Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Estudios de Posgrado

CARRERA DE ESPECIALIZACIÓN EN MÉTODOS CUANTITATIVOS PARA LA GESTIÓN Y ANÁLISIS DE DATOS EN ORGANIZACIONES

TRABAJO FINAL DE ESPECIALIZACIÓN

Predicción de la cancelación de reservas en hoteles
a través de técnicas de aprendizaje automático

AUTOR: LISSETTE BOLINAGA OTERO

DICIEMBRE 2019

Resumen

Las cancelaciones de reservas tienen un fuerte impacto en la industria del turismo. En el caso particular de los hoteles, es importante contar con información precisa y anticipada de las reservas y cancelaciones, dado que una incorrecta gestión de las mismas- como por ejemplo con las sobreventas- puede afectar tanto las ganancias como la reputación del hotel. El presente trabajo tiene por objeto predecir las cancelaciones de reservas en hoteles y busca contribuir a entender los posibles factores o causas que explican dicha cancelación. Asimismo, predecir de manera precisa las cancelaciones es una gran herramienta para poder desarrollar un sistema de gestión de la demanda -*revenue management*-. Se analiza un conjunto de datos de hoteles ubicados en Argentina que forman parte de una cadena hotelera, en donde se definieron las cancelaciones como un problema de clasificación para modelar cuáles reservas son plausibles a ser canceladas y cuáles no.

Palabras clave: Big Data, *Machine learning, Revenue Management, Data Science,* Reservas.

Estructura

Introducción

El manejo del gran volumen y variedad de datos que se generan a gran velocidad resulta un desafío para las organizaciones. Desde hace varios años, los datos se han convertido en uno de los activos más importantes a explotar por las empresas (Fosso Wamba, Akter, Edwards, Chopin y Gnanzou, 2015).

La incorporación de Big Data trae la oportunidad de cambiar los modelos de negocio (Schmarzo, 2013) y la forma de tomar decisiones basándose en datos. La creciente combinación de fuentes, herramientas, aplicaciones y datos que se generan, tienen implicancias profundas tanto en la gestión como en la forma de tomar de decisiones de las organizaciones. El uso de Big Data está impactando y transformando muchos sectores como la salud, las finanzas, la implementación de políticas públicas y el sector del turismo en general y el hotelero en particular. A su vez, está cambiando la forma en que se toman las decisiones de las organizaciones tanto del sector privado como del sector público (Kim, Trimi, y Chung, 2014).

La industria del turismo tiene gran importancia en la economía mundial. Según el Consejo Mundial del Viaje y el Turismo (WTTC) durante 2018 esta industria creció 3.9% -por encima de la tasa de crecimiento del Producto Interno Bruto Mundial que fue del 3.2%- y generó más de 319 millones de empleos en el mundo[1]. En Argentina este sector también tiene una gran importancia económica y es un reconocido generador de empleo. En 2018 el Producto Interno Bruto relacionado al Turismo creció un 3.5% mientras la economía en su conjunto cayó 2.8%[2]. A pesar de que para el 2019 se espera otro año de caída del producto, se estima que el sector va a ser de los pocos con crecimiento.

Las cancelaciones tienen un fuerte impacto en la industria del turismo en general -para las aerolíneas, las empresas de alquiler de autos y agencias de viaje-, y particularmente para la industria hotelera. Las mismas tienen efecto tanto en los costos como en las decisiones de asignación de habitaciones y de manejo de la demanda. Es en esta razón que resulta de gran importancia la creación de modelos para predecir dichas cancelaciones, de

[1] WTTC Global Economic Impact & Trends 2019 (Marzo 2019). Accesible desde https://www.wttc.org/economic-impact/country-analysis/

[2] WTTC Argentina Report (Marzo 2018). Accessible desde https://www.wttc.org/economic-impact/country-analysis/country-reports/

manera de poder construir pronósticos de demanda más certeros y desarrollar así mejores sistemas de gestión de la demanda -*revenue management*-, para perfeccionar la política de precios, optimizar la capacidad y la disponibilidad de los hoteles, mejorar la estrategia de marketing y la política de cancelaciones, entre otras razones.

Nuno, de Almeida y Nunes (2017) señalan que las cancelaciones de las reservas afectan el negocio de la industria hotelera y limitan la posibilidad de proyectar y gestionar la demanda correctamente. Adicionalmente a ello, contar con malas predicciones puede impactar tanto en las ganancias como en la reputación del hotel.

A partir del análisis de un conjunto de datos de hoteles localizados en Argentina que forman parte de una cadena hotelera, se busca generar un modelo para predecir las reservas que van a ser canceladas. A través de dichas predicciones se busca apoyar la toma de decisiones de la organización y contribuir a entender los posibles factores o causas que llevan a una cancelación.

En la actualidad, el sector de turismo fue desarrollando maneras para disminuir el riesgo de las cancelaciones, en el caso de los hoteles, pueden mencionarse la sobreventa -overbooking- de habitaciones o el endurecimiento de las condiciones de devolución de las reservas. Sin embargo, en un contexto de oferta amplia de hoteles y de fácil acceso mediante las *Online Travel Agencies* (OTAs), endurecer las condiciones de devolución dificulta la venta del servicio y pone en situación de desventaja a los hoteles que lo implementan.

Es por eso que, para optimizar las estrategias de Gestión de la demanda, se busca crear un modelo predictivo de cancelaciones a través de la prueba de diferentes algoritmos y aprovechando los grandes volúmenes de datos relacionados con los que cuentan los hoteles, para producir mejores proyecciones que contribuyan en la toma de decisiones de la organización.

Es decir, a través de un modelo de predicción se pretende anticipar cuáles reservas van a cancelarse de manera de poder actuar comercialmente, buscando alternativas de marketing para evitar que dicho cliente cancele la reserva, o utilizando el espacio para otro cliente. Se busca encontrar cuáles son los métodos de aprendizaje automático de clasificación que ayudan a predecir con mayor precisión la cancelación de reservas, a la vez

de tratar de reconocer cuáles son los factores o variables que más influyen en las cancelaciones.

Las organizaciones se encuentran con desafíos constantes en lo que respecta a la utilización, análisis y a la gestión de los grandes volúmenes de datos. El sector hotelero, al igual que otros del ámbito del sector público como privados, se topan con limitaciones para poder implementar Big Data tanto a nivel organizativo como cultural. En los últimos tiempos, se suma la responsabilidad que hay que tener en el manejo de los datos, desde el punto de vista de la preocupación por la seguridad, por la privacidad como también en el manejo ético de los datos.

A través del desarrollo de este análisis se busca mostrar que los métodos de clasificación de aprendizaje automático permiten predecir las cancelaciones de reservas de hoteles -en base a información del hotel y a datos cualitativos de los clientes y de los hoteles- con un alto AUC-ROC -que representa la probabilidad de que un clasificador ordene una instancia positiva elegida aleatoriamente más alta que una negativa, utilizado generalmente para la comparación de modelos de aprendizaje automático-.

El objetivo general del presente trabajo es desarrollar un modelo para predecir con mayor precisión las cancelaciones de reservas para un conjunto de hoteles ubicados en Argentina, comparando la capacidad predictiva de distintos algoritmos de clasificación.

Para cumplir con el objetivo general, se plantean los siguientes objetivos específicos: Analizar el manejo de grandes volúmenes de datos en el sector hotelero, Recopilar y preparar los datos para el armado del modelo, Desarrollar diferentes modelos y Analizar cuáles son las características que mejor explican la cancelación de las reservas y cuál es el modelo que predice con mayor AUC-ROC la cancelación de las reservas.

