ESCOLA TÉCNICA SUPERIOR
DE ENXEÑARÍA

Trabajo Fin de Máster

Desarrollo de un modelo de predicción de cancelaciones de reservas en el sector turístico

Noel Fabello Quintana

Tutores
Eduardo Manuel Sánchez Vila
Alejandra Comesaña García
Ignacio Freire Martínez

1 de julio de 2024

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Máster Interuniversitario en Tecnologías de Análisis de Datos Masivos: Big Data

Trabajo Fin de Máster

Desarrollo de un modelo de predicción de cancelaciones de reservas en el sector turístico

Noel Fabello Quintana

1 de julio de 2024

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Resumen

En la industria hotelera, la incertidumbre producida por las cancelaciones de reservas dificulta la previsión de disponibilidad en el establecimiento, lo cual produce pérdidas significativas de beneficios. En este proyecto se lleva a cabo un estudio profundo sobre los patrones de anulación de reservas para Eurostars Hotel Company, utilizando métodos avanzados de *Big Data* para limpiar, transformar y analizar grandes volúmenes de datos históricos. A partir del conocimiento obtenido, se desarrolla un conjunto de modelos de detección de cancelaciones y se discute su rendimiento atendiendo tanto a cuestiones técnicas como de negocio. Una vez seleccionado el mejor modelo, se compara con un clasificador implementado por una empresa externa. Finalmente, se diseña una herramienta visual que facilita al personal del sector el análisis de las predicciones, permitiendo la elaboración de estrategias efectivas para mitigar el impacto de las cancelaciones y la estimación del beneficio aportado por la aplicación de las mismas.

Confidencialidad de datos y resultados mostrados

Dada la confidencialidad e importancia comercial de los datos tratados y los resultados obtenidos en este trabajo, se deben anonimizar u ocultar los datos relativos a clientes, hoteles y reservas con el objetivo de proteger los intereses comerciales de Eurostars Hotel Company. Por este motivo, algunas gráficas, por ejemplo, no contendrán valores en los ejes. En cambio, esta práctica no afecta a las conclusiones que se pueden extraer de los resultados mostrados.

Índice

Capítulo 1

Introducción

La industria hotelera es un pilar fundamental de la economía española, representando una parte significativa del Producto Interior Bruto (PIB) y proporcionando empleo a millones de personas. En 2023, tras el impacto de la pandemia, España recibió cerca de 85 millones de turistas internacionales, consolidándose como uno de los destinos más populares a nivel mundial. La contribución del turismo al PIB se situó en torno al 12,8 %, destacando la relevancia de este sector en la economía nacional [1]. España también figura como el segundo país con mayor inversión en esta industria, superando los 4.000 millones de euros. El sector cuenta con más de 13.000 establecimientos y ofrece alrededor de 670.000 habitaciones, alcanzando una facturación total de 18.600 millones de euros en 2023 [2].

Sin embargo, uno de los principales retos que enfrentan los hoteles es la gestión de cancelaciones de reservas. Las cancelaciones representan una fuente considerable de pérdidas para los establecimientos, no solo debido a los ingresos de las reservas anuladas, sino también por la dificultad de revender la habitación liberada con un período reducido de tiempo. Se estima que las cancelaciones suponen un 20 % de las reservas, que asciende incluso a un 35 % en las reservas realizadas a través de OTAs (*Online Travel Agency*), lo que implica una disminución de la facturación significativa para una cadena hotelera [3].

Además, las políticas de cancelación han sufrido cambios importantes en los últimos años, tendiendo a flexibilizarse a favor del cliente [4]. La pandemia de *COVID-19* ha sido un factor determinante en este cambio, ya que muchos hoteles adoptaron condiciones más permisivas para ofrecer a los turistas seguridad a la hora de realizar sus planes de viaje. Estas políticas incluyen la posibilidad de cancelar sin costo hasta 24 horas antes de la llegada, la opción de modificar las fechas de la reserva sin penalización y la eliminación de tarifas no reembolsables [5]. Por otro lado, la aparición de plataformas que ofrecen habitaciones y apartamentos como Airbnb ha aumentado la competencia en este sector. Esta competencia ha presionado a los hoteles a mejorar sus servicios y a ofrecer condiciones más atractivas y menos restrictivas para los clientes.

En la actualidad, el auge de la aplicación del *Big Data* en el ámbito empresarial ha dado lugar a la ciencia de datos, una disciplina que utiliza métodos, procesos, algoritmos y sistemas para extraer conocimiento a partir de los datos. Este análisis posibilita la toma de decisiones informadas, estratégicas y óptimas a nivel de recursos, lo cual ofrece un alto valor a la entidad. En los últimos 10 años, esta disciplina ha demostrado ofrecer altos beneficios a las empresas, incrementando su eficiencia operativa, mejorando la personalización de servicios, optimizando campañas de marketing y permitiendo la identificación de nuevas oportunidades de negocio. La capacidad de predecir comportamientos del cliente, gestionar riesgos de manera más efectiva y

aumentar la rentabilidad son solo algunas de las ventajas que han convertido a la ciencia de datos en un recurso indispensable para las organizaciones modernas [6].

En concreto, en la industria hotelera se ha aplicado la ciencia de datos para implementar estrategias de sobreventa de habitaciones y evitar anulaciones inesperadas [7]. En el presente, la mayoría de soluciones de este tipo se basan en distribuciones matemáticas, estudios de tendencias y tasas históricas de cancelación [8]. Los beneficios aportados por dichas estrategias son abundantes, aunque dependen del nivel de reventa aplicado y no se realiza el análisis individual de las características de la reserva [9].

Considerando la problemática mencionada y el contexto actual, Eurostars Hotel Company, la mayor cadena hotelera española en cuanto a número de establecimientos, con más de 250 hoteles en 18 países, propone la realización de un análisis del registro histórico de datos sobre cancelaciones en la compañía. Este estudio tiene como propósito obtener información útil que aporte valor a la cadena, permitiéndole establecer cambios en la estrategia de venta de habitaciones y en la creación de políticas de cancelación adecuadas.

1.1. Objetivos del proyecto

En este proyecto se han establecido diversos objetivos con el fin de aportar valor a la empresa y atajar el problema presentado. Estos objetivos están diseñados para mejorar la eficiencia operativa y la toma de decisiones en Eurostars, mediante el análisis y el uso efectivo de los datos.

- **Identificación de patrones de cancelación:** El estudio de los datos debe permitir obtener información valiosa acerca del comportamiento de los clientes en la cancelación de reservas, lo que ayudará a identificar las causas más comunes y a desarrollar estrategias efectivas para mitigarlas.
- **Desarrollo de un modelo de predicción de cancelaciones:** La creación de un clasificador de reservas que ayude a Eurostars en tareas como la previsión de facturación y la ocupación disponible en el hotel, lo que permitirá a la empresa tomar decisiones más informadas y mejorar su capacidad para adaptarse a los cambios en el mercado.
- **Creación de una herramienta visual:** Se diseñará una herramienta que permita al personal de negocio analizar de forma simple y eficaz las tendencias de cancelación de reservas. Este desarrollo puede ser de gran utilidad a la hora de detectar oportunidades de mejora e implementar estrategias efectivas.

Capítulo 2

Desarrollo general

Durante este capítulo se describirán las tres fases iniciales del proyecto, fundamentales para establecer una base de conocimiento que permitirá avanzar con las tareas de los siguientes capítulos. La primera fase consiste en el entendimiento del negocio, un paso crucial para comprender la utilidad e intención de los objetivos. La segunda fase se centra en la comprensión de los datos, explorando la base de datos y estudiando su estructura y contenido para extraer información útil para la creación del modelo de predicción. Finalmente, la tercera fase abarca la limpieza y transformación de los datos, lo que facilita la aplicación de técnicas de *machine learning* (ML).

2.1. Entendimiento del negocio

Todo proyecto relativo a la analítica de datos y a la aplicación de modelos de *machine learning* debe incluir una fase de comprensión de las características y peculiaridades del negocio. Sin este paso, las acciones y conclusiones obtenidas pueden no guardar una relación con los objetivos o carecer de sentido debido a la falta de contexto. Para comenzar a comprender estas características, se debe tener claro el transcurso del cliente por el hotel y cómo y cuándo se efectúa el registro de la información.

Realización de la reserva

El primer paso dentro del ciclo de vida de un cliente en un hotel es la selección del hotel y la reserva de habitación. Este proceso puede ser completado a través de diferentes métodos:

- Agencias de viajes online (OTAs): Páginas web y aplicaciones donde el cliente puede consultar la disponibilidad y los precios de los diferentes hoteles en el destino y fecha deseada. Es el método más común de reserva hoy en día.
- Web propia: Página web oficial de la compañía donde se puede obtener información sobre las habitaciones de los hoteles y crear la reserva directamente.
- Agencia de viajes: Empresas dedicadas a programar viajes, que pueden ofertar alojamiento además del desplazamiento al cliente.
- Sistema de distribución global (GDS): Software que actúa como intermediario entre los hoteles y las agencias de viajes. Permite comprobar la disponibilidad y los precios de las habitaciones de manera cómoda para poder realizar una reserva a nombre del cliente final.
- Operadores turísticos: Entidades que se relacionan con los hoteles para crear paquetes de viajes que contienen el desplazamiento, el alojamiento y actividades u otros servicios.

- Walk-in: Situación en la que el cliente se presenta en el hotel sin reserva previa solicitando una habitación.

Una vez efectuada la reserva, se almacenan los datos referentes a ella en la base de datos. En este momento solo se conoce del cliente su nombre y las características propias de la reserva, como pueden ser la fecha de llegada, el tipo de habitación o el régimen.

Llegada al hotel

A la llegada al hotel, se solicitan los datos personales de cada huésped, se le informa del número de habitación asignado y se hace entrega de la llave de acceso, constituyendo así el procedimiento conocido en el sector como *check-in*. En ese instante, se actualiza el registro almacenando la nueva información. Estos datos, tales como el documento de identidad o el correo electrónico, son esenciales para que la empresa cumpla con la normativa vigente, además de contribuir a la mejora de la experiencia del cliente y permitir la creación de un histórico de actividad del huésped.

Salida del hotel

A este último paso se le conoce como *check-out*, y se da en el final de la estancia del cliente. En este momento, se realiza la devolución de la llave de acceso y se entrega la factura con los costes de la estancia (si no han sido pagados antes) y cualquier cargo adicional, se completa el pago y se confirma la salida del huésped.

Una vez completada la observación de cómo y cuándo se registran los datos, se puede comprender que, para cumplir los objetivos de predicción de cancelación, se pueden utilizar aquellos relativos a la propia reserva y que se obtienen en la primera de las fases, ya que la información restante se consigue una vez el huésped está en el hotel y, por tanto, no ha cancelado su estancia.

