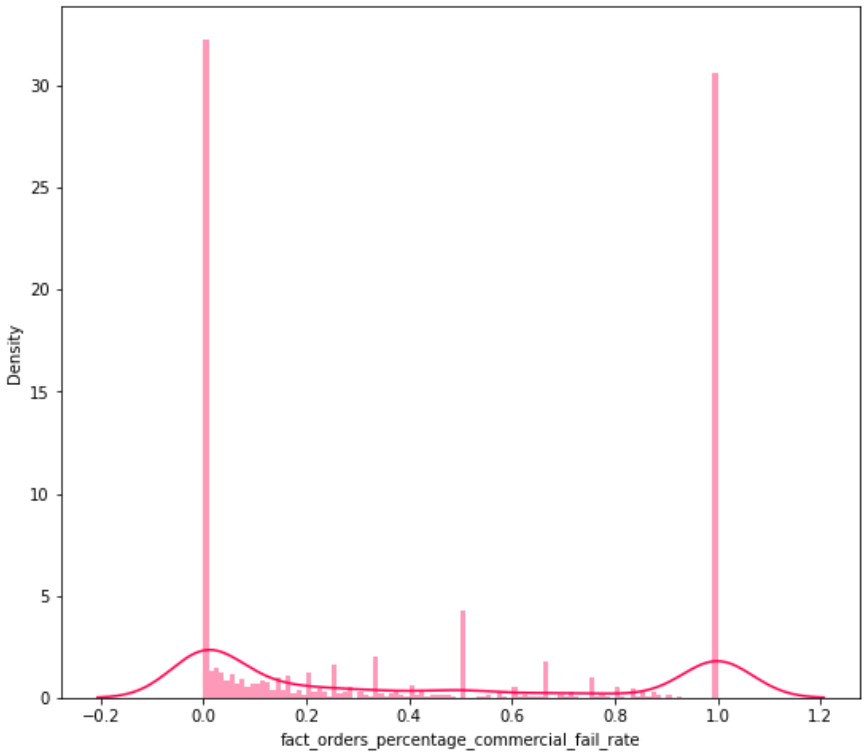


Figura 20: densidad tasa de cancelaciones para restaurantes que no hicieron churn

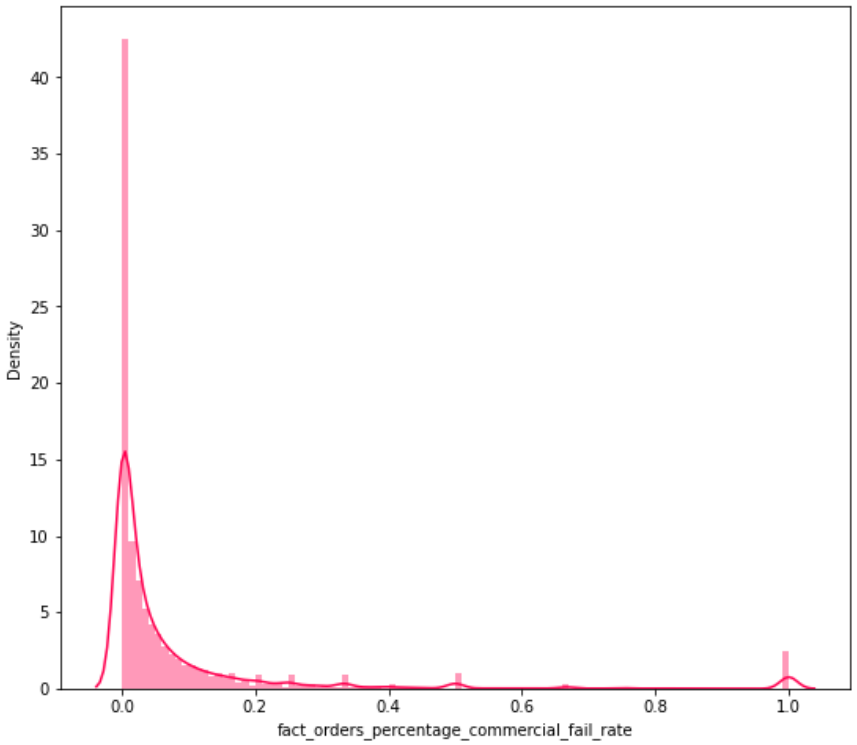


Tabla 12: Comparación análisis descriptivo tasa de cancelaciones en restaurantes que hicieron churn vs donde no lo hicieron

Restaurant Fail Rate	*Churn*	*No churn*
Count	*47,691*	*698,749*
Mean	*0.4266*	*0.0814*
std	*0.4323*	*0.1813*
Min	*0*	*0*
25%	*0*	*0*
50%	*0.25*	*0.0174*
75%	*1*	*0.0699*
max	*1*	*1*

Los restaurantes que terminan saliendo de la aplicación tienen una tasa de cancelaciones mucho mayor, al tener mayores problemas operativos y falta de stock. Esto se puede deber en parte a una falta de interés de continuar operando en la plataforma

4 - Resultados

4.2 - Análisis de resultados sin variables de tendencia

Para comenzar se dividió entre los modelos de entrenamiento, validación y testeo. Dada las características de la base de datos, y de la metodología de uso en producción del algoritmo, se eligió entrenar el modelo y elegir la combinación óptima de hiperparametros usando como base de entrenamiento el mes de noviembre. Luego se usó diciembre como base de validación, y por último se usó enero como testeo dado que es el último mes donde tenemos información disponible.