Para cumplir con el objetivo de analizar el manejo de grandes volúmenes de datos en el sector hotelero en un primer apartado se realizará un recorrido por la literatura para definir que es Big Data y cómo es su incorporación en los negocios, además se señalan las diferentes aplicaciones de Big Data y técnicas de aprendizaje automático en el sector hotelero y se analiza el caso de predicción de cancelaciones de reservas y su utilización en el manejo de Gestión de la Demanda.

En un segundo apartado se abordará la parte metodológica y el análisis de los datos utilizados para el desarrollo del modelo de predicción de cancelaciones, para lo que se explicará cómo se obtuvieron los datos provenientes de la base de reservas de la cadena de hoteles, se describirán los datos y cómo fueron preparados y se realizará la estadística descriptiva y análisis de los parámetros.

Por último, en el tercer apartado se comparan los diferentes algoritmos para predecir la cancelación de reservas, se selecciona el modelo que mejor performance tiene y se establecen algunos lineamientos para la implementación del mismo en la organización.

1. El manejo de grandes volúmenes de datos en el sector hotelero

1.1.Definición de Big Data y su incorporación en los negocios

El término Big Data está muy de moda, sin embargo, la mayoría de las veces no se usa de manera precisa y se refiere a diferentes conceptos entre los que se incluyen grandes volúmenes de datos, datos en tiempo real, datos provenientes de diferentes fuentes como redes sociales, clics, páginas web, sensores, internet de las cosas, etc., generando confusión.

Existen diferentes definiciones en la literatura, Kusnetzky (2010) explica a Big Data como el conjunto de herramientas, procesos y procedimientos que permiten a una organización crear, manipular y administrar grandes conjuntos de datos e instalaciones de almacenamiento. Se aplica el término también a la información que no puede ser procesada o analizada mediante procesos tradicionales.

McAfee y Brynjolfsson (2012) caracterizan a Big Data como el gran volumen de información, que se genera a gran velocidad y de gran variedad, que requiere nuevas formas de procesamiento para poder extraer conocimientos, optimizar procesos y guiar la toma de decisiones (Provost y Fawcett, 2013).

Por otro lado, un estudio entre IBM y la Escuela de Negocios Saïd de la Universidad de Oxford (Miele y Shockley, 2013), señala que una manera útil de caracterizar las tres dimensiones de Big Data (llamado las tres V) son: volumen, variedad y velocidad. La característica que se asocia con mayor frecuencia a Big Data es el volumen y hace referencia a las cantidades masivas de datos que las organizaciones intentan aprovechar para mejorar la toma de decisiones. Los volúmenes de datos continúan aumentado a un ritmo sin precedentes, cada día las empresas registran un aumento significativo de sus datos (terabytes, petabytes y exabytes) creados por personas y máquinas.

En relación con cuál es el tamaño que es considerado un gran volumen de datos, de acuerdo con el mencionado estudio de IBM (Miele y Shockley, 2013), que reúne las conclusiones sobre una encuesta realizada a 1.144 profesionales procedentes de 95 países y 26 sectores, más del 50% de los encuestados consideran que conjuntos de datos de entre

un terabyte y un petabyte es Big Data, mientras que otro 30% simplemente no sabía cuantificarlo.

En lo que respecta a la Variedad, se refiere a diferentes tipos y fuentes de datos y cómo gestionar la complejidad de múltiples tipos de datos -datos estructurados, semiestructurados y no estructurados-. Los datos que se generan presentan muchas formas entre las que se destacan texto, datos web, tuits, datos de sensores, audio, vídeo, secuencias de clics, entre otros. Las organizaciones buscan resolver la complejidad de integrar y analizar datos provenientes de distintas fuentes tanto internos de la propia empresa, como externos.

Por último, la característica de Velocidad se refiere a que los datos están en continuo movimiento. La velocidad a la que se crean, procesan y analizan los datos está aumentando, lo cual es un desafío para las empresas que buscan contar con los datos en tiempo real, e incorporar los mismos a los procesos de negocio y a la toma de decisiones.

Sin embargo, tanto el estudio de IBM (Miele y Shockley, 2013) como otros autores incorporan una cuarta dimensión para caracterizar el fenómeno Big Data que es la Veracidad, dado que se introduce cierto grado de incertidumbre durante el proceso de adquisición, proceso y análisis de grandes volúmenes de datos. La calidad de los datos es un requisito importante y un reto fundamental que tienen las organizaciones que quieren incorporar herramientas de Big Data para la toma de decisiones.

Para IBM el uso de Big Data crea una oportunidad para que las organizaciones puedan obtener una ventaja competitiva. Desde el punto de vista de cambiar la manera en que se relacionan con sus clientes, cómo les prestan servicio o incluso solucionando problemas que antes se desconocían. Aunque no todas las organizaciones tienen las mismas posibilidades de implementar herramientas de Big Data, se señala que en todos los sectores existe la posibilidad de utilizar las nuevas tecnologías y analíticas de Big Data para mejorar la toma de decisiones y el rendimiento.

Existen varios estudios y autores que se centran en analizar los efectos que tiene la incorporación de Big Data en las empresas. En este sentido, se destaca el estudio llevado a cabo por el MIT *Center for Digital Business* y recogido por la Harvard Business Review (McAfee y Brynjolfsson, 2012). El mencionado trabajo aportó evidencias empíricas de la

importancia de la explotación de Big Data para diferentes sectores, a través de una encuesta a 330 empresas en relación con sus prácticas gerenciales en materia organizacional y tecnológica.

McAfee y Brynjolfsson (2012) muestran que la toma de decisiones basada en datos mejora el desarrollo del negocio. Las empresas que se consideran basadas en datos muestran mejor performance en cuanto a medición de objetivos y resultados operacionales. Las compañías que se encuentran en el top tres de su industria en el uso de datos para la toma de decisiones, son 5% más productivas y 6% más rentables que sus competidores, en promedio.

Los autores resaltan dos casos en los que el uso de Big Data impactó fuerte en las empresas que fueron pioneras en su implementación. Uno de ellos, fue la experiencia de una de aerolínea, para la elaboración de pronósticos acertados en el tiempo estimado de llegada de los aviones -ETA- y la importancia que tuvo para el negocio. Una de las aerolíneas más grandes de Estados Unidos, a través de un estudio, encontró que en aproximadamente el 10% de sus vuelos, en el *hub* más importante, había un gap de 10 minutos entre el ETA y el tiempo de llegada efectivo (y que un 30% de los vuelos tenía un gap de menos de 5 minutos). En base la información obtenida de los datos es que decidieron accionar. El uso de los datos llevó a que pudieran realizar mejores predicciones y las mejores predicciones llevaron a mejores decisiones, resultando en un ahorro de millones de dólares.

El estudio realizado conjuntamente por IBM *Institute for Business Value* y la Escuela de Negocios Saïd en la Universidad de Oxford (Miele y Shockley, 2013) basado en el uso de Big Data en el mundo real con las empresas más innovadoras, muestra estado de avance de la incorporación de Big Data y *Data Analytics* en las organizaciones. Las empresas más innovadoras extraen valor de datos inciertos, muestra que la mayor parte de las empresas en 2012 se encontraban en las primeras fases del desarrollo de Big Data, la mayoría de ellas centradas en comprender los conceptos (24%) o definir una estrategia relacionada con Big Data (47%) y el 28% de los encuestados trabaja en empresas de a en las que están desarrollando pruebas de conceptos o ya implementaron soluciones a escala.