Otro factor a analizar al realizar la predicción es el impacto de una clasificación errónea. El objetivo de etiquetar las reservas futuras como canceladas es prever con mayor precisión la disponibilidad de habitaciones en una fecha específica y, por lo tanto, tener la capacidad de revender una habitación que se encuentra reservada, asumiendo que será cancelada. Esta práctica se denomina en el sector como *overbooking*. Si esta clasificación es errónea, y se revenden demasiadas habitaciones, se produce una sobreventa u *overselling*. Esta situación ocurre cuando se venden más habitaciones de las disponibles en un hotel y no se producen suficientes cancelaciones como para alojar a todos los clientes. Si esto sucede, el hotel debe compensar al cliente con alguna de las siguientes opciones: (1) reubicación en un hotel cercano de igual o mayor categoría (número de estrellas), asumiendo los costos de transporte si es necesario; (2) compensación económica equivalente al total o a una parte de la reserva; (3) compensación en forma de servicios adicionales o aplazamiento de la fecha de la reserva; o (4) bonos de descuento o incremento de los puntos de lealtad en el programa de fidelización de la compañía. En cambio, si la predicción refleja menos cancelaciones de las reales, se produce una pérdida de oportunidad de beneficio. Este conocimiento debe ser tomado en cuenta y será utilizado a la hora de fijar la tolerancia de error al modelo y el cálculo de beneficios.

2.2. Comprensión de los datos

Una vez entendidos los factores de negocio que se deben analizar, se lleva a cabo un estudio de la información disponible en la base de datos. La intención final de este estudio es la identi-

ficación de qué datos serán más útiles para alcanzar el objetivo principal del proyecto.

Como ya se ha comentado, existe un histórico almacenado en la base de datos de la compañía relativo a las reservas realizadas. En él, se incluye un identificador de reserva, la fecha de generación de la misma, las fechas de *check-in* y *check-out*, así como características concretas que se explicarán en esta sección. Además, para ejecutar un estudio más detallado y un mejor análisis, se han explorado otras tablas de la base de datos que permiten obtener contexto sobre el momento y el motivo por el que se realizó una reserva. Algunas de estas características son los precios de los competidores, la evolución de la ocupación de los hoteles y la evolución de los precios de las habitaciones.

La base de datos está alojada en PostgreSQL y será la fuente de información principal a lo largo de todo el proyecto. Debido a la gran cantidad de datos almacenados por la compañía (250 hoteles y casi diez años de registro de reservas), se ha realizado una selección de 16 hoteles y se ha decidido analizar las reservas correspondientes a los últimos tres años, obteniendo así una muestra de alrededor de medio millón de registros. Este intervalo de tiempo se ha creado con la intención de evitar registros afectados por las políticas existentes durante la pandemia. La cantidad de datos resultante permite una fluidez mayor a la hora de ejecutar operaciones y mantiene un tamaño considerable sobre el que aplicar técnicas de *machine learning*.

Para la exploración de la base de datos se han ejecutado consultas SQL y sus resultados se han exportado como archivos CSV. Posteriormente, para la realización de tareas analíticas sobre estas extracciones, se ha utilizado el lenguaje de programación R.

A continuación, se resumirán las variables extraídas de la base de datos que se utilizarán para el estudio y el modelado del problema.

- Características relativas al hotel: El identificador del hotel, el tipo de hotel según su principal fuente de reservas (turismo, empresa, eventos, etc.), las estrellas o la capacidad.
- Variables temporales: La fecha de reserva, la de *check-in* y *check-out* y la de la última modificación del registro.
- Atributos propios de la reserva: Por ejemplo, el origen, el régimen, el tipo de habitación, el número de huéspedes y si la reserva fue o no cancelada.
- Variables relativas al precio: Como el precio de la reserva, el descuento aplicado y el precio de los hoteles cercanos.
- Variables de contexto: Entre ellas están el porcentaje de ocupación y el nombre y prioridad de los eventos y festividades cercanos al hotel que coinciden con la estancia.

Durante el estudio de las características presentadas, se ha visto la oportunidad de crear variables derivadas que pueden ofrecer información útil a la hora de comprender los datos y obtener patrones. Algunas de estas son las siguientes:

- Antelación: Calculada como los días que transcurren entre la fecha de creación de la reserva y el *check-in*.
- Hora de reserva: Hora a la que se efectuó la reserva.
- Días de semana y fin de semana: Variables que hacen referencia a cuántos días de la estancia coinciden con días laborales y cuántos en fin de semana.

- Coincidencia de festivo o evento: Representa la existencia de festivos o eventos que afecten al hotel durante la estancia.
- Diferencia de precio y ocupación: Indican la diferencia porcentual del precio del hotel con respecto a la competencia y la diferencia de precio u ocupación de las habitaciones entre el día de la reserva y el día de *check-in*

A partir de las variables presentadas, se han obtenido diversas conclusiones que resultarán útiles para la creación del modelo de predicción. Este conocimiento facilita un mejor preprocesamiento de los datos y una selección de características favorable para el rendimiento del modelo. Seguidamente, se detallan algunos de estos resultados.

Variables sesgadas

A raíz del estudio se han detectado variables que se sobrescriben una vez conocido el estado final de la reserva, por lo que no son válidas a la hora de crear un modelo, ya que incluyen información que no estará disponible a la hora de realizar predicciones sobre datos desconocidos. Ejemplos de estas variables son el beneficio bruto y la fecha de modificación del registro, ya que la primera contiene valores negativos en algunos hoteles si la reserva no se completó y la segunda es anterior a la fecha de *check-in* en el caso de que se haya llevado a cabo una cancelación.

Evolución de las cancelaciones

El porcentaje de cancelaciones se mantiene relativamente constante durante el año respecto a la cantidad de reservas y no sufre variaciones entre épocas de alta y baja demanda. Como se puede observar en la Figura 2.1, solo se aprecian ligeros picos en diciembre de 2022 y 2023 y julio de 2022, pero estos suponen un impacto menor al 2 % en el ratio de cancelaciones.

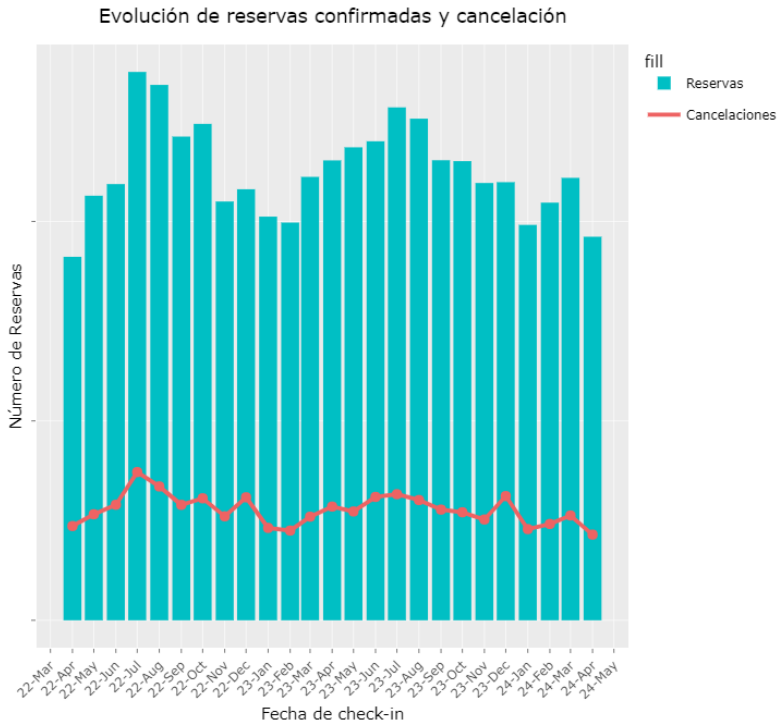

Figura 2.1: Evolución de la cantidad de reservas y cancelaciones.

Antelación

Un factor importante en la cancelación es la antelación, es decir, los días que ocurren entre la fecha de reserva y la fecha de *check-in*. En la Figura 2.2, se puede notar que existe una alta correlación entre ambos factores. Este comportamiento puede deberse a que la cantidad de días entre la reserva y la estancia aumenta la posibilidad de aparición de imprevistos u otros compromisos. En la gráfica también se muestra el porcentaje de reservas que ocupa cada división, en la que destaca el hecho de que casi un 45 % de las reservas se realizan con menos de una semana de antelación.

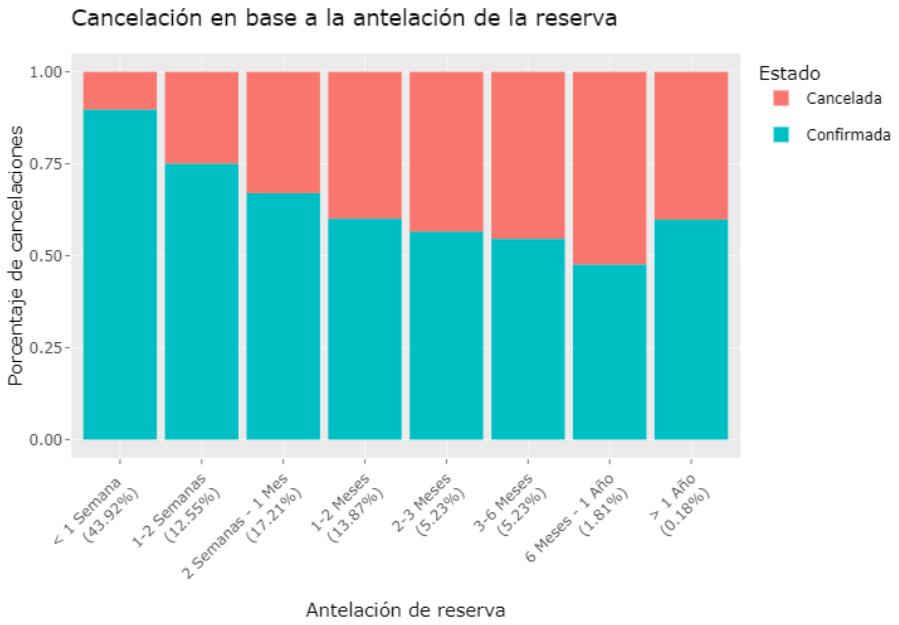

Figura 2.2: Proporción de cancelaciones en base a la antelación de la reserva.

Descuento

Existen dos tipos de descuento para una reserva, el que es ofertado por la propia cadena hotelera y el que realizan las agencias u OTAs a través de sus plataformas. Tras el estudio del comportamiento de las cancelaciones, se ha identificado que la existencia o no de una reducción en el precio es mucho más importante que la cantidad de esta. Por este motivo se ha creado una variable que indica si la reserva disfruta de una rebaja en su precio. En la Figura 2.3 se aprecia un ratio de cancelaciones un 26 % menor cuando el cliente ha aplicado un código de descuento en su reserva.

Figura 2.3: Proporción de cancelaciones en base a la existencia de descuento en la reserva.

Duración de la estancia

La duración de la estancia también es un factor importante a analizar, y como se puede observar en la Figura 2.4, la correlación entre la proporción de cancelaciones y el número de noches en el hotel es evidente. Presumiblemente, esto se debe a que cuanto más larga sea la estancia en el hotel, mayor es la probabilidad de que al cliente le surja algún motivo que lo lleve a cancelar la reserva.