Para entrenar los modelos, se utilizó la función “Parametergrid” del paquete “sklearn” para iterar a través de las posibles combinaciones de hiperparametros hasta conseguir la composición que optimice el área bajo la curva ROC en la muestra de testeo.
Para el modelo de XGBoost, se encontraron los siguiente hiperparámetros como óptimos:

Tabla 13: Listado de hiperparametros y el valor encontrado en la optimización para XGBoost

Hiperparametro	valor óptimo
Max_Depth	5
learning_rate	0.05
n_estimators	300
reg_lambda	10
subsample	0.9
colsample_bytree	0.5
Eval_metric	‘mLogLoss’
scale_pos_weight	número de muestras que no hacen churn / número de muestras que si lo hacen (Kinnander, M. 2020)

Tabla 14: Valor del área bajo la curva ROC para cada muestra del modelo "xgboost" sin variables de tendencia

Muestra	AUC ROC
Train	0.96767
Validation	0.81957
Test	0.75982

Para el modelo de Light GBM, se obtuvo como resultados los siguientes hiperparametros:

Tabla 15: Listado de hiperparametros y el valor encontrado en la optimización para Light GBM sin variables de tendencia

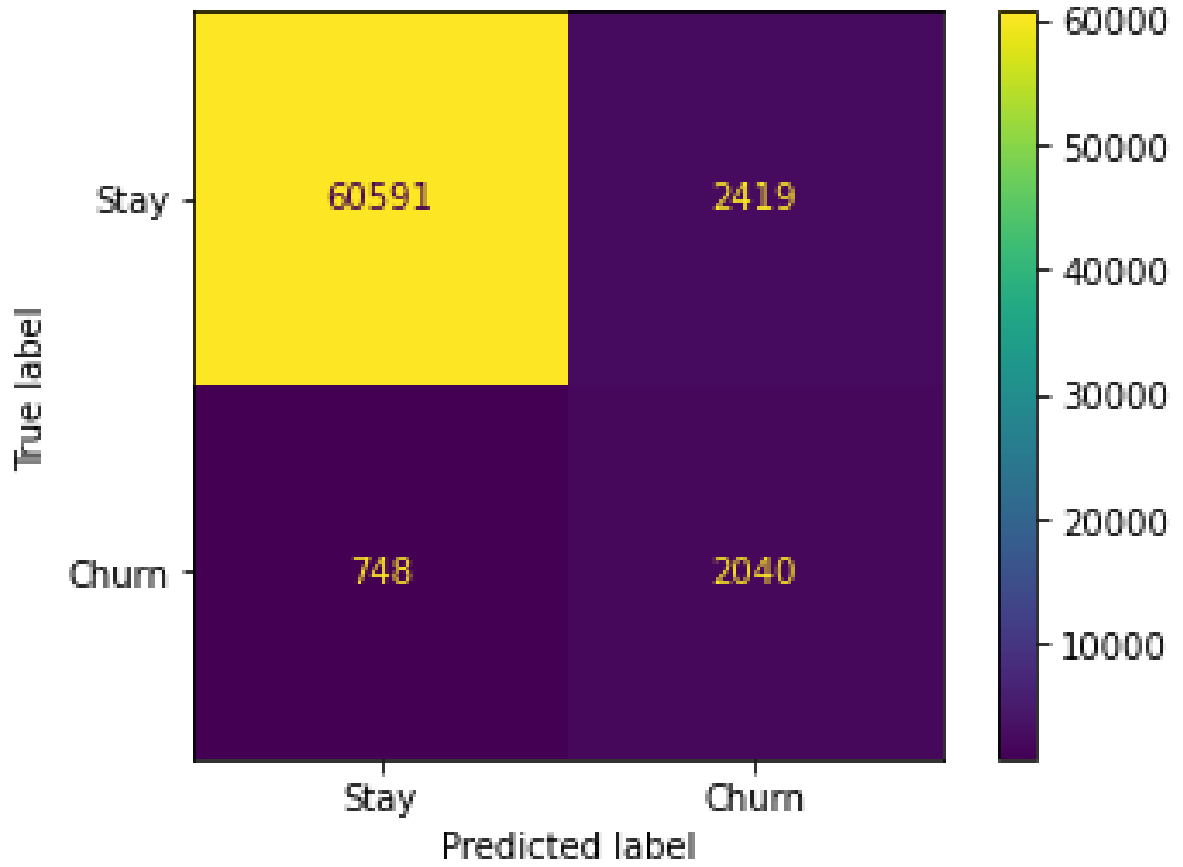
Hiperparametro	valor óptimo
Max_Depth	-1
num_leaves	20
learning_rate	0.2
n_estimators	700
reg_lambda	6
subsample	0.5
colsample_bytree	1
boosting_type	'goss'
scale_pos_weight	número de muestras que no hacen churn / número de muestras que si lo hacen (Kinnander, M. 2020)