Más del 50% de los encuestados consideran que Big Data proporciona la capacidad para comprender y predecir mejor los comportamientos de los clientes. Los encuestados clasifican a los objetivos funcionales para Big Data en sus empresas, de mayor a menor

importancia, los siguientes: resultados centrados en el cliente, optimización operativa, gestión financiera/de riesgos, nuevo modelo empresarial y colaboración de los empleados.

De acuerdo con ese análisis, el 58% de las empresas que señalan que pusieron en marcha iniciativas de Big Data cuenta con procesos de seguridad y de *data governance*. Sin embargo, resaltan que estos dos aspectos son muy importantes dado que hay que tener en cuenta las nuevas consideraciones éticas, de privacidad, y posiblemente normativas y jurídicas. La utilización de Big Data introduce nuevos riesgos y compromisos para las empresas, ya que algunas organizaciones perdieron el control sobre los datos o los utilizaron de formas cuestionables.

Por otro lado, el trabajo de la OBS Business School sobre la situación de Big Data 2018-2019 señala que el 71% de las empresas a nivel global creen que su apuesta por Big Data se va a acelerar en los próximos tres años. El 57% de las empresas tienen la figura del *Chief Data Officer* (CDO), liderando y siendo la piedra angular en la ayuda de la integración, la democratización de los datos y capacidad de análisis a lo largo de toda la organización. Otro de los datos interesantes es que el 52% de las empresas están aprovechando el potencial del Big Data y el análisis de datos predictivo para proveer mejores ideas y más inteligencia a sus procesos operativos.

También muestra que las empresas que adoptan una cultura basada en datos y llegan a implantarla están alcanzando mayores ingresos. El 57% de las organizaciones que tienen implantadas esa cultura tienen una ventaja competitiva, que se traduce en una mayor rapidez y efectividad en el proceso de toma de decisiones. El 63% de las organizaciones considera que la cultura *Data-Driven* ayuda a mejorar la eficiencia y la productividad.

Según el estudio, el nivel de implantación de estrategias Big Data está siendo cada año más alto y el nivel de confianza de los empleados sobre estas iniciativas también. Los resultados de las encuestas muestran que el nivel de adopción está en 70-80%. Pero con relación a cómo ven a su empresa en gobernanza y el manejo de datos, los resultados son más bajos alcanzando el 64% en 2018.

Los resultados del 2018 señalan que el 60% de los encuestados considera que el uso que se le está dando a Big Data es para impulsar procesos y reducir costos. En un segundo

lugar, el 57% están utilizando Big Data como estrategia que les ayude a lanzar el cambio en sus organizaciones y adquirir una posición de ventaja frente a la competencia. Las oportunidades que está brindando Big Data a las organizaciones va desde mejorar la gestión de sus riesgos, mejorar la productividad, obtener más y mejor información sobre el mercado, mejorar la relación y retención de los clientes, mejorar sus productos y servicios y mejorar sus rendimientos financieros.

1.2.Aplicaciones de Big Data y técnicas de aprendizaje automático en el sector hotelero

Las organizaciones cuentan cada vez más con grandes volúmenes de datos -o con la capacidad de generarlos y recolectarlos-, de gran variedad y que se generan a mucha velocidad. A través del procesamiento de estos datos mediante algoritmos, las empresas buscan generar información como factores de decisión para optimizar procesos, mejorar resultados y/o generar nuevos negocios. El aumento de la capacidad de análisis y de darle valor a los datos fue creciendo junto con la capacidad de almacenamiento y procesamiento de los mismos.

Los hoteles no se encuentran aislados de esta situación y existe un gran potencial a partir del uso de Big Data para producir información que sea pertinente, de calidad y oportuna, para fundamentar y orientar decisiones; o para diagnosticar problemas que no se conocían y entonces era imposible accionar.

En la actualidad los hoteles poseen una variedad y cantidad de datos de diferentes fuentes: datos que puede obtener y recolectar el mismo hotel de clientes y reservas, información de las redes sociales, hasta datos externos como la previsión meteorológica, de la competencia, etc.

Las empresas del sector hotelero que pueden integrar Big Data en sus negocios tienen la ventaja competitiva que las que no -con una probabilidad del más del 55%- de acuerdo con Gupta y otros (2017).

Las posibles aplicaciones de Big Data en el sector hotelero son muy amplias entre las cuáles se destacan: políticas de marketing personalizadas (Marr, 2016), precios dinámicos y en tiempo real (Huet, 2015 y Marr, 2016), predicción de la ocupación de los

hoteles (Yang y Pan, 2017), explotación de datos de redes sociales y revisión de comentarios de *feedback* de clientes (Song y Liu, 2017), mejora de conocimiento y satisfacción de clientes (Marr, 2016), gestión de stock (Waller y Fawcett, 2013), hasta en el aumento de la eficiencia energética (Kahn y Liu, 2016).

En muchos casos, el propio hotel aún no utiliza ni genera valor a los datos propios. Es por ello que se recomienda comenzar por organizar los datos existentes previo a incorporar otros puntos de datos. Marr (2016) señala que, en este sector, si bien el uso de Big Data es relativamente reciente, resulta determinante para identificar las características y necesidades heterogéneas de cada cliente, y poder cumplir con sus expectativas particulares para conseguir fidelizarlos, creando así una ventaja competitiva frente a la competencia. El autor también establece que Big Data permite a los hoteles obtener información para tomar decisiones en tiempo real, anticiparse a las necesidades de los clientes, y calcular el precio óptimo de las habitaciones, lo que les permite situarse en una posición privilegiada en el mercado.

En este sentido, Magnini, Honeycutt y Hodge (2003) analizaron la importancia que tiene la extracción e interpretación de datos de clientes en el sector hotelero. El conocimiento de los clientes - de dónde son, cuanto, cuando y en qué gastan su dinero - puede ayudar a los hoteles a implementar mejores estrategias de marketing y maximizar los beneficios.

El proceso de extracción de datos y el análisis permite identificar patrones de comportamiento significativos y construir modelos predictivos de la conducta del cliente, clasificados por segmentos.

El uso de Big Data en hoteles puede ofrecer la evidencia basada en datos necesaria para tomar decisiones en base a números y análisis, más que en anécdotas, intuición o en su experiencia pasada (Frederiksen, 2012), lo que puede conducir a establecer análisis más precisos, una toma de decisiones más segura, mayor eficiencia operativa y a una reducción de costos y de riesgos (De Mauro et al., 2015).

Gupta, Gauba, Jain (2017) señalan que las herramientas de Big Data mejoran la calidad del servicio, así que podría llegar a implementarse en casi cualquier sector y que un ejemplo de esto es en el sector hotelero. Entre las oportunidades que brinda se puede

desatacar la mejora continua de la atención al cliente y servicios que se le puede ofrecer a medida que más se lo conoce, hacer más eficiente sus operaciones, crear mejores estrategias de marketing (saber el momento indicado, a quien dirigir la campaña y a qué responde cada segmento de clientes) y a partir de todo esto, aumentar las ganancias del hotel.

Los autores Gupta y otros (2017) mencionan que Big Data está transformando varias áreas del sector de alojamientos entre las que resaltan las siguientes: la mejora de la experiencia del cliente a través de un trato personalizado (incluso el hotel puede analizar sus gustos en la comida, o los patrones de gasto), estrategias de marketing (con las técnicas correctas de Big Data el hotel puede hacer campañas a través de diferentes canales cambiando si es necesario en tiempo real, y así darle visibilidad al hotel incluso en lugares que no creía que se podía llegar); optimización de costos y capacidad (los hoteles pierden dinero si no logran calcular de manera precisa la demanda para sus servicios).