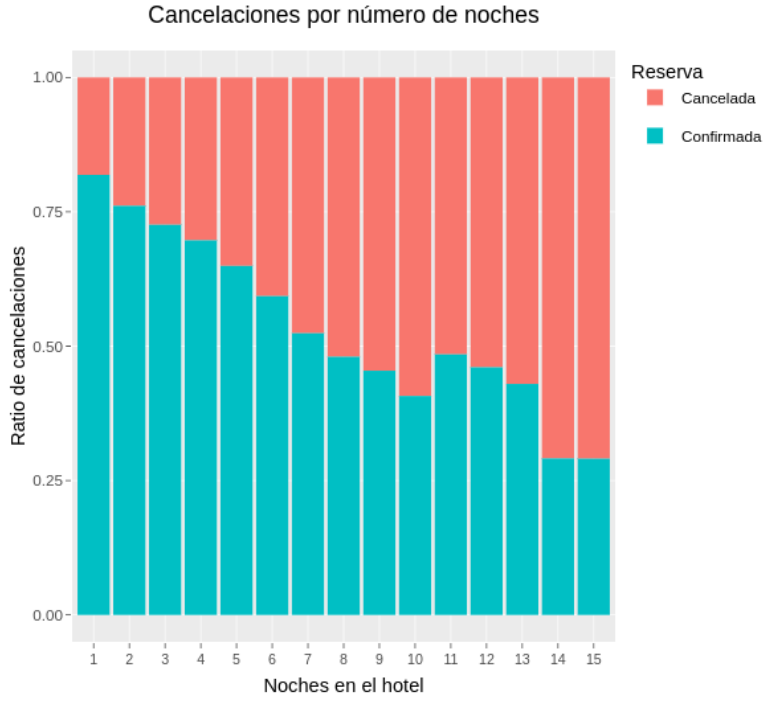

Figura 2.4: Proporción de cancelaciones en base al número de noches.

Hora de reserva

Debido a la naturaleza de la variable referente a la fecha de reserva, almacenada en la base de datos como *timestamp*, se puede analizar la posible correlación entre la hora en la que el cliente realiza la reserva y la frecuencia de cancelación. Esta variable podría parecer irrelevante en un primer instante pero, tras su estudio, se ha observado que hay una proporción de cancelaciones mayor cuando se realiza entre las dos y las siete de la madrugada, como se puede observar en la Figura 2.5. Durante el día, la hora no parece afectar significativamente a la proporción de cancelaciones.

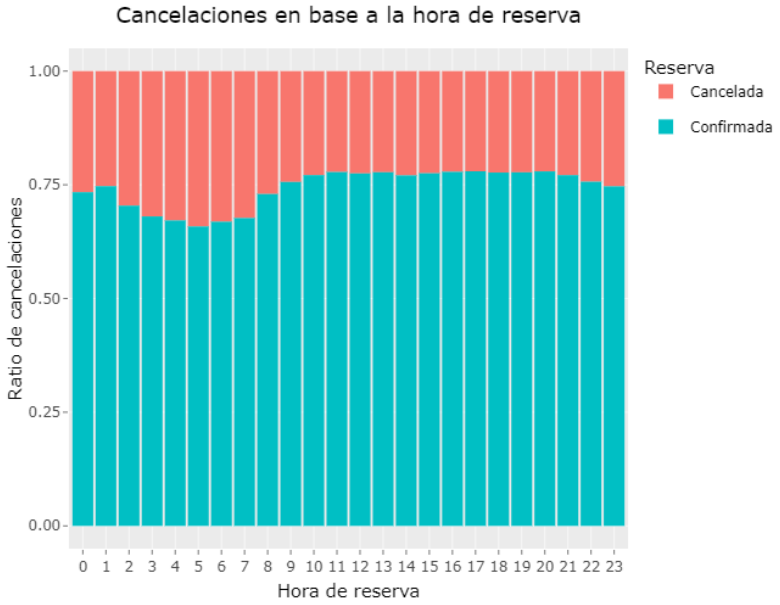

Figura 2.5: Proporción de cancelaciones en base a la hora de reserva.

El análisis de las variables presentó conclusiones interesantes para los capítulos posteriores. Se detectaron patrones y comportamientos que serán claves a la hora de seleccionar las características con las que se construirá el modelo, además de descartar las variables sesgadas debido a la actualización de los datos. Este estudio también ha permitido encontrar valores atípicos y estudiar las distribuciones de los datos, lo que será de gran ayuda en la limpieza y transformación de datos.

2.3. Preparación de los datos

En esta etapa, se realiza una limpieza y se acondicionan los datos para poder realizar con éxito el modelado en el siguiente capítulo. Esta ha sido una etapa transversal a todo el proyecto, ya que durante la realización del mismo se han añadido nuevas variables y se han comparado resultados con diferentes preprocesamientos. La intención de este tratamiento es poder aportar al modelo los datos de forma que pueda extraer la mayor información posible y, para ello, se debe estudiar y comprender el formato en el que el clasificador necesita que se reciban y cómo obtiene patrones de ellos.

Para la fase de limpieza se ha estudiado la distribución de cada variable y el significado de los valores faltantes que existen en ellas. En relación con las distribuciones, se han implementado diversas medidas. Por un lado, se han eliminado aquellas características que presentan una

variabilidad muy baja debido a que no aportan información. Un ejemplo de este escenario son las habitaciones vendidas, ya que más del 98 % de las reservas incluyen una sola habitación. Por otro lado, en el caso de las variables con numerosos valores atípicos, se ha aplicado la regla del rango intercuartílico al conjunto de entrenamiento.

Regla del rango intercuartílico

La regla del rango intercuartílico es una técnica utilizada para identificar y manejar valores atípicos en un conjunto de datos. Esta técnica se basa en el cálculo del rango intercuartílico (IQR, por sus siglas en inglés), es la diferencia entre el tercer cuartil (Q3) y el primer cuartil (Q1).

Para aplicar esta regla, se siguen los siguientes pasos:

1. Calcular el primer cuartil (Q1), que es el valor por debajo del cual se encuentra el 25 % de los datos.
2. Calcular el tercer cuartil (Q3), que es el valor por debajo del cual se encuentra el 75 % de los datos.
3. Determinar el rango intercuartílico (IQR) restando ambos valores:

$$\text{IQR} = Q3 - Q1$$

4. Identificar los límites inferior y superior para detectar valores atípicos:

$$\text{Límite inferior} = Q1 - 1{,}5 \times \text{IQR}$$

$$\text{Límite superior} = Q3 + 1{,}5 \times \text{IQR}$$

5. Eliminar cualquier valor que se encuentre por debajo del límite inferior o por encima del límite superior, ya que se considera atípico.

Al eliminar o tratar los valores atípicos identificados mediante esta regla, se puede reducir la influencia de estos en el modelo, mejorando así su estabilidad y robustez frente a nuevos datos [10].

Las variables con presencia de valores ausentes encontradas en la base de datos se encuentran en alguno de los siguientes escenarios: (1) variables con datos que no se han podido obtener y por lo tanto no se conocen, (2) variables con nulos, cadenas vacías o códigos como indicadores de ausencia de información. A continuación se detallarán las medidas tomadas para cada caso.

(1) Valores que no se han podido obtener: En ocasiones, datos como los precios de los hoteles vecinos no están disponibles; en otras, se guardan valores erróneamente debido a fallos en el sistema, como el tipo de régimen del cliente o el idioma en el que se hizo la reserva. Para reducir o eliminar los valores nulos se han realizado diferentes acciones dependiendo de la naturaleza de la variable. Para las continuas, como el precio, registradas periódicamente, se ha realizado una interpolación temporal. Para las categóricas, dado que no se han encontrado variables de con un alto porcentaje de nulos, se han sustituido estos valores por la moda.

(2) Valores por defecto para indicar ausencia: Existen características donde la falta de valor tiene un significado propio. Por ejemplo, el nombre de la festividad debe estar vacío si

no existe ninguna que coincida con la estancia del cliente. En este caso no se imputa ningún valor, simplemente se crea una nueva variable lógica indicando si existe o no festividad en esa fecha.

Además, cada variable ha sido estudiada de manera independiente prestando atención al formato en el que está almacenada. La naturaleza de algunas características requiere que se realice alguna transformación de forma que se obtengan valores en un formato cómodo y manejable. Este proceso se puede ejemplificar con los códigos de descuento. Cuando un cliente usa varios descuentos a la hora de realizar una reserva, los códigos quedan almacenados como *código1/código2/códigoN* y los porcentajes asociados a cada uno de ellos, en el mismo orden, como *porcentaje1/porcentaje2/porcentajeN* en la base de datos. A esta variable se le aplica una transformación para obtener el porcentaje total en un formato numérico, utilizando la fórmula del descuento compuesto:

$$D_{\text{total}} = 1 - \left(\prod_{i=1}^{n} (1 - d_i) \right),$$

donde d_i representa cada uno de los descuentos sucesivos.

Una vez realizada la limpieza, es necesario preparar los datos para el entrenamiento de los modelos. Se ha aplicado la estandarización mediante la normalización Z-score para asegurar que todas las características numéricas tengan una media de 0 y una desviación estándar de 1. Este paso es clave para eliminar las disparidades en las escalas de las variables, permitiendo así una comparación más efectiva y uniforme entre diferentes atributos en los algoritmos de *machine learning* [11].

El siguiente paso ha sido manipular las variables categóricas (tipo *factor* en R). Estas variables han sido transformadas utilizando *One Hot Encoding*, dado que se prevé utilizar algoritmos que no aceptan variables categóricas directamente. Una vez completado este proceso, se ha detectado un problema de sobredimensionalidad, ya que existen variables categóricas con alta cardinalidad. Este problema se ha resuelto escogiendo un subconjunto de valores dentro de cada variable categórica basándose en la ganancia de información [12].

Los procesos de normalización y la creación de columnas con *One Hot Encoding* se han ejecutado sobre el conjunto de entrenamiento y luego se han aplicado al conjunto de prueba. Esta metodología de preparación de datos es importante para evitar la introducción de sesgos en el modelo y asegurar una evaluación justa y precisa.

Capítulo 3

Modelado

En este capítulo se describirá el proceso ejecutado para entrenar y evaluar los modelos de predicción. Esta ha sido una fase importante del proyecto y ha requerido gran parte del tiempo, ya que se ha llevado a cabo un análisis profundo del estado del arte de los modelos de predicción de cancelaciones actuales. Además, se han probado diferentes metodologías y preprocesamientos para buscar el mejor rendimiento posible.

3.1. Exploración de posibles modelos

El primer paso es decidir qué modelos se van a entrenar y evaluar; la selección realizada determinará el tipo de preprocesamiento que se aplique a los datos e incluso el método para evaluar y buscar sus hiperparámetros.

En la actualidad, los modelos de predicción de cancelaciones que proporcionan los mejores resultados son los modelos de clasificación basados en árboles, especialmente aquellos que utilizan técnicas de *gradient boosting*, como XGBoost, LightGBM o AdaBoost [13][14]. Para este proyecto, se han descartado algoritmos como el de K-vecinos más cercanos o las máquinas vectores de soporte debido a la ineficiencia y malos resultados en este ámbito [15]. Teniendo en cuenta estos datos, se ha decidido realizar el entrenamiento de varios modelos y comparar sus resultados y su velocidad de entrenamiento. Los algoritmos seleccionados han sido los siguientes: Random Forest [16], XGBoost [17], LightGBM [18] y la implementación del paquete *caret* de R de un perceptrón multicapa (MLP) [19].

Random Forest

Motivación: Random Forest es un método ampliamente utilizado en aprendizaje automático que combina múltiples árboles de decisión para mejorar la precisión y reducir el riesgo de sobreajuste. Este enfoque es particularmente eficaz en el manejo de datos con muchas características y el descubrimiento de relaciones complejas.

Ventajas: Entre sus principales ventajas se incluyen el soporte para datos faltantes y desequilibrados, una buena capacidad de generalización y la facilidad para interpretar la importancia de las variables.