Al revisar los resultados, se encontraron las siguientes áreas bajo la curva:

Tabla 16: Valor del área bajo la curva ROC para cada muestra del modelo "lgb" sin variables de tendencia

Muestra	AUC ROC
Train	0.98489
Validation	0.84666
Test	0.79595

Figura 21: Matriz de confusión: modelo "lgb" sin tendencia en la muestra de validación:



Analizando ambos resultados, vemos que el modelo entrenado con el algoritmo XGBoost tiene un menor poder de predicción en los tres conjuntos de datos. Adicionalmente el tiempo de cómputo para buscar los hiperparametros óptimos fue ampliamente mayor, demorando casi 24 horas en recorrer todas las posibilidades.

Por su parte, el modelo entrenado usando LightGBM mostró resultados prometedores en esta primera instancia, con un mejor poder de predicción y un tiempo de cómputo inferior (alrededor de las 8 horas).
Es por estas razones que se decidió avanzar con el modelo de Light GBM para seguir en la búsqueda del modelo óptimo que consiga predecir el churn.

4.2 - Análisis de resultados con variables de tendencia

Cuando se agregaron las variables de tendencia a la base de datos, dio como resultado la siguiente selección óptima de hiperparametros:

Tabla 17: Listado de hiperparametros y el valor encontrado en la optimización

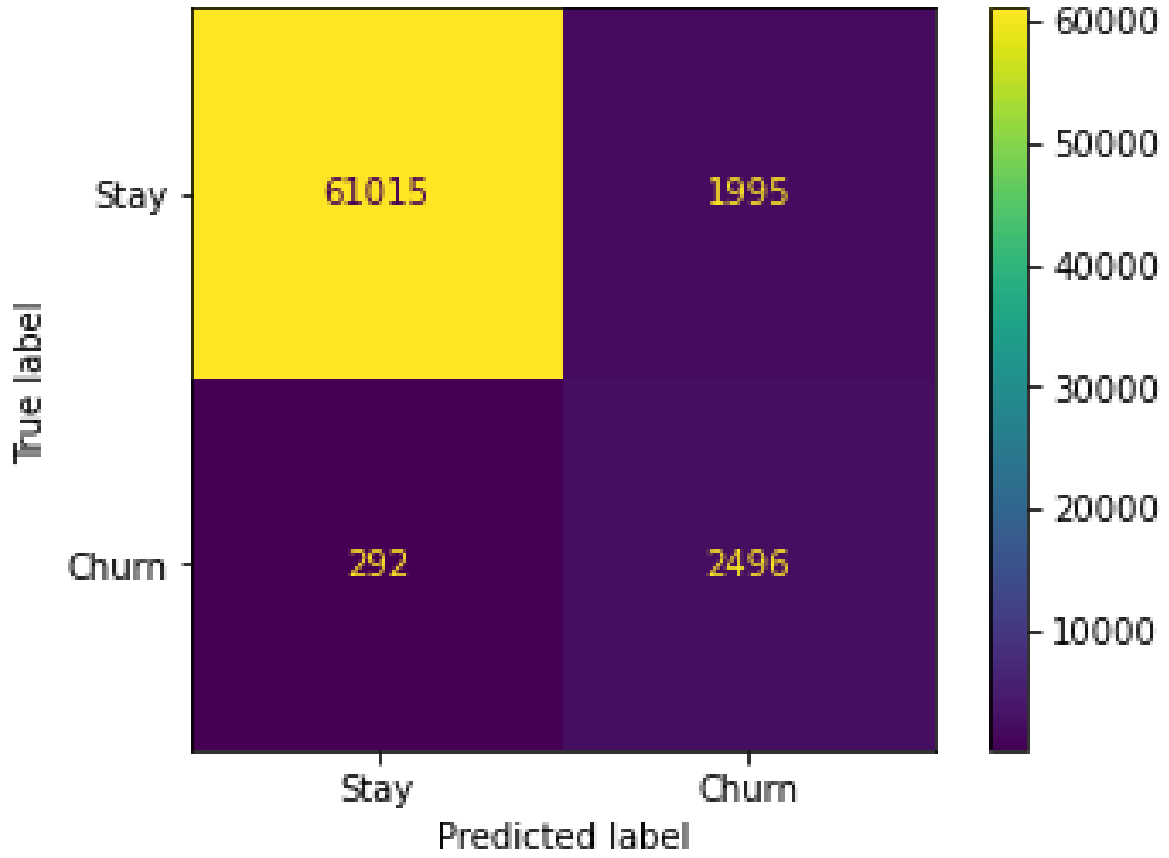
Hiperparametro	valor óptimo
Max_Depth	-1
num_leaves	10
learning_rate	0.1
n_estimators	500
reg_lambda	8
subsample	0.7
colsample_bytree	0.9
boosting_type	‘goss’
scale_pos_weight	número de muestras que no hacen churn / número de muestras que si lo hacen (Kinnander, M. 2020)