1.3. El caso de predicción de cancelaciones de reservas y su utilización en la Gestión de la Demanda

Hasta el momento el sector hotelero no se ha enfocado mucho en el uso de Big Data para mejorar la experiencia del cliente y la operativa del hotel -de acuerdo con datos de la empresa STR que analiza y brinda información de los mercados a sectores internacionales de la industria de la hospitalidad[3]-, pero sí existe cierto avance en su utilización en estrategias de gestión de la demanda.

Las cancelaciones tienen un fuerte impacto en la industria del turismo en general (para las aerolíneas, las empresas de alquiler de autos y agencias de viaje), y particularmente para la industria hotelera. Las mismas tienen efecto tanto en los costos como en las decisiones de asignación de recursos y habitaciones y de manejo de la demanda. Por eso una aplicación importante de Big Data en el sector hotelero es para la creación de modelos para predecir dichas cancelaciones, de manera de construir pronósticos de demanda más certeros y desarrollar así mejores sistemas de gestión de la demanda -*Revenue Management*-, para perfeccionar la política de precios, optimizar la capacidad y la disponibilidad de los hoteles, mejorar la estrategia de marketing y la política de cancelaciones, entre otras razones.

[3] https://www.hosteltur.com/126734_debilidades-hoteles-su-estrategia-big-data.html

La gestión de la demanda -o *revenue management*- se define como la aplicación de sistemas de información y estrategias de precios para vender el producto justo, al cliente justo, al precio justo en el momento indicado (Kimes, Wirtz, 2003), permitiendo mejorar las ganancias de los hoteles mediante la diferenciación de precios y el manejo adecuado de la disponibilidad de las habitaciones. Es una práctica que empezó a desarrollarse en sus comienzos para la industria aerocomercial en la década del ‘60 y posteriormente fue adaptándose a otras industrias de servicios (Antonio, de Almeida, Nunes, 2017). La definición de *revenue management* fue readaptada para el caso de la industria hotelera como: tener disponibilidad en la habitación justa, para el cliente justo, al precio justo, por el canal de distribución justo y en el momento indicado (Mehrotra, Ruttley, 2006).

El ejemplo de elaboración de mejores y más precisas predicciones por parte de las aerolíneas y su traducción a tomar acciones en el negocio es uno de los tantos casos donde se muestra el potencial del uso de Big Data en los negocios y en la toma de decisiones, y ocurre lo mismo para las cancelaciones en el sector hotelero. La cancelación de reservas afecta negativamente a la producción de pronósticos precisos, por lo que es necesario entender que es una herramienta crítica en la industria hotelera.

Como señalan Antonio, de Almeida y Nunes (2019) muchas investigaciones muestran que con el avance tecnológico computacional y los algoritmos de aprendizaje automático posible construir modelos para predecir la cancelación de reservas. Las predicciones precisas son una herramienta indispensable para asegurar un buen desenvolvimiento del *revenue management*.

De acuerdo con Antonio, de Almeida, Nunes (2017) varios estudios abordan temas relacionados con los métodos empleados para tratar de morigerar las consecuencias de las cancelaciones en las ganancias y asignación de plazas, en las políticas de cancelación y en las estrategias de sobreventa. Según los autores, en los comienzos estos estudios fueron desarrollados para aerolíneas, pero ahora están incrementándose para el caso de la industria hotelera. Sin embargo, en su mayoría son modelos basados en metodologías de estadísticas tradicionales y muy pocos aprovechan las ventajas de técnicas de aprendizaje automático.

También señalan que la mayoría de los estudios sobre la predicción de la cancelación de reservas son planteados como problemas de regresión. Solo unos pocos lo abordan como un problema de clasificación y ponen el foco en predecir la tasa de cancelación global en vez de la confianza con la que se va a cancelar cada una de las reservas. Antonio, de Almeida y Nunes (2017) mostraron que se puede predecir con elevada precisión -*Accuracy*- la cancelación de las reservas.

En base a estas predicciones, puede calcularse la demanda neta que surge de deducir de la demanda del hotel la suma de todas las reservas con alta confianza a ser canceladas. Contar con datos precisos y confiables ayuda a que las organizaciones avancen hacia la toma de decisiones basadas en datos, o como es definido por McAfee y Brynjolfsson (2012) -*data driven decision making*- referido a basar las acciones del negocio en evidencias y no en la intuición.

Por otro lado, las técnicas de minería de datos, definidas como el proceso de selección, investigación y modelado de grandes volúmenes de datos para descubrir regularidades y relaciones que en principio son desconocidas, con el objetivo de obtener resultados claros y útiles (Giudici, 2003), permiten realizar el análisis de datos y descubrir patrones importantes en los mismos, lo que contribuye en gran medida a mejorar estrategias comerciales, ampliar las bases de conocimiento y de la investigación científica (Han, Kamber, 2011). En base a esto es que resulta esencial la aplicación de minería de datos en las organizaciones dado que la toma de decisiones basada en datos mejora el desarrollo del negocio.

En línea con esto, las organizaciones cuentan cada vez más con grandes volúmenes de datos -o con la capacidad de generarlos y recolectarlos-, de gran variedad y que se generan a mucha velocidad. A través del procesamiento de estos datos mediante algoritmos, las empresas buscan generar información que sirvan como factores de decisión para optimizar procesos, mejorar resultados y/o generar nuevos negocios. Los hoteles no se encuentran aislados de esta situación y existe un gran potencial a partir del uso de Big Data para producir información que sea pertinente, de calidad y oportuna, para fundamentar y orientar decisiones; o para diagnosticar problemas que no se conocían y entonces era imposible accionar. El aumento de la capacidad de análisis y de darle valor a los datos fue creciendo junto con la capacidad de almacenamiento y procesamiento de los mismos.

A través de minería de datos, que es el análisis de conjuntos de datos -principalmente de grandes volúmenes- para el descubrimiento de patrones y para resumir los datos de formas novedosas que sean comprensibles y útiles para el usuario de los datos en diferentes organizaciones (Hand, Mannilla y Smith, 2001), se busca predecir las cancelaciones.

Los modelos de minería de datos utilizan técnicas estadísticas en el modelado y algoritmos para su desarrollo. El proceso mediante el cual a través de minería de datos se busca obtener información para la toma de decisiones abarca desde la parte de recolección, selección, limpieza e integración de datos, incluyendo el análisis y validación del modelo pudiendo utilizar como parte del proceso diferentes lenguajes de programación y herramientas de software.

Como ya fuese mencionado en varias oportunidades, las cancelaciones tienen un fuerte impacto en la industria del turismo en general -incluyendo las aerolíneas, empresas de alquiler de autos -, y particularmente para la industria hotelera, tanto en los costos como en las decisiones de asignación de plazas y de manejo de la demanda.

Es en esta razón que resulta de gran importancia la creación de modelos para predecir dichas cancelaciones, de manera de poder construir mejores pronósticos de demanda, mejorar la política de precios, optimizar la capacidad y la disponibilidad de los hoteles, mejorar la estrategia de marketing y la política de cancelaciones de los hoteles, entre otras. Durante los últimos años, las cancelaciones de los hoteles se vieron incrementadas por el uso de diferentes canales de venta en línea y a través de distintas plataformas. Las cancelaciones pueden representar el 20% del total de las reservas realizadas en hoteles (Morales and Wang, 2010) y de acuerdo con datos de D-EDGE, proveedor de tecnologías de distribución hotelera para el sector turístico, la tasa de cancelación de reservas global en hoteles alcanzó el 40% en promedio en 2018.