XGBoost

Motivación: XGBoost (*Extreme Gradient Boosting*) es una técnica avanzada de *gradient boosting* que ha demostrado ser extremadamente efectiva en una amplia gama de problemas de

aprendizaje supervisado. Su diseño optimizado le permite manejar grandes volúmenes de datos con alta precisión.

Ventajas: Este modelo destaca por su manejo eficiente de memoria, capacidad de paralelización y una amplia serie de parámetros ajustables que permiten una gran flexibilidad y optimización.

LightGBM

Motivación: LightGBM (*Light Gradient Boosting Machine*) es otra técnica de *gradient boosting* diseñada para ser extremadamente rápida y eficiente. Es especialmente adecuada para conjuntos de datos grandes y de alta dimensionalidad.

Ventajas: LightGBM ofrece una velocidad de entrenamiento superior, menor uso de memoria y una capacidad para manejar grandes volúmenes de datos con una precisión comparable a la de XGBoost. También cuenta con una extensa variedad de parámetros que permiten flexibilidad y funciones como el manejo de valores faltantes o la robustez frente al balanceo.

Perceptrón Multicapa (MLP)

Motivación: El perceptrón multicapa es una red neuronal artificial capaz de capturar relaciones no lineales complejas en los datos. Su estructura de múltiples capas le permite modelar interacciones sofisticadas entre las variables de entrada, lo que es esencial para tareas de alta complejidad.

Ventajas: Las principales ventajas del MLP incluyen su gran capacidad para modelar relaciones no lineales y la flexibilidad en la arquitectura del modelo (número de capas y neuronas por capa).

Al seleccionar estos modelos, se busca no solo evaluar la precisión y la capacidad predictiva, sino también considerar la velocidad de entrenamiento y la eficiencia en el uso de recursos. Esto permitirá identificar el modelo que mejor se adapte a las necesidades y restricciones del problema en estudio.

3.2. Metodología

En esta sección se explicará la metodología aplicada para el entrenamiento y comparación de modelos de predicción. Es fundamental prestar atención a cada paso del proceso de modelado, ya que una ejecución incorrecta puede producir evaluaciones imprecisas o modelos incapaces de generalizar correctamente a datos no vistos.

El proceso de creación del modelo pasa por las siguientes fases: división del conjunto de datos, selección de estrategia de clasificación, búsqueda de hiperparámetros, entrenamiento del modelo y evaluación.

División del conjunto de datos

Para realizar una evaluación del modelo justa, se deben fragmentar los datos en dos conjuntos diferentes, el de entrenamiento y el de prueba. Para el proyecto se han planteado dos estrategias

para la división de datos, el muestreo aleatorio y la división temporal (*time-based split* en inglés).

El muestreo aleatorio es el método más común en el contexto del *machine learning*; se basa en la selección imparcial de datos, asegurando que cada instancia del conjunto original tenga la misma probabilidad de ser incluida en el conjunto de entrenamiento que en el de prueba. Este enfoque minimiza el sesgo y mejora la capacidad del modelo para generalizar a datos no vistos [20].

La división temporal, por otro lado, se utiliza cuando los datos tienen una estructura temporal inherente, como en las series de tiempo. En este método, los datos se dividen en función de un punto en el tiempo: los datos anteriores a este punto se utilizan para el entrenamiento y los posteriores para la prueba. Esta estrategia es fundamental para problemas con dependencias temporales significativas, ya que refleja mejor las condiciones en las que el modelo se aplicará en la práctica y evita la fuga de datos temporales que podría llevar a una sobreestimación del rendimiento [21].

Teniendo en cuenta las características de los datos, que incluyen fechas de reserva y de *check-in*, se ha decidido utilizar la división temporal. El clasificador a entrenar debe tener la capacidad de generalización suficiente para tener un buen rendimiento en fechas para las que no fue entrenado. Un muestreo aleatorio podría causar un sobreaprendizaje en el intervalo temporal perteneciente al conjunto de entrenamiento y no ofrecer buenos resultados sobre intervalos no vistos.

Estrategia de clasificación

En el desarrollo de modelos de predicción, es crucial decidir cómo se interpretarán las salidas del modelo. Existen dos enfoques principales: la clasificación directa y el cálculo de probabilidad de pertenencia a una clase [22].

La clasificación directa es el método tradicional donde el modelo predice la clase a la que pertenece una instancia. Este enfoque es sencillo y proporciona una respuesta clara y rápida. Cada instancia es asignada a una clase específica sin ambigüedad, lo cual es útil en aplicaciones donde se requiere una decisión binaria inmediata. Sin embargo, este método no proporciona información sobre la incertidumbre de la predicción, lo que puede ser una limitación en escenarios donde es importante comprender la confianza del modelo en su decisión.

Por otro lado, la clasificación basada en probabilidad implica que el modelo devuelve la probabilidad de pertenencia a cada clase para una instancia dada. Posteriormente, se define un umbral para determinar la clase final. Este enfoque ofrece varias ventajas, como la capacidad de ajustar el umbral para equilibrar la precisión y sensibilidad del clasificador según las necesidades específicas del problema. Además, permite interpretar la predicción en términos de probabilidad, proporcionando una medida de confianza sobre la clasificación. Esta información adicional es valiosa en aplicaciones donde las consecuencias de las decisiones incorrectas son significativas y es necesario un análisis más detallado de los resultados del modelo.

Dada la importancia del análisis de los resultados y la evaluación de la confianza, se han optado por el cálculo de la probabilidad de pertenencia a una clase. Al utilizar un enfoque basado en probabilidad, se puede ajustar el umbral en base a objetivos de negocio y la capacidad de asumir el riesgo de equivocarse. Esta flexibilidad y la información adicional proporcionada por las probabilidades mejoran la capacidad de tomar decisiones informadas y precisas, maximizando así el valor del modelo predictivo en la práctica.

Búsqueda de hiperparámetros

Para maximizar el rendimiento del modelo, es necesario seleccionar los hiperparámetros correctos. En este proyecto, se han contemplado dos métodos: *Grid Search* y el algoritmo genético.

Grid Search es un método exhaustivo que implica definir un conjunto de valores posibles para cada hiperparámetro y luego evaluar el rendimiento del modelo para cada combinación posible de estos valores. Aunque este método puede ser computacionalmente costoso, garantiza encontrar la mejor combinación dentro del espacio de búsqueda definido. Cada combinación se evalúa usando validación cruzada para asegurar la fiabilidad de las métricas obtenidas [23].

El algoritmo genético, por otro lado, es un método inspirado en los principios de la evolución natural. Este enfoque comienza con una población inicial de posibles soluciones (combinaciones de hiperparámetros) y las evalúa usando una función de aptitud. Las mejores soluciones se seleccionan para "reproducirse", combinando y mutando sus valores para crear una nueva generación de soluciones. Este proceso se repite durante varias generaciones, con el objetivo de encontrar una combinación óptima de hiperparámetros. Los algoritmos genéticos son particularmente útiles para espacios de búsqueda grandes y complejos donde métodos como *Grid Search* serían impracticables debido a su alta demanda computacional [24].

Para el proyecto, se ha utilizado el método de *Grid Search* para los modelos Random Forest y MLP, debido a su menor cantidad de hiperparámetros, y se ha optado por el algoritmo genético para LightGBM y XGBoost, debido a su gran espacio de hiperparámetros y al coste computacional de aplicar *Grid Search* en estos modelos.

Entrenamiento

La elección del método de entrenamiento adecuado es esencial para asegurar que el modelo predictivo tenga un buen desempeño y que no sufra sobreaprendizaje. En este proyecto, se ha considerado apropiado el uso de la validación cruzada (*cross-validation*).

La validación cruzada es una técnica robusta para evaluar el rendimiento de un modelo predictivo y asegurar la obtención de métricas fiables. Consiste en dividir el conjunto de datos en varios subconjuntos o *folds*. En cada iteración, se entrena el modelo con todos los *folds* excepto uno, que será el subconjunto de validación. Este proceso se repite k veces (donde k es el número de *folds*), de forma que cada subconjunto se usa exactamente una vez para validación y las pruebas restantes para entrenamiento. Finalmente, se promedian los resultados de todas las iteraciones para obtener una estimación más precisa y robusta del rendimiento del modelo. Esta técnica es ejemplificada en la Figura 3.1.

Dado el gran tamaño del conjunto de datos empleado, se ha decidido establecer un número de *folds* reducido, de $k = 3$, para mitigar el coste computacional. Con esta configuración, se equilibra la necesidad de una evaluación robusta y fiable del modelo con la realidad de los recursos computacionales disponibles, asegurando así un proceso de validación eficiente y efectivo [25].

Figura 3.1: Funcionamiento de la validación cruzada [26].

La implementación de la validación cruzada depende del método de búsqueda de hiperparámetros. Para la búsqueda por *Grid Search*, la validación cruzada se aplica a la hora de probar cada configuración de hiperparámetros del espacio definido; la evaluación de los resultados de cada entrenamiento determinará la configuración final para el modelo. En el caso del algoritmo genético, esta técnica es aplicada en la función de aptitud que permite al algoritmo decidir la configuración óptima del modelo.

Evaluación

Para la evaluación de los modelos desarrollados se ha utilizado un conjunto de métricas seleccionadas que permiten identificar cuándo un clasificador es más apropiado que otro para este proyecto. Los principales criterios de evaluación utilizados han sido el área bajo la curva (AUC) y el *F1-score*. Además, se han calculado otras métricas como la precisión, la sensibilidad, la exactitud y la tasa de falsos negativos (FNR).

1. Precisión

La precisión es la proporción de verdaderos positivos entre todas las predicciones positivas realizadas por el modelo. En este proyecto, refleja el porcentaje cancelaciones reales que existen entre todas las reservas clasificadas como cancelación. Una alta precisión ofrece seguridad en el criterio del modelo para detectar positivos. La fórmula de la precisión es:

$$\text{Precisión} = \frac{TP}{TP + FP},$$

donde TP es el número de verdaderos positivos y FP es el de falsos positivos.

2. Sensibilidad

La sensibilidad, también conocida como *recall* o tasa de verdaderos positivos, mide la proporción de verdaderos positivos correctamente identificados por el modelo. En este proyecto, una alta sensibilidad indica que el modelo es capaz de identificar la mayoría de las cancelaciones. La fórmula de la sensibilidad es:

$$\text{Sensibilidad} = \frac{TP}{TP + FN},$$

donde TP es el número de verdaderos positivos y FN es el número de falsos negativos.

3. Tasa de falsos negativos

La Tasa de falsos negativos (FNR) mide la proporción de instancias positivas que fueron incorrectamente clasificadas como negativas por el modelo. Es una métrica importante en este proyecto debido a que indica la cantidad de cancelaciones que resultan identificadas como confirmadas, lo cual resulta en habitaciones libres que se esperaban ocupar. La fórmula de la tasa de falsos negativos es:

$$\text{FNR} = \frac{FN}{FN + TP},$$

donde FN es el número de falsos negativos y TP es el número de verdaderos positivos.

4. Exactitud

La exactitud (*accuracy*) mide la proporción de predicciones correctas sobre el total de predicciones realizadas. Es una métrica sencilla y ampliamente utilizada, que en este proyecto hace referencia a la cantidad de reservas clasificadas correctamente. La fórmula de la exactitud es:

$$\text{Exactitud} = \frac{TP + TN}{TP + TN + FP + FN},$$

donde TP es el número de verdaderos positivos, TN es el número de verdaderos negativos, FP es el número de falsos positivos y FN es el número de falsos negativos.