Al revisar los resultados, se encontraron las siguientes áreas bajo la curva:

Tabla 18: Valor del área bajo la curva ROC para cada muestra del modelo "lgb" con variables de tendencia

Muestra	AUC ROC
Train	0.9737
Validation	0.9318
Test	0.8735

Figura 22: Matriz de confusión: modelo "lgb" con tendencia en la muestra de validación:



Cuando comparamos ambos resultados, podemos ver que el modelo sin variables de tendencia tiene una AUC ROC más grande para la muestras de entrenamiento, pero no logra ejecutar a la altura en las muestras de validación y testeo, al estar haciendo overfitting. En cambio, el modelo con las variables de tendencia muestra valores consistentes en las muestras de validación y testeo, al tener un mayor poder predictivo.

Tabla 19: Resumen AUC ROC para cada muestra del modelo "lgb" sin y con las variables de tendencia

Muestra / AUC ROC	Sin variables de tendencia	Con variables de tendencia
Train	0.9849	0.9737
Validation	0.8467	0.9318
Test	0.7960	0.8735

4.3 - Performance en testeo

Hasta el momento se encontraron las variables más importantes para predecir, se crearon las variables de tendencia, y la combinación óptima de hiperparametros que ayudan a llegar a la mejor predicción posible.

Para pasar a la etapa de producción del algoritmo, donde se quiere predecir el churn del próximo mes, es necesario usar los hiperparametros optimizados con los datos de entrenamiento (noviembre) para entrenar un nuevo modelo con los datos de validación (diciembre), y luego ver la performance en testeo (enero). Al usar información más cercana al período a predecir, este ajuste nos asegura tener información más relevante a la hora de entrenar el modelo que luego se va a usar en el caso real a predecir.

Tabla 20: Valor del área bajo la curva ROC para cada muestra del modelo "lgb_test"

Muestra	AUC ROC
Validation	0.9862
Test	0.9199

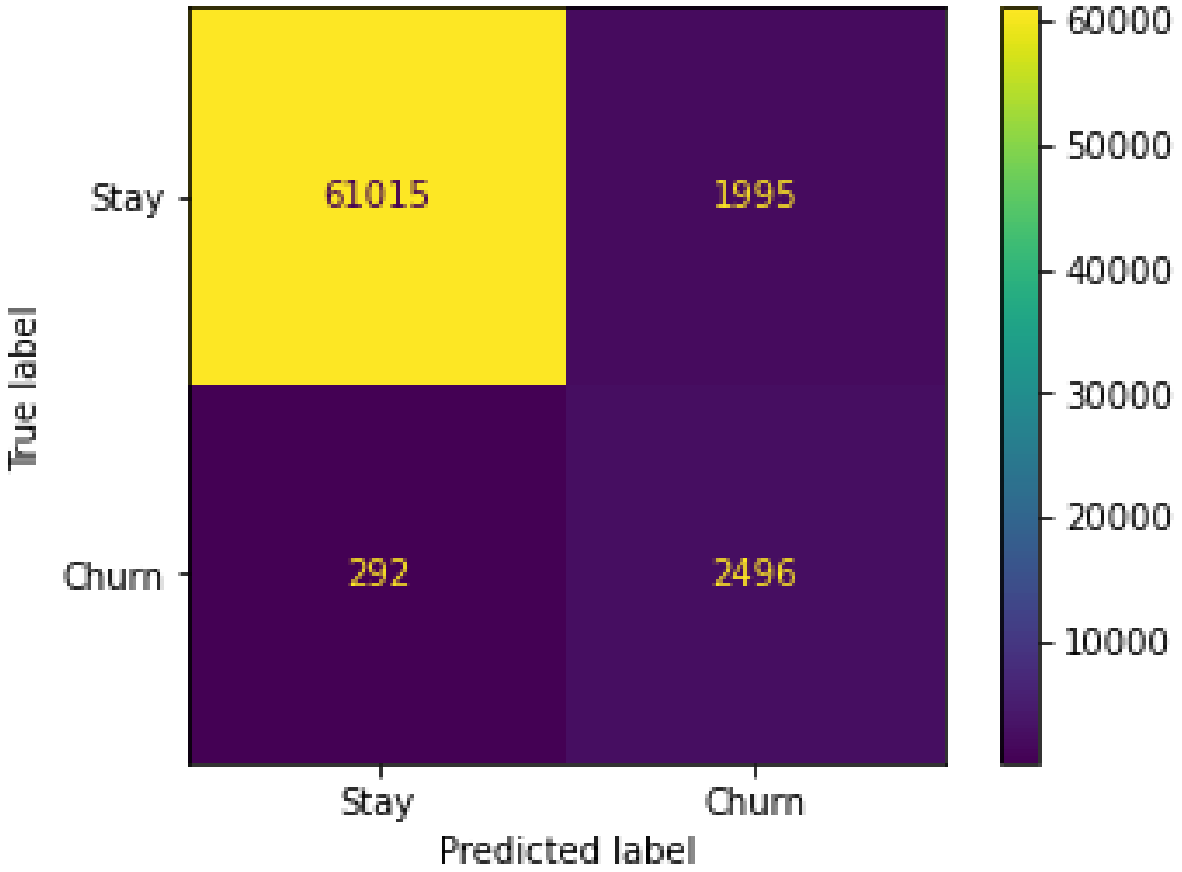
Tabla 21: Precision, recall & F-1score para el modelo "lgb_test" con la base de testeo:

	Precision	Recall	F-1 Score	Support
0	0.99	0.96	0.98	64.066
1	0.56	0.88	0.68	3.618

Tabla 22: Accuracy para el modelo "lgb_test" con la base de testeo:

	Precision	Recall	F-1 Score	Support
Accuracy			0.96	67.684
macro avg	0.78	0.92	0.83	67.684
weighted avg	0.97	0.96	0.96	67.684

Figura 23: Matriz de confusión: modelo "lgb_test" con tendencia en la muestra de testeo:

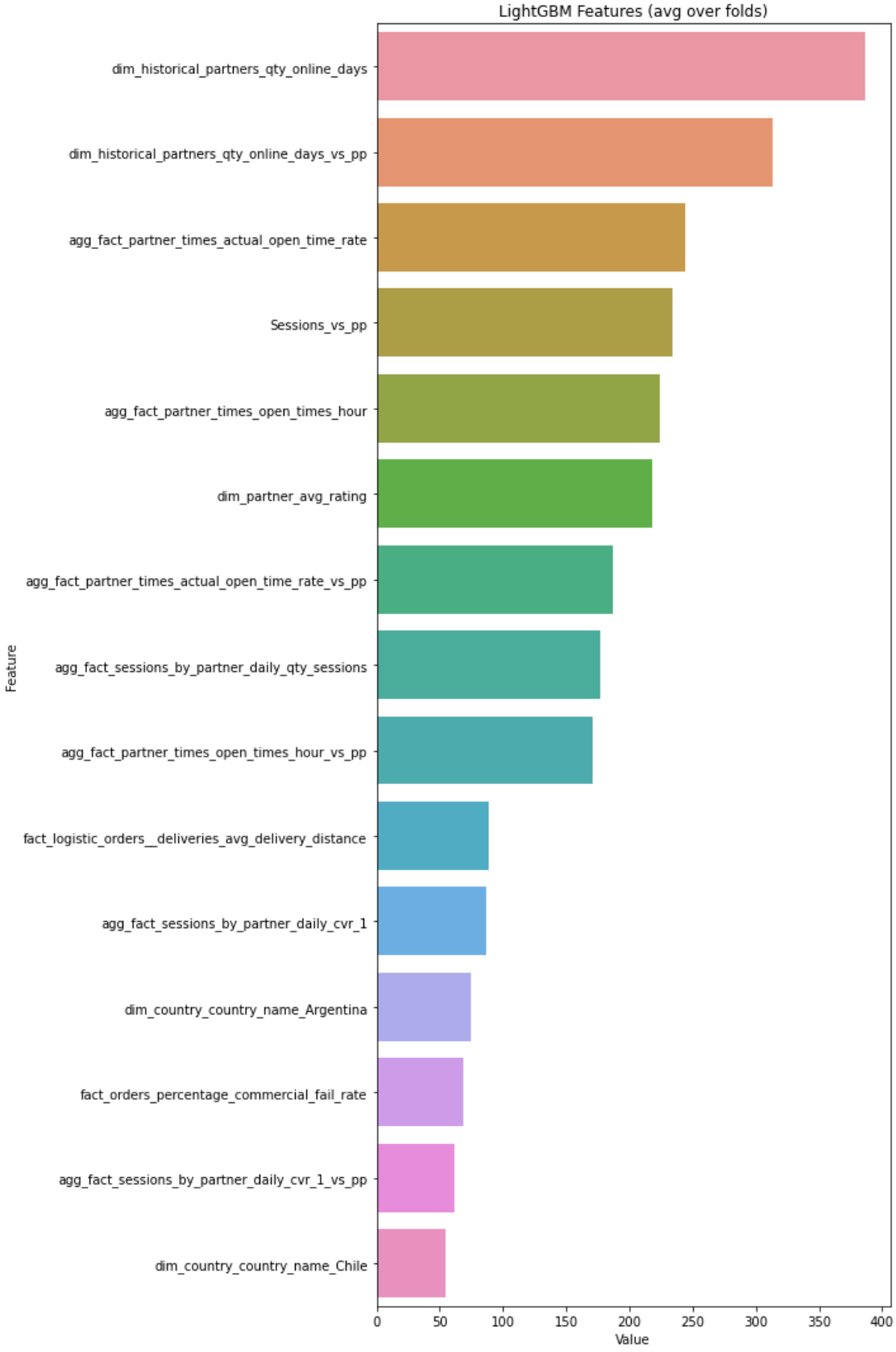


4.4 - Importancia de atributos:

En esta sección se intentará entender cuales son las variables que mejor explican el poder de predicción del modelo a hora de entender el churn de los restaurantes. Es importante interpretar la importancia que tienen las variables en el modelo para poder entender cuales son las acciones realizables para prevenir las bajas futuras.