Los hoteles cuentan con una serie de canales de comercialización, mediante los cuales los consumidores puedan realizar las reservas: a) el canal directo (vía telefónica, en oficinas propias, recepción del hotel, etc.); b) a través de las OTAs (*Online Travel Agencies*); c) a través de agencias de viaje tradicionales; d) de manera online en sitios propios. La

posibilidad de reservar por adelantado y mediante nuevos canales ampliaron las razones por las que puedan generarse cancelaciones.

Adicionalmente a los motivos clásicos de cancelación, tales como cambios en las reuniones de negocio, enfermedades, cambios en las vacaciones o cuestiones meteorológicas, se incorporaron nuevas razones que incluyen clientes que continúan buscando mejores ofertas de hospedaje aún después de haber reservado una habitación. Esto incluye la búsqueda de una mejor opción -un hotel que les brinde mejores servicios- o en algunos casos, incluso se reservan habitaciones en más de una locación, para luego seleccionar una de ellas cerca de la fecha del *check-in*. Dichas búsquedas posteriores al momento de la reserva también se vieron potenciadas por la proliferación de sitios donde los hoteles ofrecen cancelación gratuita de manera rápida y segura online, lo cual obliga a los hoteles a considerar las potenciales cancelaciones en su proyección de demanda de manera que esto no afecte el negocio ni sobreestime el uso futuro de instalaciones o personal.

De acuerdo con la información de reservas todos los hoteles de la cadena bajo estudio, se observa que entre el 16,2% de las reservas se cancelaron durante el período de junio de 2018 a octubre 2019.

Existen maneras de intentar disminuir el riesgo de las cancelaciones para los hoteles, como ya se mencionó anteriormente, entre las que aparecen la sobreventa -*overbooking*- o endurecer las condiciones de devolución de las reservas. En un contexto de oferta amplia y de fácil acceso mediante las OTAs, endurecer las condiciones de devolución dificulta la venta del servicio. La implementación de condiciones duras para las cancelaciones podría generar el efecto opuesto al buscado, disminuyendo las reservas y bajando el nivel de ocupación del hotel.

Por otro lado, un mal manejo del *overbooking*, al implementar un sistema de sobreventa de habitaciones sin la suficiente información, podría causar un daño aún mayor al hotel, dado que podría afectar la reputación del hotel en caso. Este último punto debe tomarse con mayor relevancia en esta industria en donde la imagen de un hotel en internet -tanto en buscadores de metadatos, como comentarios de usuarios en sitios de turismo y redes sociales- tienen un fuerte impacto en la toma de decisiones del consumidor.

Un hotel con menores ventas, en algunos casos, puede afectar menos financiera como comercialmente que uno sobrevendido, que afecte negativamente en la reputación global de la marca -en caso de una cadena- y del hotel en particular. Existen muchos estudios realizados en diferentes países sobre la predicción de cancelaciones de hoteles para *revenue management*, para disminuir la incertidumbre, entre varias aplicaciones. La realización de un modelo de predicción de cancelaciones a través de un modelo supervisado podría realizarse de varias maneras, entre ellas: a) a través de la predicción del porcentaje de cancelaciones respecto del total de reservas actuales, o b) mediante la clasificación de la reserva, que implica tratar de entender cómo se comportan las variables, ver si los clientes pueden segmentarse en diferentes clústeres y así predecir si una reserva especifica va a cancelarse o no, se busca predecir la clase de cada una de las reservas.

Gupta y otros (2018) señalan que muchos hoteles presentan problemas al intentar incorporar Big Data. En la actualidad, una de las principales limitaciones con las que se enfrenta la industria es la seguridad y ética del manejo de los datos, lo cual en muchos casos limita la cantidad de actores dispuestos a incorporar *big data* en su negocio. Asimismo, muchas empresas del sector de acuerdo con datos de la consultora Deloitte, se encuentran ocupados por cumplir con la legislación del Reglamento General de Protección de Datos (GDPR) de la Unión Europea, cuyas reglas se establecieron en Europa, pero tienen un impacto global, y cuyo incumplimiento puede incurrir en elevadas multas que comienzan con el 4% de la facturación anual de la compañía.

El estudio de IBM (Schroeck, Shockley, Smart, Romero-Morales y Tufano, 2012) muestra que existen distintos desafíos que obstaculizan la adopción de Big Data y que difieren a medida que las empresas avanzan a lo largo de las cuatro diferentes fases de adopción de Big Data, que se enumeran a continuación:

1) Educar. Es la fase que se centra en la concientización y el desarrollo de conocimiento. La mayoría de las empresas que se encuentran en esta fase están estudiando las posibles ventajas de las tecnologías y la analítica de Big Data e intentando entender cómo puede ayudarles a abordar importantes oportunidades de negocio en sus propios sectores o mercados.

2) Explorar. Es la fase en la cual las empresas se enfocan en desarrollar la hoja de ruta de la empresa para el desarrollo de Big Data para un caso de negocio cuantificable. Se tienen en cuenta los datos, la tecnología y las habilidades existentes y, a continuación, se establece dónde comenzar y cómo desarrollar un plan en consonancia con la estrategia de negocio de la empresa.

3) Interactuar. En esta etapa las empresas empiezan a comprobar los beneficios que tiene para el negocio el uso de Big Data, así como a llevar a cabo una valoración de sus tecnologías y habilidades. Se trabaja dentro de un ámbito definido y limitado para comprender y probar las tecnologías y habilidades necesarias para aprovechar nuevas fuentes de datos.

4) Ejecutar. Es la última fase donde el nivel de operatividad e implementación de las funciones analíticas y de Big Data es mayor.

McAfee y Brynjolfsson (2012) muestran que las empresas -al querer incorporar estrategias de Big Data- se encontraban con problemas de acceso a los datos, de gestión del volumen, limitaciones en las capacidades de infraestructura y de falta de personal capacitado para las nuevas necesidades.

El trabajo de la OBS Business School muestra que, durante el año 2018, las empresas se encuentran con otros inconvenientes, como las limitaciones a nivel de silos funcionales que impiden que las iniciativas puedan tener un recorrido a lo largo de toda la organización. Además, señalan que no se logra generar estructuras de trabajo colaborativo de carácter multidisciplinario dentro de las organizaciones, ya que tanto la estrategia organizativa como también la operativa siguen bajo una filosofía tradicionalista de carácter vertical. Muestra que tan solo el 21% de las organizaciones disponen de una arquitectura sofisticada e integrada con los marcos de seguridad y gobernanza de los datos de la organización.

A medida que las organizaciones avanzan en la implementación de sus estrategias de Big Data y aprovechan más sus datos para fortalecer sus negocios es que se van encontrando con diferentes barreras. El informe resalta tres desafíos que afectan a todas las organizaciones en la actualidad: la privacidad de los datos y las preocupaciones de seguridad; la democratización limitada de los datos; y la falta de capacitación sobre cómo aprovechar

el potencial de los datos. Hacen hincapié, además, en la relevancia que está tomando para las organizaciones la gobernanza de datos y que es necesario que se adopte un cambio cultural para implementarlo. El 45% de las organizaciones establece que menos de la mitad de sus datos está siendo gobernado (con algún tipo de autoridad certificada, siguiendo políticas corporativas y teniendo una única versión).

Ese estudio también muestra que las organizaciones con un nivel de madurez más alto en la implementación de Big Data están trabajando en resguardar la privacidad del dato, mientras que aquellas empresas que todavía están en vías de desarrollo o en procesos de inicio de incorporación de una estrategia de Big Data, están encontrando más problemas en cómo almacenar y hacer útil los datos y, en no tener clara una estrategia de gestión y explotación del dato.