5. Área bajo la curva

El Área Bajo la Curva (AUC) es una métrica que evalúa el rendimiento de un clasificador al medir el área bajo su curva ROC (*Receiver Operating Characteristic*). La curva ROC representa la relación entre la tasa de verdaderos positivos (sensibilidad) y la tasa de falsos positivos a varios umbrales de decisión [27]. Un valor de AUC cercano a 1 indica un buen rendimiento del modelo. El AUC se interpreta como la probabilidad de que el clasificador asigne una puntuación más alta a una muestra positiva que a una negativa [28]. En este proyecto ha sido la métrica a maximizar a la hora de entrenar los modelos de clasificación. Esta elección se fundamenta en su capacidad para evaluar efectivamente el rendimiento del modelo incluso con el desequilibrio de clases existente, así como en su habilidad para evaluar clasificadores probabilísticos sin depender de un umbral de clasificación específico.

6. F1-score

El *F1-score* es la media armónica de la precisión y la sensibilidad (*recall*). Es una métrica útil cuando se necesita un equilibrio entre precisión y *recall*, especialmente en casos de datos desbalanceados. Un *F1-score* alto indica que el clasificador tiene tanto una alta precisión como una alta sensibilidad. Para este proyecto ha sido una métrica clave, ya se ha utilizado para la selección del umbral de clasificación. El motivo del uso de este estadístico en la selección del umbral radica en el propósito de equilibrar los falsos positivos y los falsos negativos en el modelo. Cuando estas dos medidas son equivalentes, en la práctica, los errores se contrarrestan provocando que la disponibilidad de habitaciones y la cantidad de cancelaciones sea igual a la esperada. La fórmula del *F1-score* es:

$$\text{F1-score} = 2 \times \frac{\text{Precisión} \times \text{Recall}}{\text{Precisión} + \text{Recall}}.$$

3.3. Resultados

En esta sección se mostrarán los resultados obtenidos del entrenamiento de los modelos seleccionados. Todos los clasificadores han sido creados utilizando los datos de las reservas de 2022 como conjunto de entrenamiento. El umbral de clasificación se ha calculado maximizando el *F1-score* en los 4 primeros meses de 2023, y se ha utilizado el resto de reservas (desde mayo del 2023 hasta mayo del 2024) como conjunto de prueba.

Modelo	AUC	F1-Score	Umbral	Precisión	Sensibilidad	FNR	Exactitud
Random Forest	0.8390	0.6330	0.30	58.83	68.52	10.81	80.51
MLP	0.8017	0.6014	0.4	52.83	69.80	11.32	79.03
LightGBM	0.8383	0.6414	0.60	59.82	69.15	10.57	81.03
XGBoost	0.8396	0.6417	0.30	57.57	72.48	9.77	80.14

Cuadro 3.1: Resultados proporcionados por las distintas implementaciones.

Una vez realizada la evaluación de los clasificadores, se procede a seleccionar el más apropiado para este problema. Para ello, se comienza observando el Cuadro 3.1, en el que se presentan los resultados de los diferentes clasificadores. Pese a que sus rendimientos son bastante similares en cuanto a métricas, se puede notar que el MLP es inferior al resto de modelos, por lo que será descartado. A continuación, se procede a comparar el resto de clasificadores en cuanto a la precisión en las probabilidades que estiman. En una predicción ideal, las reservas etiquetadas con un 30 % de probabilidad de cancelación, resultan canceladas el 30 % de las ocasiones. Para valorar este comportamiento, se ha creado un gráfico que muestra el porcentaje de reservas que resultan canceladas en función de la probabilidad de cancelación estimada. En las Figuras mostradas más adelante, se representa en azul el porcentaje de ocasiones en la que una reserva con x probabilidad de cancelación (%) resulta finalmente cancelada. En negro, por otro lado, se muestra la recta $y = x$, que representa la referencia en cuanto a lo explicado anteriormente. En este proyecto, es interesante la capacidad de variar el criterio de clasificación, subiendo o bajando el umbral definido sin producir comportamientos irregulares. Esto se utilizará para ajustar el riesgo con el que se decide que un cliente no va a completar su reserva. Para ello, la gráfica debe presentar una distribución suave y con un crecimiento uniforme de la proporción de cancelación.

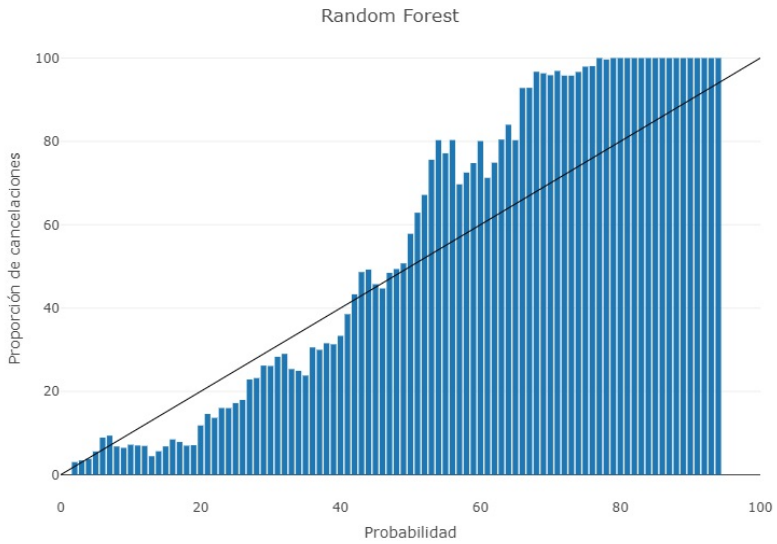

Figura 3.2: Distribución de probabilidad del modelo Random Forest.

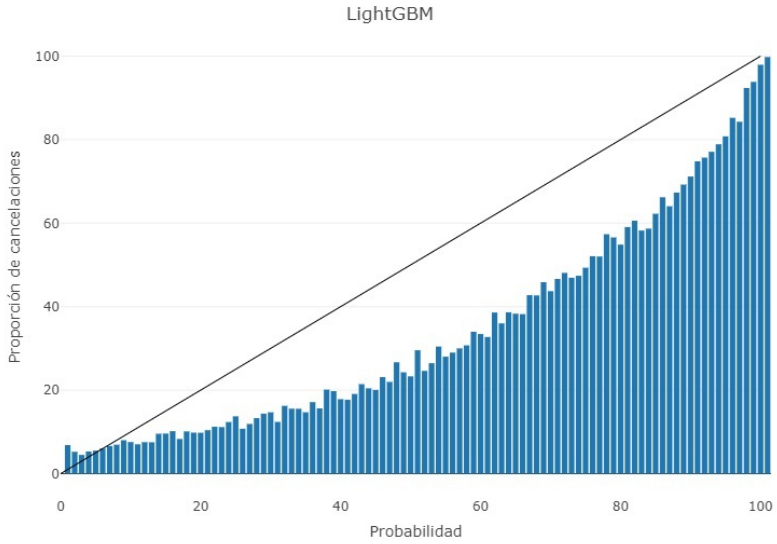

Figura 3.3: Distribución de probabilidad del modelo LightGBM.

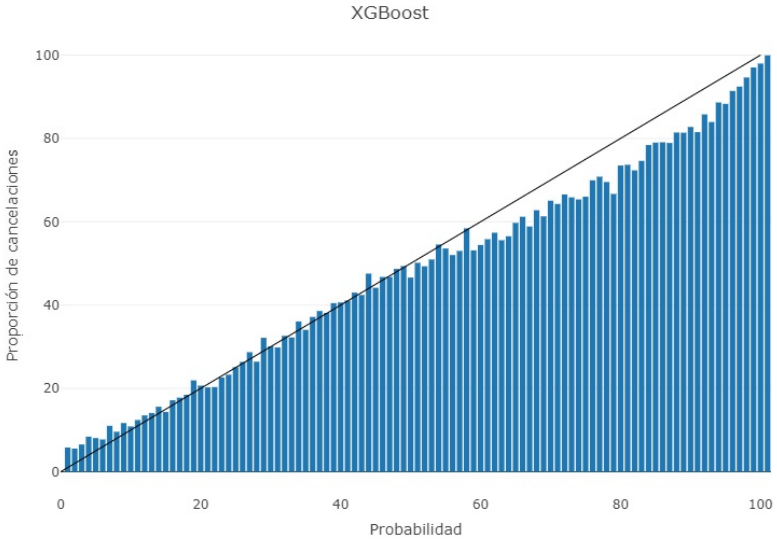

Figura 3.4: Distribución de probabilidad del modelo XGBoost.

En la Figura 3.2, se puede visualizar una distribución muy irregular en el Random Forest, lo cual hace al modelo inestable ante cambios en el umbral. Por otro lado, en la Figura 3.3, LightGBM muestra un crecimiento suave, pero alejado de la recta $y = x$. Por último, XGBoost presenta una distribución muy próxima a la del modelo de referencia, por lo que su estimación de probabilidades es acertada y representativa.

Tras la visualización de estas Figuras, se puede concluir que el modelo XGBoost mostrado en la Figura 3.4 es la mejor opción para este problema. A continuación, en la Figura 3.5, se visualiza la importancia de las variables numéricas en este clasificador. En ella destacan las variables de la antelación y la existencia de descuento en la reserva, dos factores que ya se habían comentado anteriormente.

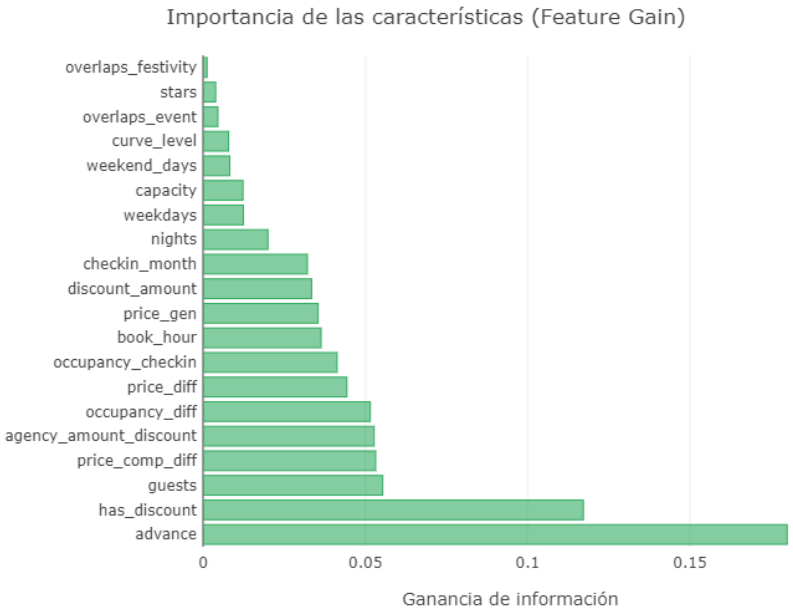

Figura 3.5: Importancia de las características en XGBoost.