En este caso vamos a usar la métrica de importancia de variables que está incluida en el paquete de LightGBM, que cuenta cuantas veces se utilizó la variable en el modelo.

Figura 24: gráfico de importancia de atributos para el modelo "lgb_test" con la muestra de testeo:

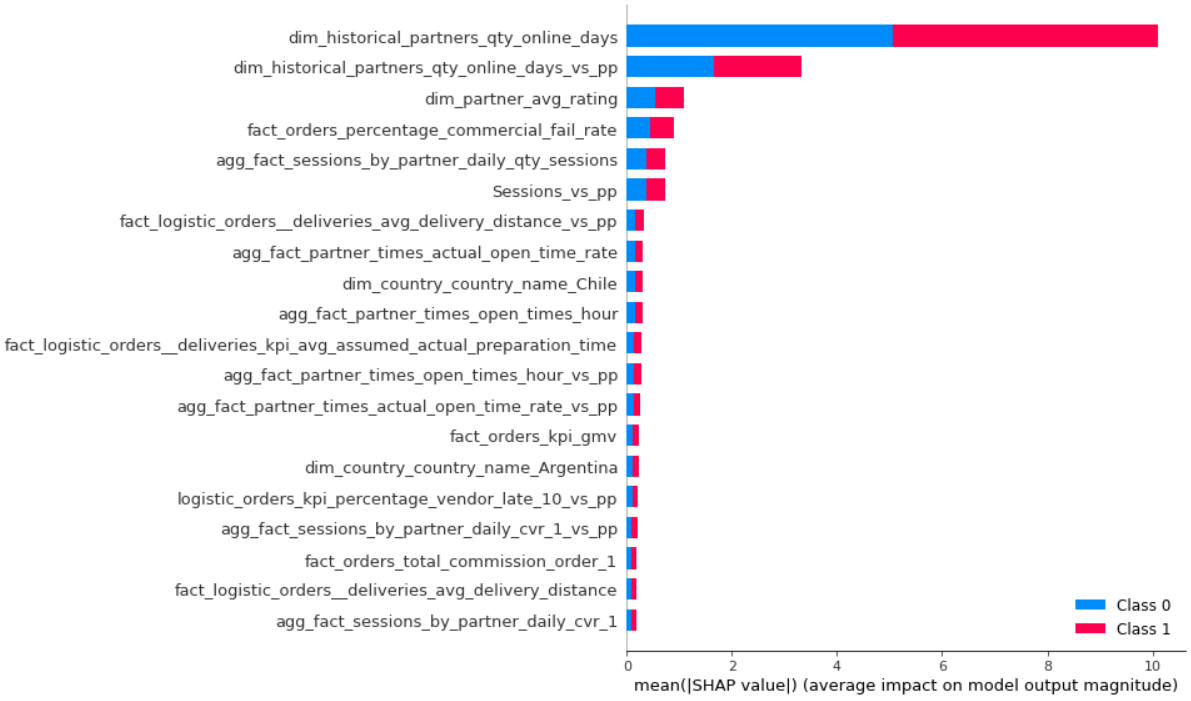


En este gráfico podemos ver que las dos variables de mayor importancia son la cantidad de días Online, y su variación contra el mes anterior. La tercera variable en el listado es el porcentaje de tiempo que está abierto el restaurante sobre la cantidad de tiempo que se comprometió a abrir en su contrato. Intuitivamente esto tiene sentido, dado que cuando un restaurante deja de prender su perfil en la aplicación de manera regular, significa que las ventas a través de la aplicación ya no le resulta importante para su negocio.
Luego tenemos la cantidad de sesiones en el perfil y su variación vs el período anterior, el rating en la aplicación, la distancia promedio de los pedidos, la conversión de sesiones y la tasa de cancelaciones del restaurante. Estas métricas están relacionadas con una mala operación del restaurante, donde la calidad de sus platos está en decadencia, aumentando sus probabilidades de darse de baja de la aplicación

4.5 - SHAP:

Otra manera más reciente y moderna de calcular la importancia de las variables es a través del método SHAP (Shapey Additive Explenations). SHAP asigna a cada variable
un valor de importancia para cada predicción particular. (Lundberg y Lee, 2017).
En el siguiente gráfico se muestran las variables usadas en órden de importancia, junto con sus valores SHAP para ambas clases

Figura 25: SHAP value por variable



En este gráfico podemos ver que se mantienen resultados muy similares a través de la metodología presente en el paquete LightGBM. Adicionalmente podemos ver que las variables tienen el mismo impacto en el modelo para predecir cuando hace churn (Clase 1) como cuando no lo hace (Clase 0).

5 - Discusión

5.1 - Principales resultados

El modelo final con los hiperparametros optimizados usando la base de noviembre, entrenado con la de diciembre y usando enero como test, tuvo un área bajo la curva ROC de 0.9862 en la muestra de validación M-1 (diciembre) y 0.9199 en la muestra final de testeo M-0 (enero).