Li, Xu, Tang, Wang y Li (2018) señalan que las desventajas del uso de Big Data en las investigaciones del sector de turismo son: la preocupación por la calidad del dato, el costo de los datos y las cuestiones de privacidad. En lo que respecta a la calidad del dato, por ejemplo, en el caso del procesamiento de texto de lo publicado online (ya sea en redes sociales comentarios de páginas web) es difícil confiar en los mismos, teniendo cuenta que hay *fakes reviews*. También establecen que los datos provenientes del procesamiento de imágenes o *roaming* son menos confiables que los datos de GPS, o que las búsquedas en Google pueden estar sesgadas.

Sobre el costo de los datos, los autores destacan que obtener información de sensores o GPS es caro, sin embargo, el 58% de las aplicaciones de Big Data en turismo utilizan las búsquedas en páginas webs, ya que es mucho más económico.

Por último, señalan que la preocupación sobre la privacidad es un tema tanto de las agencias de viaje como de los hoteles, los turistas, los sectores de gobierno, entre otros. El uso de la mayoría de los dispositivos (*roaming*, datos de WIFI, bluetooth) están siendo trabajados con cuidado debido a que implican datos privados que revelan los movimientos de los turistas. Lo mismo sucede con los datos transaccionales, como los de tarjetas de crédito.

Cabe mencionar, que existen casos de robo de información masiva en el sector, como lo sucedido a la cadena Marriott, que fue una de las mayores filtraciones de datos corporativos. Alrededor de 327 millones de personas se vieron afectadas con el robo de datos de las reservas (que incluyó nombres, números de teléfonos, correos electrónicos, números de pasaportes y fechas de nacimientos) y en el caso de otros millones de clientes también se robaron los datos de las tarjetas de crédito. Todo esto puede ser usado para robo de identidad, apertura de cuentas bancarias, obtener tarjetas de crédito, en muchos países hasta préstamos, afectando negativamente la reputación de la cadena, además del costo económico.

2. Metodología y datos utilizados para el desarrollo del modelo de predicción de cancelaciones

2.1. La cancelación de las reservas en la industria hotelera y su implicancia en la gestión de la demanda

Una reserva representa un contrato entre dos partes, por un lado, el establecimiento hotelero y, por el otro, el cliente. La misma le da a este último, el derecho de utilizar el servicio del hotel en un futuro a un precio pactado. Usualmente, los hoteles brindan la opción de cancelar la reserva antes del *check-in*. Esa alternativa de cancelar el servicio antes de que sea provisto, significa que los hoteles tienen que garantizar habitaciones tanto para los huéspedes que cumplen con la reserva como soportar el costo de tener vacante la habitación cuando una reserva se cancela o el cliente no se presenta a la prestación de un servicio previamente contratado -*no show*-. En esos casos, se deposita todo el riesgo en los hoteles.

La industria hotelera fue generando diversas herramientas para paliar las pérdidas económicas por las cancelaciones como el endurecimiento de las condiciones de cancelación y la implementación de la sobreventa de habitaciones - *overbooking* -.

De acuerdo con la investigación realizada por Riasi (2018) los resultados indican que los hoteles que implementan la sobreventa tienen una mejor performance en relación con aquellos que no sobrevenden habitaciones.

La venta excesiva que se realiza sobre un servicio con el objeto de garantizar la plena ocupación debe hacerse sobre proyecciones precisas sobre cancelaciones; en caso contrario estas estrategias pueden impactar negativamente en las finanzas y/o en la imagen de la organización. Una sobreventa puede causar que el hotel tenga que negar la provisión del servicio a un cliente al no tener disponibilidad de habitaciones, esto provoca una relación de descontento por parte del cliente hacia el hotel – que pone en la obligación del hotel de recompensar de alguna manera al cliente- y genera costos al tener que afrontar el gasto de la estadía en otro hotel alternativo. Esta relocalización además puede llegar a significar pérdidas futuras de reservas, ya que se da a conocer otro hotel que el cliente puede llegar a disfrutar.

Por otro lado, Chen y otros (2011) y Smith y otros (2015) señalan que las políticas de cancelaciones duras reducen las ganancias de los hoteles no sólo por tener que aplicar precios de descuento para estas tarifas sino también porque reducen la cantidad de reservas, particularmente las no reembolsables y las políticas de cancelación hasta 48 horas antes del *check-in*.

En la actualidad, las *Online Travel Agencies* - OTAs - han facilitado la cancelación de reservas. Además, al brindar la opción de cancelar sin costo alguno, se fomenta la reserva de varios hoteles en un mismo destino y con mucha antelación, para decidir más cerca de la fecha sin tener que preocuparse por no encontrar una alternativa en el destino del viaje. Muchos clientes reservan por las dudas, y al cancelar las reservas cerca de la fecha de viaje impactan significativamente en las ganancias de los hoteles. Estas OTAs tienen estrategias de venta que pueden llegar a influir en el comportamiento de los clientes al mostrar en sus páginas web que sólo queda una habitación disponible en ese hotel, o que cierta cantidad de personas reservaron últimamente ese hotel. Tienen como objetivo generar que los clientes reserven con cierta urgencia y sin pensarlo demasiado, intentando quitarle racionalidad al proceso de compra.

Esto hace que existan diferencias en los hoteles entre las reservas en base al origen de estas, dado que las que se reciben a través de una OTA muchas veces no van a concretarse en función de que muchas veces no existe una intención real en hospedarse y la reserva fue hecha de manera preventiva. Todo lo anteriormente mencionado es un reto para la industria, teniendo que predecir las reservas que tienen mayor probabilidad de ser canceladas y cuales

no para poder planificar la ocupación, los precios y ventas, y mitigar las pérdidas a las que pudieran incurrir con las cancelaciones.

La gestión de la demanda se torna un factor relevante en lo que respecta al manejo de los ingresos de una organización, y la parte relativa a optimizar los ingresos de la organización es considerada una de las áreas de aplicación operativa más exitosa de gestión de demanda. A pesar de haber comenzado como una práctica especializada en la gestión de tráfico aéreo en la década de 1960, con el paso del tiempo se fue extendiendo hacia otras ramas del sector de turismo y viajes, incluyendo la gestión hotelera.

Talluri y Van Ryzin (2005) explican que la gestión de la demanda es utilizada en industrias donde el manejo táctico de la demanda resulta esencial, y dado que la demanda presenta determinantes multidimensionales resulta complejo su tratamiento y su estimación. Alguna de las variables que pueden ser determinantes en la gestión de la demanda son: el tipo de producto vendido (tipo de habitación); el tipo de consumidor, sus preferencias y comportamiento de compra; el momento de la compra; los canales de distribución, entre otros.

Algunas ramas de la teoría económica tales como el racionamiento, la discriminación de precios y el monopolio proveen referencias que resultan relevantes en el manejo de los ingresos dado que brindan sustento a los conceptos básicos de determinación de precios (Ivanov, 2014), entre los que se encuentran la capacidad limitada, altos costos fijos con bajos costos variables, demanda volátil en el tiempo, capacidad de realizar estimaciones de demanda, capacidad de segmentación de mercados y diferenciada elasticidad de precios en distintos segmentos.

Algunos segmentos de la teoría económica son más teóricos que prácticos -por ejemplo, la elasticidad de la demanda, que permite entender que los consumidores tienen hoteles alternativos en caso de un aumento de precios del hotel, o que pueden decidir quedarse en el hotel por fidelización de marca a pesar de que otros hoteles ofrezcan servicios similares por menos valor-. Es por ello por lo que la integración de los datos de reservas hoteleras sumado a información cualitativa de los usuarios en un sistema computacional puede brindar información que de otro modo la empresa no lograría descifrar.