Por último, cabe destacar que los resultados de la implementación basada en XGBoost se han comparado con los proporcionados por una empresa experta en inteligencia artificial que ha realizado una prueba para Eurostars. Las métricas calculadas son referentes a la predicción de las cancelaciones de un hotel de Barcelona en abril de 2024, con alrededor de 3.000 reservas. El enfoque de la empresa externa fue el de crear un modelo único de predicción de cancelaciones para dicho hotel, entrenándolo con unos 80 mil registros. En cambio, el clasificador entrenado en este proyecto, es capaz de clasificar reservas en 16 hoteles de la compañía. Ambos modelados se han realizado con las mismas variables de origen y las reservas pertenecientes al mismo intervalo temporal. A priori, un modelo único para el hotel de interés debe ofrecer mejor rendimiento que un clasificador dedicado a varios hoteles, ya que puede obtener patrones propios del comportamiento del establecimiento y no requiere la misma capacidad de generalización. En cambio, en el Cuadro 3.2 se puede apreciar que, pese a estar entrenado para un conjunto de hoteles mucho mayor, el modelo realizado en este proyecto presenta resultados notablemente mejores que el ofrecido por la empresa externa.

Modelo	F1-Score	Precisión	Sensibilidad	FNR	Exactitud
XGBoost	0.3172	55.03	22.28	12.72	85.27
Empresa externa	0.1906	44.35	12.14	2.97	83.17

Cuadro 3.2: Comparativa entre los resultados obtenidos con XGBoost y el modelo de la empresa externa.

Esto evidencia los buenos resultados del clasificador desarrollado en este proyecto, superando por más de 10 puntos la precisión y la sensibilidad del modelo creado por una empresa experta en inteligencia artificial, mientras ofrece una capacidad de generalización mayor, permitiendo realizar predicciones sobre una selección de hoteles variados en la compañía.

Capítulo 4

Creación de herramienta visual

En este capítulo se describe la herramienta creada para su uso por parte del departamento de *revenue*. La aplicación debe proporcionar un uso sencillo y una visualización clara y útil para perfiles que no tienen conocimientos avanzados sobre *machine learning*.

4.1. Intención de negocio

La interfaz de visualización desarrollada ha sido creada con la supervisión del departamento de *revenue*, con el que se colaboró para poder ofrecer la mayor utilidad posible y facilitar tareas diarias. Esta interacción entre departamentos ha sido muy enriquecedora para el proyecto, debido a que ha ayudado a comprender con mayor profundidad las consecuencias de una reserva clasificada como cancelación en la estrategia de negocio.

Como resultado, se han concretado los casos de uso que debe cumplir la interfaz. La herramienta debe permitir una visualización clara de la predicción de reservas a los *revenue managers*. Esta información será utilizada por el departamento para brindar descuentos, modificar los precios de las habitaciones y prever gastos en los hoteles. Además, debe ser capaz de mostrar una evaluación del rendimiento de la predicción y del beneficio estimado aportado para fechas pasadas. Para mejorar la flexibilidad del modelo y de la estrategia de negocio, también se ofrece la posibilidad al usuario de modificar la prudencia con la que el modelo clasifica una reserva como cancelación.

El cálculo del beneficio estimado se realiza teniendo en cuenta los aciertos y errores del modelo, observando las consecuencias de cada predicción. Para hallar el beneficio se supone que: (1) todas las reservas clasificadas como cancelaciones se revenden, (2) las habitaciones revendidas se rebajan un 10 % para incentivar la venta, (3) los costes del *overselling* son iguales al precio de la habitación que ha reservado el cliente y (4) los costes de preparar una habitación son de 10 €. Asumir estos supuestos permite observar en la práctica el beneficio máximo que se obtendría si la predicción fuese acertada y evaluar los riesgos de una clasificación errónea. Con una predicción precisa, se revenden las mismas habitaciones que se cancelan, manteniendo así el beneficio esperado antes de las cancelaciones. En cambio, para una clasificación errónea existen varios escenarios diferentes que se estudiarán a continuación. En caso de que la clasificación contenga muchos falsos negativos, lo que se produce es una reducción de la ocupación en el hotel, ya que se esperaban reservas confirmadas que resultaron canceladas. Para el hotel, esto es una pérdida de oportunidad de reventa, ya que se obtiene menos beneficio del esperado y resultan más habitaciones libres de las deseadas. Por otro lado, si la clasificación contiene una gran cantidad de falsos positivos, lo que ocurre es que se revenden habitaciones que estarán

ocupadas, por lo que existe un peligro de sobreventa. El hotel tendrá más ocupación de la esperada y, si esta cantidad es mayor que la suma de las habitaciones disponibles y los falsos negativos, resultará en *overselling*, con las consecuencias ya explicadas.

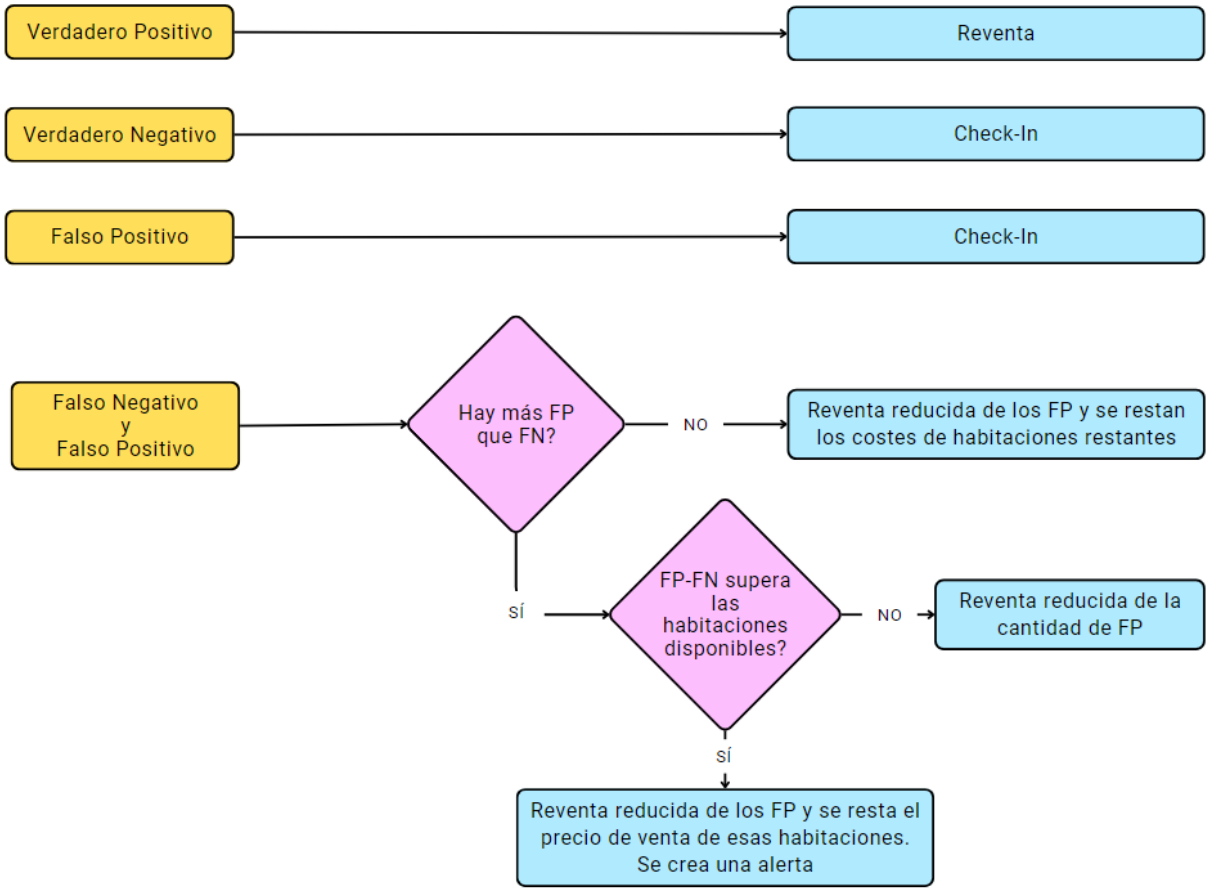

Figura 4.1: Diagrama explicativo del cálculo de beneficio.

En la Figura 4.1 se muestra la lógica aplicada para calcular el beneficio. Como se puede observar, todas las reservas confirmadas (verdaderos negativos y falsos positivos) completan el *check-in* en el hotel, por lo que se suman al beneficio. En el caso de los verdaderos positivos, se realiza una reventa y, por ende, se añaden de nuevo a la ganancia total. Por otro lado, se debe prestar especial atención al balance de las reservas mal clasificadas. Primero, se evalúa si el hotel obtuvo una ocupación mayor o menor a la esperada, lo cual depende directamente de la diferencia entre las predicciones incorrectas. Si la cantidad de falsos negativos es mayor que la de falsos positivos, significa que hay más reservas canceladas que confirmadas mal clasificadas. Esto reduce la ocupación y, por lo tanto, la ganancia esperada, pero no supone ningún riesgo de sobreventa. En el caso contrario, se analiza si esta diferencia de ocupación supera las habitaciones disponibles, lo que provocaría *overselling*. Esto se determina comparando las confirmaciones y las cancelaciones inesperadas. Si esta diferencia no supera la disponibilidad de habitaciones, no hay sobreventa, por lo que se suma a la ganancia la reventa de las habitaciones (la cantidad de falsos positivos). Si esta diferencia es mayor a las habitaciones disponibles, se añade a la ganancia la cantidad de falsos positivos hasta completar la ocupación y se restan del beneficio las habitaciones sobrevendidas, ya que es necesario reubicar a los clientes en otro hotel. Además, en este caso, se genera una alerta en la aplicación para que el departamento de *revenue* pueda estudiar el suceso con más atención.

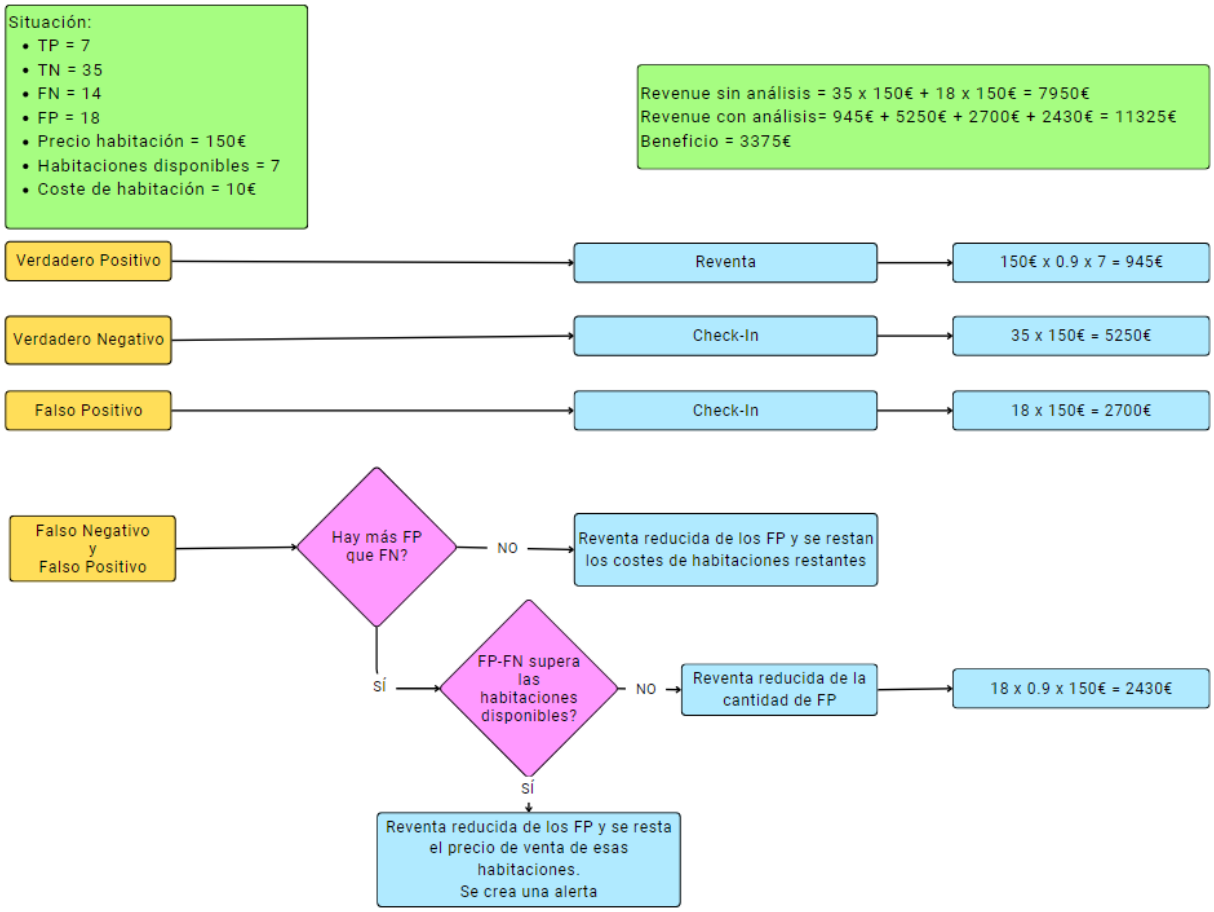

Figura 4.2: Ejemplo de cálculo de beneficio.