5.2 - Aplicación en la industria

El objetivo de este trabajo fue poder elaborar una herramienta que le permita a la compañía poder atacar el problema del churn en restaurantes, un problema fundamental del negocio, de manera proactiva.

El output del modelo es un listado de los restaurantes con una variable booleana indicando si va a hacer churn (1) o si no lo va a hacer (0) en el siguiente mes. De esta manera se podrán direccionar acciones puntuales para los partners en riesgo de darse de baja para intentar contrarrestar la decisión. Para poder decidir qué acción tiene más impacto, es importante revisar el gráfico de los SHAP values para entender cuál es la variable que más afecta a la hora de hacer churn (ver figura 25).

En primera instancia, los días online y su variación contra el período anterior son las principales variables que afectan los restaurantes que hacen churn. Para contrarrestar esta tendencia la compañía podrá contactarse vía e-mail, WhatsApp, o a través de sus asesores comerciales con los restaurantes para notificarles que están abriendo pocos días al mes, e incentivar que se prendan los perfiles con más frecuencia ofreciendo descuentos de comisión para los restaurantes que puedan dar vuelta la tendencia.

Siguiendo en importancia viene el rating y el porcentaje de órdenes canceladas sobre el total. Para poder mejorar estas métricas se le puede ofrecer al partner capacitaciones culinarias desde el equipo de calidad para poder mejorar su operación, para poder entregar platos de mejor calidad, sin problemas operativos que puedan terminar en una cancelación.

En tercer lugar está la cantidad de sesiones y su variación contra el período anterior. En este caso la compañía puede subsidiar la compra de espacios publicitarios dentro de la app, que ubiquen los perfiles de los restaurantes en los primeros lugares del listado, o que le den visibilidad en la página principal de la aplicación. De esta manera el perfil de estos

restaurantes tendrá más tráfico y así generará más sesiones que puedan terminar en una venta.

Por último, una posible acción para generar más conversión es financiar descuentos de precios en productos claves pagos entre el partner y la compañía. Esto generará un mayor incentivo para los clientes al ver productos más atractivos. Así al tener más órdenes dependerá más del negocio online.

Estas estrategias para retener restaurantes tienen costos muy inferiores a la de adquirir nuevos operadores (Blattberg, R. C., & Deighton, J.,1996). Es por eso que es fundamental entender las razones por las que están haciendo churn, y poder optimizar el uso de recursos para disminuirlo con acciones puntuales direccionadas a los restaurantes con mayor riesgo de dejar de operar en la aplicación.

5.3 - Limitaciones y trabajos futuros

Uno de los principales desafíos en este trabajo fue el armado de la base de datos con información de distintas fuentes. La aplicación analiza y almacena grandes cantidades de datos de los restaurantes que operan en la misma desde hace más de 10 años.

La selección de las variables a utilizar en el modelo se hizo en base a la percepción de cuáles pueden ser las que ayuden a predecir el churn, pero por cuestiones prácticas hubo cientos de variables que no se tuvieron en consideración. En caso de haberse usado todas las variables, la base hubiese tenido un tamaño imposible de analizar en una computadora de uso cotidiano. Entonces es posible que hayan quedado variables que puedan aumentar el poder de predicción del modelo, que hayan quedado por fuera de este análisis.

Por otro lado, el modelo usó exclusivamente la información recabada por la aplicación, que pesa entre el 10% y el 30% de las ventas de un restaurante en promedio (Hirschberg, C., Rajko, A., Schumacher, T., & Wrulich, M., 2016). Esto quiere decir que hay muchísima información por fuera de la app que podría ser de vital importancia para analizar el churn. Esto puede ser cantidad de empleados, ubicación, posición de marca, balances financieros, datos macroeconómicos del país, etc. Este tipo de información si bien sería difícil de recopilar, ayudaría mucho a mejorar la predicción del modelo

Otra rama posible de explorar sería usar otro tipo de algoritmos para predecir el churn en series temporales. Si bien Light GBM es una herramienta moderna y eficiente para este desafío, también se podría usar algoritmos similares de árboles ensamblados como XGBoost o AdaBoost. Pero tal vez la vía de afrontar el problema de predicción de churn más interesante podría ser a través del uso de redes neuronales, que son computacionalmente más complejas y avanzadas.

Adicionalmente, se podría buscar la manera de crear un modelo que busque predecir las adquisiciones que nunca van a poder adaptarse a la plataforma, y que terminen haciendo churn al poco tiempo. Este modelo sería útil para poder filtrar potenciales restaurantes, ahorrando recursos y tiempo a la compañía para poder tener una cartera más eficiente de restaurantes. En la práctica dicho modelo parece difícil de implementar dado que el proceso de selección es de manera telefónica, y no hay manera de poder tener un conjunto de datos confiable sobre estos restaurantes que ayude a tomar la decisión.