El *revenue management* explora este mundo multidimensional y busca optimizar los ingresos y resultados de las compañías al colaborar en la toma de decisiones estructurales, de precio, timing y cantidades a ofertar, lo cual involucra un proceso cíclico que permita la toma de decisiones y retroalimentación en base a los resultados obtenidos. Dentro de este contexto, el objetivo de predecir las cancelaciones de reservas en conjunto con modelos de *revenue management* resultan relevantes para optimizar los resultados. Para ello es donde las herramientas de predicción mediante ciencia de datos pueden resultar de utilidad.

El hecho de poder clasificar con cierta confianza las reservas más plausibles de ser canceladas, permite a los equipos responsables de manejo de demanda realizar accionables. Por un lado, pueden aplicar estrategias de sobreventa -liberando una proporción de esas potencialmente canceladas reservas-, o también el uso de medidas precautorias que permitan reducir la posibilidad de cancelación de dichas reservas. De las acciones que pueden tomar, se pueden listar la oferta de servicios adicionales, mejoras en el tipo de habitación reservada, descuentos o entradas a eventos locales, entre otras.

Otro factor relevante para el *revenue management* en el sector hotelero es el del cálculo de la demanda neta -esto es, del total de habitaciones reservadas, deducir aquellas clasificadas como cancelables- y a partir de allí establecer los requerimientos de personal, alimentos y uso de proveedores externos de una manera más eficiente, reduciendo a la mínima expresión la existencia de recursos ociosos cuya demanda por parte del hotel permite flexibilidad.

2.2. Obtención de la base de datos de reservas de hoteles y preparación de los datos

El presente análisis se realiza en base a los datos de reservas de una cadena de hoteles de 3 a 5 estrellas ubicados en la Argentina. Los mismos provienen de distintas bases, en distintos formatos y con distinta calidad, dependiendo del origen de dichas reservas. Es por eso que, como primer paso, el proceso de recopilación de información requiere unificar en una misma base los datos de cada una de estas diferentes fuentes - provenientes de las bases de datos primarias de los hoteles y bases de datos secundarias -, con información proveniente de fuentes externas -como es el caso de la valoración de los clientes en los sitios de evaluación y comparación de hoteles online -.

Cabe señalar que la información utilizada fue provista con consentimiento de la cadena de hoteles -como Anexo se adjunta el documento firmado que autoriza la utilización y manejo de la información- y que los datos se encuentran anonimizados para preservar la identidad y privacidad de los huéspedes. Asimismo, se decidió no revelar el nombre de la organización.

El presente trabajo se desarrolla en un marco de manejo responsable de la información. Los datos proporcionados por la organización son utilizados exclusivamente para realizar estudios estadísticos y analíticos con fines académicos. Para el desarrollo del proyecto los datos se almacenan de manera segura, habiendo tomado todas las medidas de precaución para proteger la información contra adulteraciones, pérdidas, consultas, uso o acceso no autorizado o fraudulento.

Según Steinmann, Adam Matei y Collmann (2016) las metodologías de *Big Data* tienen un potencial para mejorar los avances en investigaciones científicas, pero también pueden crear dilemas éticos significativos al respecto. Se presenta el desafío de cómo afrontar y reflexionar sobre los dilemas éticos en la utilización de Big Data, desde distintos aspectos y teniendo en cuenta su impacto en la privacidad de las personas, que abarca e integra además un conjunto de otros valores.

En la actualidad, los dilemas relacionados con la privacidad son probablemente los más importantes al trabajar con grandes volúmenes de datos. Asimismo, se sugiere poner énfasis en el contexto en el que se utiliza Big Data, porque de acuerdo a eso varían los cuidados a los que hay que enfrentarse y las diferentes preocupaciones que surgen de su utilización y del análisis. Los diferentes contextos en los que se puede usar son: sociales, gubernamentales, científicos y comerciales.

Steimann y otros (2016) plantean un marco teórico para reflexionar sobre las cuestiones éticas relacionadas con el uso de Big Data en investigaciones científicas. Para ello hay que tener en cuenta el desafío de las 4R: Reusar/Reutilizar, Readaptar/Reasignar, Recombinar y Reanalizar de los datos.

De acuerdo con los autores, para entender el potencial impacto que tiene el uso de Big Data, primero hay que conocer los diferentes manejos que pueden influir en la privacidad de las personas. Asimismo, proponen un modelo marco para el manejo ético de los datos, que es un análisis del tipo *trade-offs*, a través de una matriz de privacidad que contiene los diferentes contextos de uso y dimensiones de privacidad y los distintos trade-offs que se pueden dar.

Es importante ser cuidadosos en la utilización de Big Data en lo que respecta a los desafíos de las 4R que proponen Steimann y otros (2016), entender si los datos están siendo reutilizados, es decir, si fueron recopilados para un propósito específico y luego se están usando para otros fines en dominios comparables; o si están siendo reasignados, dado que se toman datos recopilados para un propósito específico y se usan y analizan para otros campos. Siguiendo con las 4R, hay que estar atentos en lo que respecta al recombinado de datos que puede llevar a la reidentificación de individuos a partir de datos que no contienen identificadores específicos o que han sido despojados intencionalmente de identificadores, y al reanalizado de los mismos, dado que se puede extraer nueva información de estos. El presente trabajo se realiza en el marco del manejo ético de los datos y teniendo en cuenta los desafíos de las 4R para la utilización de Big Data de Steimann y otros (2016).

En lo que respecta a la recolección de la información, se utilizan los siguientes conjuntos de datos:

- Datos de reservas entre el 19 de junio de 2018 y el 13 de octubre de 2019.

La base de datos de reservas fue provista por la cadena de hoteles e incluye información cualitativa y cuantitativa de cada reserva. Presenta un registro transaccional por cada reserva realizada, incluyendo en un campo el estatus más reciente de la misma (‘*New reservation*’, ‘*Modified reservation*’, ‘*No Show reservation*’ o ‘*Cancelled reservation*’). Asimismo, cuenta con información del id de reserva, id de cliente, el id del hotel reservado, la fecha de la reserva y del ingreso esperado al hotel, el tipo de habitación, la cantidad de adultos, niños y noches reservadas, el canal de ventas y el agente de ventas -si aplica- y si el cliente realiza la reserva por medios corporativos. Además, contempla el costo medio por habitación y el revenue esperado de dicha reserva.

- Base de datos de clientes

La base de datos de clientes también fue provista por la cadena de hoteles e incluye información de sus clientes, entre ellas: id_cliente, ciudad, país, fecha de nacimiento.

- Base de datos de clientes pertenecientes al programa de fidelización de la cadena de hoteles.

Esta base de datos fue provista por la cadena de hoteles y contempla los clientes adheridos a su programa de fidelización, junto con el estatus actual (gold, silver, etc.).

- Información de los hoteles

Tabla provista por la cadena de hoteles que incluye detalles cualitativos de los hoteles de su cadena, entre los cuales se cuenta con: id de hotel, cantidad de habitaciones, cantidad de restaurantes, si tiene sala de convención o no, en que ciudad se encuentra, etc.

- Datos de puntuación online de los hoteles.

Tabla generada a partir de información disponible en distintos portales de internet. Los datos recopilados fueron: cantidad de estrellas, puntuación en el sitio Booking.com, en Google.com y en Tripadvisor.com, relevado en octubre de 2019. En caso de implementarse el modelo, este dato debería ser relevado con una mayor frecuencia de modo de evaluar el impacto de cambios en el valor en la cancelación de reservas.