En la Figura 4.2 se observa un ejemplo práctico del cálculo de beneficio y la ventaja de realizar un análisis de predicción de cancelaciones. Como se puede advertir, existe una mayor cantidad de falsos positivos que falsos negativos y solo quedan siete habitaciones libres en el hotel. Los beneficios obtenidos de los *check-in*, es decir, los verdaderos negativos y los falsos positivos, son comunes independientemente de si se ha realizado el análisis de cancelaciones o no. Los beneficios del análisis se encuentran en la oportunidad de reventa de habitaciones. En este ejemplo hay siete verdaderos positivos y 18 falsos positivos que se transforman en reventas, esto se traduce en un ingreso potencial de 3.375 €. Esta cantidad excede la disponibilidad de habitaciones, pero debido a los 14 falsos negativos, la reventa resulta en solo cuatro habitaciones menos de las esperadas en el hotel. Se puede apreciar la importancia de maximizar el *F1-score* en este proyecto, ya que permite igualar la cantidad de falsos negativos y falsos positivos en la predicción y prevenir el *overbooking* o la obtención de demasiadas cancelaciones inesperadas. Debido a esto, el beneficio final es igual a 11325 €, lo que supone un aumento del 42 % respecto al *revenue* obtenido de no haber realizado el análisis de cancelaciones.

4.2. Implementación

La herramienta ha sido creada utilizando *Shiny App* [29]. Esta librería permite la creación de aplicaciones web interactivas de forma sencilla, facilitando la integración de visualizaciones y análisis de datos en R definiendo interfaces reactivas y dinámicas.

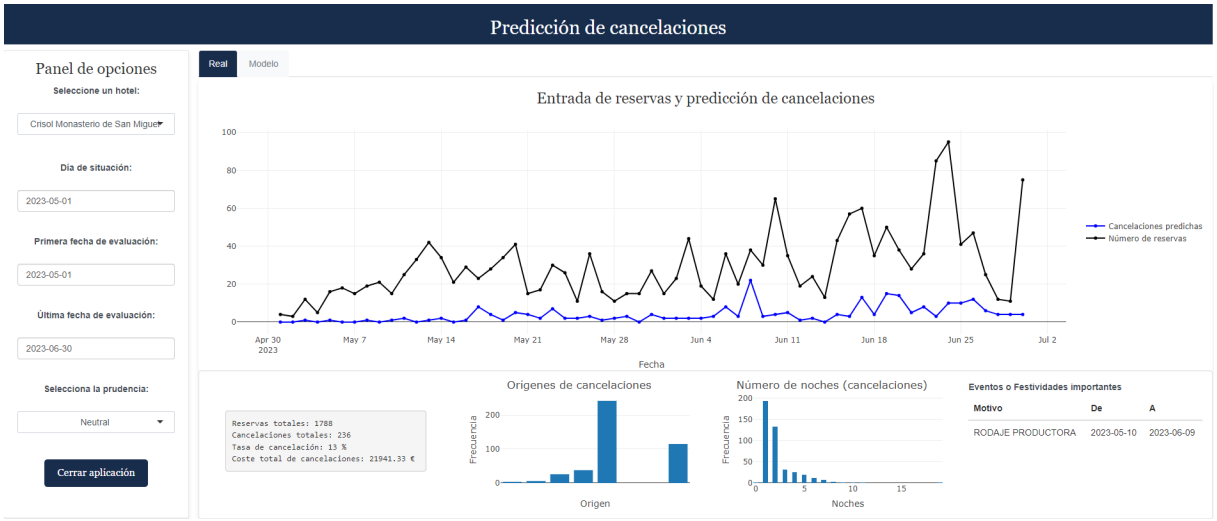

Figura 4.3: Página principal de la herramienta.

En la Figura 4.3 se observa la pantalla principal de la aplicación, en la que se identifican tres espacios: el panel de opciones a la izquierda, la ventana principal con la línea de reservas y cancelaciones y la sección de gráficas y análisis. En el panel de opciones, el usuario puede seleccionar el hotel, la fecha de situación, el intervalo a examinar y la prudencia. La fecha de situación es la fecha desde la cual quiere realizarse el análisis. Al seleccionarla, solo se tienen en cuenta las reservas efectuadas antes de esa fecha que aún no se hayan cancelado, lo cual resulta útil para que el usuario pueda examinar el proceso y analizar la evolución de las reservas y la predicción. La prudencia modifica el criterio de clasificación del modelo subiendo o bajando el umbral de clasificación un 10 %, permitiendo así ser más precavido a la hora de clasificar una reserva como cancelada. La ventana principal contiene dos pestañas. La primera de ellas muestra la entrada de reservas y la acumulación de cancelaciones para las fechas seleccionadas. Además, permite ampliar información sobre la cantidad, el coste total relativo a las cancelaciones o el beneficio de las reservas al colocar el cursor encima de cualquier punto de la gráfica. La segunda pestaña contiene la sección de evaluación del modelo, en la cual se visualiza, además de las líneas de reservas y predicción, la línea de cancelaciones reales, tal y como se observa en la Figura 4.4. Esta sección se utiliza para estudiar el rendimiento obtenido en periodos anteriores y permite al departamento de *revenue* evaluar la clasificación y estudiar posibles cambios en la estrategia de negocio.

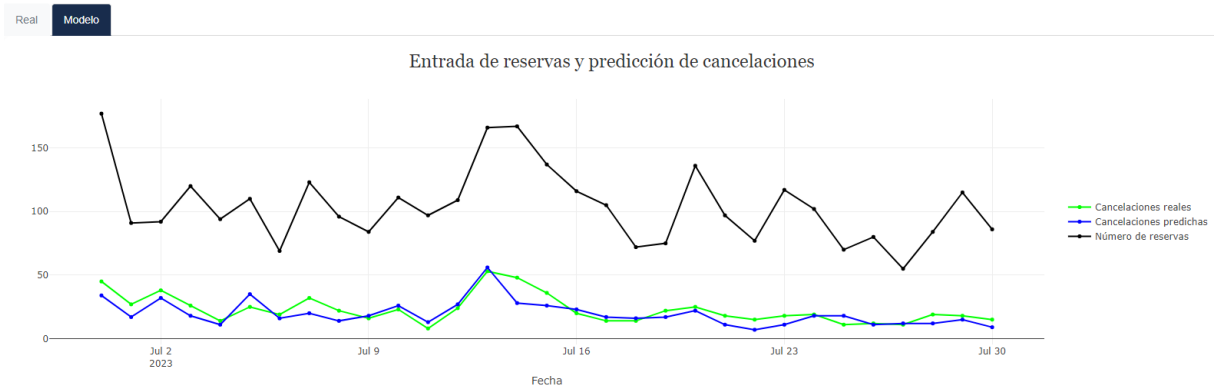

Figura 4.4: Pestaña de evaluación del modelo.

La sección de gráficas y análisis muestra datos relativos a la pestaña que se esté seleccionada en ese momento. Para la de predicción, se ofrece información sobre la cantidad de reservas totales, la tasa de cancelación y la suma del coste de las cancelaciones, representada en la Figura 4.5. También se visualizan dos gráficas que resumen los orígenes y la duración de las reservas canceladas. Finalmente, se genera una tabla que lista los eventos y festividades relevantes durante el período analizado. Esto facilita al usuario estudiar posibles razones detrás de los picos en las reservas o de comportamientos inesperados.

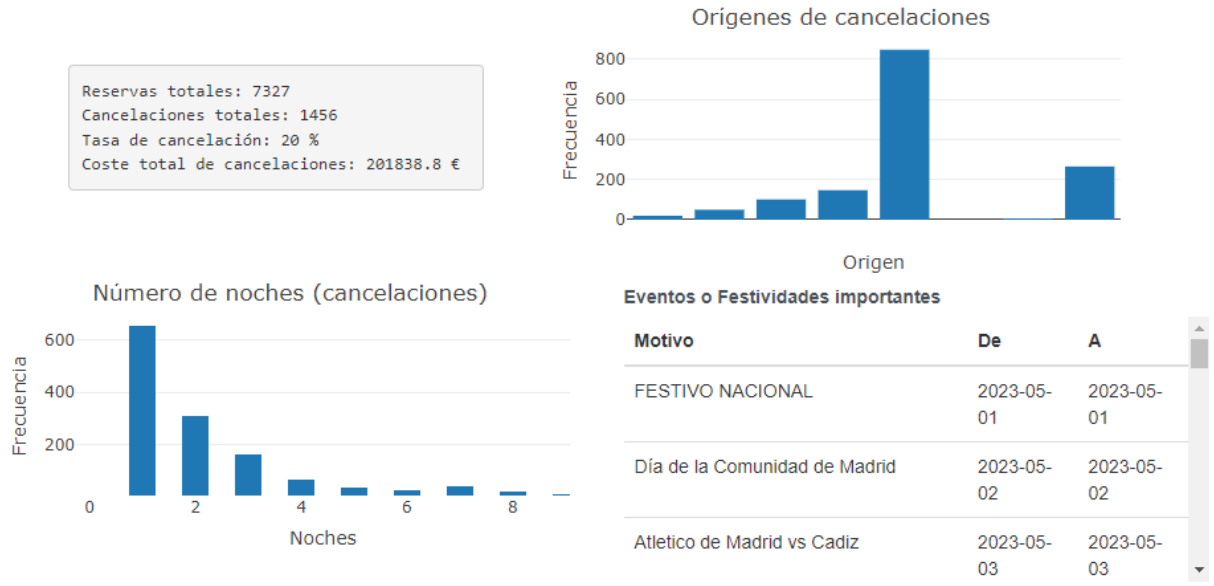

Figura 4.5: Sección de análisis de la predicción.