Por último, cabe destacar que en este trabajo se utilizó el área bajo la curva ROC como función objetivo a optimizar. Si bien esta función es estándar en este tipo de algoritmos, es posible que no sea ideal para este caso de negocio. En una próxima iteración se podría trabajar utilizando como función objetivo a optimizar las órdenes o la facturación potencial de cara restaurante, para así lograr enfocarse en los mejores candidatos para operar en la aplicación.

5.4 - Conclusión

En este trabajo se buscó encontrar un método para poder entender de manera proactiva el conjunto de restaurantes que van a apagar sus perfiles en el mes siguiente.

Para ello, fue necesario llevar a cabo un análisis previo de los datos históricos de estos restaurantes para entender sus principales características y métricas operativas.

Luego, utilizando el algoritmo LightGBM, se efectuó la predicción de churn, entrenando un modelo con información del mes M-2 para elegir la mejor combinación de hiperparametros. Después se entrenó un nuevo modelo con información M-1 con esa combinación de hiperparametros, y se usó para predecir en el mes M-0, que es el último mes con información disponible.

Usando el área bajo la curva ROC como medida de performance, se encontró que el modelo tuvo un resultado de 0.99 en la muestra de validación M-1 y 0.92 en la muestra final de testeo M-0.

Estos resultados favorables tienen uso práctico concreto, dado que se pueden impulsar acciones puntuales de fidelización enfocada únicamente en estos restaurantes para bajar la tasa de churn. De esta manera se puede optimizar el gasto en retención de restaurantes, generando mayor eficiencia en el uso de los recursos y un mayor impacto en la operación para un problema que es fundamental en la industria.

6 - Bibliografía

1. Al Daoud, Essam. "Comparison between XGBoost, LightGBM and CatBoost using a home credit dataset." International Journal of Computer and Information Engineering
2. AL-Shatnwai, A. M., & Altibbi, M. F. (2020). Predicting Customer Retention using XGBoost and Balancing Methods. International Journal of Advanced Computer Science and Applications
3. Bhatti, W. A., Glowik, M., & Vahlne, J. E. (2020). Delivery Hero: Instant Internationalization Through Digitalization.
4. Blattberg, R. C., & Deighton, J. (1996). Manage marketing by the customer equity test. Harvard business review
5. Bommireddipalli,R. (2022). Council Post: What's Next For Q-Commerce: The Golden Child Of E-Commerce.
6. Davis J, Goadrich M. (2006). The Relationship Between Precision-Recall and ROC Curves. Proceedings of the 23rd International Conference on Machine Learning. Association for Computing Machinery
7. Davis J, Goadrich M. (2006). The Relationship Between Precision-Recall and ROC Curves. Proceedings of the 23rd International Conference on Machine Learning. Association for Computing Machinery.
8. de la Llave Montiel, M. A., & López, F. (2020). Spatial models for online retail churn: Evidence from an online grocery delivery service in Madrid. Papers in Regional Science
9. Dunn, E. (2018). How delivery apps may put your favorite restaurant out of business. The New Yorker
10. Feldman, P., Frazelle, A. E., & Swinney, R. (2019). Can Delivery Platforms Benefit Restaurants?. Available at SSRN 3258739.
11. Gupta, M. (2019). A Study on Impact of Online Food delivery app on Restaurant Business special reference to zomato and swiggy. International Journal of Research and Analytical Reviews

12. Hancock, J., & Khoshgoftaar, T. M. (2021, August). Leveraging LightGBM for Categorical Big Data. In 2021 IEEE Seventh International Conference on Big Data Computing Service and Applications (BigDataService). IEEE.
13. Hirschberg, C., Rajko, A., Schumacher, T., & Wrulich, M. (2016). The changing market for food delivery.
14. James G, Witten D, Hastie T, Tibshirani R.(2017). An Introduction to Statistical Learning. 8 edición. Springer.
15. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30.
16. Kinnander, M. (2020). Predicting profitability of new customers using gradient boosting tree models: Evaluating the predictive capabilities of the XGBoost, LightGBM and CatBoost algorithms.
17. Kuhn M, Johnson K. (2016). Applied Predictive Modeling. 5 edición. Springer.
18. Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems
19. Marzban, C. (2004). The ROC curve and the area under it as performance measures. Weather and Forecasting.
20. Microsoft Corporation (2022), LightGBM Release 3.3.2.99 Documentation. Redthedocs
21. Moreno Padilla, G. A. (2021). Plan estratégico de marketing: Caso Delivery Hero (Bachelor's thesis, PUCE-Quito).
22. Shah,A. (2021). Q-Commerce: Fulfilling Instant Wishes, by Amirul Shah, Foodpanda. Osome.
23. Shi, H. (2007). Best-first decision tree learning (Doctoral dissertation, The University of Waikato).
24. Tan P, Steinbac M, Kumar V (2006). Introduction to Data Mining. Pearson.
25. Warmer, C., & Weber, S. (2014). Delivery Hero. In Mission: Startup, Springer Gabler, Wiesbaden.
26. Zou, Q., Xie, S., Lin, Z., Wu, M., & Ju, Y. (2016). Finding the best classification threshold in imbalanced classification. Big Data Research