Cuando la pestaña de evaluación del modelo está seleccionada, la sección de análisis muestra información útil para la evaluación del rendimiento en el intervalo escogido. Se incluye la matriz de confusión, el cálculo del beneficio estimado y el error total en las cancelaciones.

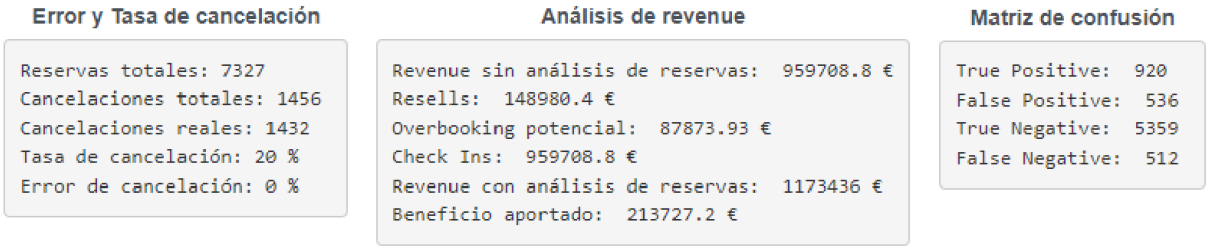

Figura 4.6: Sección de análisis del modelo.

En conclusión, la herramienta desarrollada permite al equipo de ventas visualizar y ajustar la predicción en base al riesgo permitido, de una forma sencilla, flexible y alineada con la intención de negocio. Además ofrece la información útil sobre las reservas que han sido clasificadas como posibles cancelaciones. Por último, el departamento puede comprobar si el rendimiento del modelo ha sido satisfactorio en un pasado y cuánto beneficio podría aportar la aplicación de estrategias basadas en la detección de cancelaciones.

Capítulo 5

Conclusiones y posibles ampliaciones

Como conclusión general, la realización de este proyecto ha permitido completar los objetivos establecidos para el mismo. El estudio llevado a cabo ha permitido aportar información útil al departamento sobre los patrones de cancelación y los posibles beneficios de enfocar estrategias de negocio en este análisis. Además, se ha desarrollado un modelo de predicción flexible y eficaz que incluso ha superado en rendimiento a un clasificador implementado por una empresa externa experta en inteligencia artificial. Por último, se ha creado una herramienta de visualización que permite al departamento de *revenue* ejecutar predicciones y analizar de una forma simple y rápida la evolución de las reservas en 16 hoteles de la compañía.

Por otro lado, este proyecto ha sido una experiencia muy enriquecedora para el alumno, ya que no solo ha permitido el aprendizaje de nuevas técnicas aplicables durante la preparación de los datos y el modelado, sino que también ha mejorado significativamente su desempeño en el lenguaje de programación R, con el que tenía poca experiencia. A su vez, le ha permitido asentar y aplicar conocimientos adquiridos durante el máster en un entorno real, como, por ejemplo, la visualización y minería de datos, la creación de consultas SQL, el uso de técnicas de aprendizaje estadístico y la aplicación de estas a un caso de uso empresarial.

Posibles ampliaciones

Tras el desarrollo y análisis de este proyecto, se han identificado varias áreas en las que se podrían realizar ampliaciones y mejoras futuras. Estas posibles ampliaciones incluyen:

- **Ampliar la cantidad de hoteles:** Ajustar el modelo para poder realizar predicciones precisas en más hoteles mejoraría la capacidad de aplicar estrategias basadas en el análisis de cancelaciones de Eurostars.

- **Incluir la predicción en el sistema de precios:** La variable de probabilidad de cancelación de una reserva puede ser utilizada en el sistema de precios dinámicos de la compañía. Estimar con una mayor precisión la ocupación del hotel en una fecha concreta permitiría realizar cambios en los precios para incrementar la facturación o alcanzar el beneficio esperado.

- **Estudiar un enfoque basado en el acumulado de reservas:** En este proyecto, las probabilidades se utilizan para clasificar las reserva individualmente en base a un umbral determinado. Como trabajo futuro alrededor de este proyecto, se podría estudiar el impacto de utilizar un enfoque acumulado, agrupando las reservas por fecha de *check-in* y estimando la cantidad de cancelaciones en base a las probabilidades de todo el grupo.

- **Utilizar información de cliente en el modelo:** Gracias al nuevo programa de fidelización de Eurostars, la empresa cuenta con la información personal de los clientes recurrentes que realizan la reserva iniciando sesión desde la página web. Los nuevos datos obtenidos por este medio pueden mejorar significativamente la precisión del modelo. Además, el programa de fidelización permite obtener un registro histórico del huésped, lo que puede ayudar a realizar un análisis individual.
- **Mejorar la herramienta de visualización:** Añadir a la herramienta la posibilidad de que el *revenue manager* calcule el beneficio estimado de una forma más personalizable, permitiendo establecer la cantidad y el descuento de habitaciones revendidas. Por otro lado, se podrían identificar los tipos de habitaciones que se clasifican como canceladas, para mejorar la estrategia de reventa.

Bibliografía

[1] Instituto Nacional de Estadística. "Movimientos Turísticos en Fronteras." (2023).

[2] Instituto Nacional de Estadística. "Establecimientos, plazas estimadas, grados de ocupación y personal empleado por puntos turísticos." (2023).

[3] N. Antonio, A. De Almeida y L. Nunes, "Big Data in Hotel Revenue Management: Exploring Cancellation Drivers to Gain Insights Into Booking Cancellation Behavior," *Cornell Hospitality Quarterly*, vol. 60, n.º 4, págs. 298-319, mayo de 2019. DOI: 10.1177/1938965519851466.

[4] M. Velten. "Cancellation policies in combination with scarcity- and social proof appeals : a study into the effects of cancellation policies and persuasion cues on consumer responses within the online booking industry." (jun. de 2017), dirección: http://essay.utwente.nl/72502/.

[5] B. Benítez-Aurioles, "Why are flexible booking policies priced negatively?" *Tourism Management*, vol. 67, págs. 312-325, 2018, ISSN: 0261-5177. DOI: https://doi.org/10.1016/j.tourman.2018.02.008.

[6] F. Provost y T. Fawcett, "Data Science and its Relationship to Big Data and Data-Driven Decision Making," *Big Data*, vol. 1, n.º 1, págs. 51-59, mar. de 2013. DOI: 10.1089/big.2013.1508.

[7] N. Phumchusri y P. Maneesophon, "Optimal overbooking decision for hotel rooms revenue management," *Journal of Hospitality and Tourism Technology*, vol. 5, n.º 3, págs. 261-277, 2014. DOI: 10.1108/JHTT-03-2014-0006.

[8] G. N. Vajpai y U. Ramnagar, "Managing overbooking in hotels: A probabilistic model using Poisson distribution," *Ideas and Innovation in Technology*, vol. 4, n.º 2, págs. 1375-1379, 2018.

[9] R. S. Toh, "An Inventory Depletion Overbooking Model For the Hotel Industry," *Journal of Travel Research*, vol. 23, n.º 4, 1985. DOI: 10.1177/004728758502300404.

[10] H. Vinutha, B. Poornima y B. Sagar, "Detection of Outliers Using Interquartile Range Technique from Intrusion Dataset," en *Information and Decision Sciences*, ép. Advances in Intelligent Systems and Computing, S. Satapathy, J. Tavares, V. Bhateja y J. Mohanty, eds., vol. 701, Singapore: Springer, 2018, págs. 605-618. DOI: 10.1007/978-981-10-7563-6_53.

[11] M. M. Ahsan, M. A. P. Mahmud, P. K. Saha, K. D. Gupta y Z. Siddique, "Effect of Data Scaling Methods on Machine Learning Algorithms and Model Performance," *Technologies*, vol. 9, n.º 3, 2021, ISSN: 2227-7080. DOI: 10.3390/technologies9030052.

[12] B. Azhagusundari, A. S. Thanamani et al., "Feature selection based on information gain," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 2, n.º 2, págs. 18-21, 2013.

[13] N. Antonio, A. de Almeida y L. Nunes, "Predicting Hotel Bookings Cancellation with a Machine Learning Classification Model," en *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2017, págs. 1049-1054. DOI: 10.1109/ICMLA.2017.00-11.

[14] S. Chen, E. W. Ngai, Y. Ku, Z. Xu, X. Gou y C. Zhang, "Prediction of hotel booking cancellations: Integration of machine learning and probability model based on interpretable feature interaction," *Decision Support Systems*, vol. 170, 2023, ISSN: 0167-9236. DOI: https://doi.org/10.1016/j.dss.2023.113959.

[15] A. Herrera, Á. Arroyo, A. Jiménez y Á. Herrero, "Forecasting hotel cancellations through machine learning," *Expert Systems*, DOI: https://doi.org/10.1111/exsy.13608.

[16] S. J. Rigatti, "Random forest," *Journal of Insurance Medicine*, vol. 47, n.º 1, págs. 31-39, 2017.

[17] T. Chen, T. He, M. Benesty et al., "Xgboost: extreme gradient boosting," *R package version 0.4-2*, vol. 1, n.º 4, 2015.

[18] G. Ke, Q. Meng, T. Finley et al., "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems*, vol. 30, 2017.

[19] H. Taud y J.-F. Mas, "Multilayer perceptron (MLP)," *Geomatic approaches for modeling land change scenarios*, págs. 451-455, 2018.

[20] J. Tan, J. Yang, S. Wu, G. Chen y J. Zhao, "A critical look at the current train/test split in machine learning," *arXiv preprint arXiv:2106.04525*, 2021.

[21] R. Medar, V. S. Rajpurohit y B. Rashmi, "Impact of Training and Testing Data Splits on Accuracy of Time Series Forecasting in Machine Learning," en *2017 International Conference on Computing, Communication, Control and Automation (ICCUBEA)*, 2017, págs. 1-6. DOI: 10.1109/ICCUBEA.2017.8463779.

[22] A. A. Soofi y A. Awan, "Classification techniques in machine learning: applications and issues," *J. Basic Appl. Sci*, vol. 13, n.º 1, págs. 459-465, 2017.

[23] S. M. LaValle, M. S. Branicky y S. R. Lindemann, "On the relationship between classical grid search and probabilistic roadmaps," *The International Journal of Robotics Research*, vol. 23, n.º 7-8, págs. 673-692, 2004.

[24] P. Liashchynskyi y P. Liashchynskyi, "Grid search, random search, genetic algorithm: a big comparison for NAS," *arXiv preprint arXiv:1912.06059*, 2019.

[25] S. Yadav y S. Shukla, "Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification," en *2016 IEEE 6th International Conference on Advanced Computing (IACC)*, 2016, págs. 78-83. DOI: 10.1109/IACC.2016.25.

[26] R. Patro. "Cross-Validation: K-Fold vs. Monte Carlo." Acceso: Junio 13, 2024. (2023), dirección: https://towardsdatascience.com/cross-validation-k-fold-vs-monte-carlo-e54df2fc179b.

[27] J. M. Pérez y P. P. Martin, "ROC curve," *Semergen*, vol. 49, n.º 1, 2023.

[28] J. R. Turner, "Area under the curve (AUC)," *Encyclopedia of Behavioral Medicine*, págs. 146-146, 2020.

[29] W. Chang, J. Cheng, J. Allaire et al., *shiny: Web Application Framework for R*, R package version 1.8.1.9001, https://github.com/rstudio/shiny, 2024. dirección: https://shiny.posit.co/.