Como ya fuese mencionado, algunos de los datos provienen de las bases de datos primarias de los hoteles, mientras otros fueron extraídos de distintas páginas de internet. Algunas variables fueron creadas en función de los datos provistos por el sistema del hotel y otras se dedujeron en función de la información con la que se contaba. La estructura de los datos fue armada de manera tal de enriquecer lo más posible la información que presentaba la base de los hoteles, incorporando puntos de referencia con el objeto de que el modelo tuviera mejor poder predictivo, en función también de la literatura consultada.

Para la realización de este análisis se trabajó con la base de datos de reservas con un total de 121.922 registros. De los mismos, 18.484 registros (el 15,16% del total) corresponden a reservas que fueron canceladas mientras que 103.438 registros son de reservas que no fueron canceladas (el 84,84% de la base).

A los fines de este trabajo se contemplan como cancelación tanto las reservas canceladas en un momento previo al *check-in* como también aquellos casos en donde el cliente no se presenta el día del ingreso al hotel (*no show*).

Considerando que hay reservas que todavía no se concretaron por ser realizadas para una fecha de ingreso futura -es decir, plausibles a ser canceladas hasta la fecha en que se concrete el *check-in*-, se han caracterizado los registros con las siguientes categorías:

- Registros categoría A: reservas efectivas. Esto incluye aquellas reservas realizadas con fecha de ingreso previo o igual al 13 de octubre de 2019 que no fueran canceladas. Esto incluye reservas que a dicha fecha ya hubieran registrado el ingreso o *check-out.*
- Registros categoría B: reservas canceladas. Son aquellas reservas realizadas y luego actualizadas como canceladas o No Show.
- Registros categoría C: reservas con arribo futuro. Son aquellas reservas aún no canceladas con fecha de ingreso posterior al 13 de octubre de 2019. Esto significa que no se conoce con precisión si dichas reservas van a ser efectivas o canceladas hasta el momento del *registro*.

Luego de la categorización se decide excluir del análisis aquellos registros bajo la categoría C, dado que podrían ser cancelados en un futuro. Consecuente a esta exclusión, los registros totales bajaron de 121.922 a 114.307. De esta manera, no se utilizarán para el entrenamiento del modelo estas reservas que podrían generar riesgo de *data leakage* y mal entrenamiento del modelo. El ratio de cancelaciones del dataset utilizado para el modelo es de 16,17%.

Es importante tener presente las circunstancias en las que un algoritmo de aprendizaje automático puede sobre representar el error de generalización, lo que puede causar que no prediga con precisión al momento de implementación. Uno de los potenciales problemas es llamado *data leakage*.

Muchas veces, cuando se entrena el modelo, puede suceder que el mismo performa muy bien en el set de validación o *testing*, que luego es seleccionado como modelo final y puede ser puesto en producción. Sin embargo, existe un problema en los modelos de

aprendizaje automático en los que hay una discrepancia entre la performance del modelo en el testing y la implementación y puesta en producción, llamado *data leakage.*

El problema de *data leakage* se refiere a un error que consiste en evaluar y seleccionar modelos incorporando en el entrenamiento información que no estará disponible al momento de producción. Lo que tiene como consecuencia, que se sobreestime la performance del modelo propuesto. Este error se da también cuando el modelo de *machine learning* comparte accidentalmente información entre el set de datos de entrenamiento y de prueba. Cuando el modelo se usa con datos nuevos que hasta el momento no había visto, su performance va a ser mucho más baja a la esperada, dado que información de afuera del dataset de entrenamiento es usado para la creación del modelo. Esta información adicional puede permitir al modelo aprender o saber algo que de otra manera no podría, y que invalida la estimación de la performance del modelo construido. Por ejemplo, si alguna característica no estará disponible en la práctica al momento de usan el modelo para predecir, es un atributo que puede causar *data leakage* en el modelo. También puede darse cuando al entrenar el modelo de aprendizaje automático se utilizan datos que tienen la información que se busca predecir.

Es importante tratar de evitar los casos que podrían causar *data leakage* al entrenar el modelo, en general, se debería evitar cualquier cosa que pudiera implicar que el set de entrenamiento tenga conocimiento sobre el set de prueba. Muchas veces este problema pasa inadvertido y resulta en *overfitting*. Nisbet y otros (2019) colocan a este problema entre los principales diez errores del aprendizaje automático.

Tratar de identificar de antemano el error de *data leakage* y corregirlo es parte de la definición de un problema de predicción. Existen algunas estrategias para encontrar y eliminar este problema como el análisis exploratorio de los datos, examinando a través de técnicas estadísticas y/o de visualización, permitiendo revelar patrones que de otra manera no sería fácil detectar o darse cuenta si la performance del modelo es muy buena como para ser real.

Continuando con la preparación del dataset final, a los datos de reservas resultantes con información de reservas entre el 19 de junio de 2018 y el 13 de octubre de 2019 se

procedió a realizarle diversos cambios para el enriquecimiento de datos y se generaron de nuevos campos. Entre ellos:

- Se transformó la fecha de reserva y *check-in* en día y mes con su número correspondiente.

- Se incorporó un campo que transforma en una variable *dummy* el estado de cada reserva: aquellas bajo el estatus "*New reservation*" y "*Modified reservation*" fueron categorizadas como no canceladas, mientras que aquellas bajo el estatus "*No show reservation*" o "*Cancelled reservation*" fueron categorizadas como canceladas.

- Se incorporó información del número de semana del año a la cual correspondía cada fecha.

- Se crearon variables categóricas con el día de realización de la reserva y del *check-in* (si fue lunes, martes, etc.).

- Se extrajo en campos separados la información de cantidad de adultos, niños, cuartos requeridos en la reserva y total de noches reservadas.

- Se extrajo de las fechas de *check-in -in* y *check-out* información de la cantidad de días de semana y cantidad de días de fin de semana de la reserva.

Asimismo, se creó una variable *dummy* que marca cuáles de las reservas existentes fueron realizadas por clientes asociados al programa de fidelización, en base al cruce del id de cliente de la base de reservas con aquel de la base de clientes de fidelización. En este proceso se encontró baja calidad de datos. Para implementar el modelo, deberían resolverse dichos problemas.

Por otro lado, se enriqueció el dataset final con datos de la base de clientes, al:

- Agregar el "RFM score" asociado a cada cliente. Es una medida utilizada para segmentar a los clientes en función de las actividades recientes, frecuencia y valor monetario de las compras realizadas. Es muy utilizado en la industria hotelera para la retención de clientes.

- Agregar una variable dicotómica que establece si la cadena de hoteles cuenta el mail de cliente o no.

- Eliminar el campo 'País del cliente' y 'Ciudad del cliente', dado que se encontraron muchas inconsistencias entre la ciudad y el país. Además, se observó que en repetidas ocasiones la información que contenía el país de origen del cliente estaba reemplazada por la del hotel. La información del cliente es verídica únicamente luego de que el mismo realiza

el ingreso al hotel y los datos son cargados en el sistema. Por lo tanto, la información contenida en la base no podría ayudar en la predicción de cancelaciones, y esto claramente generaría un problema de *data leakage*.

De la tabla de calificación de los hoteles se incorporaron los siguientes campos: cantidad de estrellas del hotel y calificación en los portales relevados -homogeneizando la puntuación en una escala de 1 a 10-.

Ordenado cronológicamente, se incorporaron campos referidos a la cancelación -o no- de reservas del mismo cliente. Para ellos, se incorporaron los siguientes campos:

- Incorporación del campo is_repeated_guest: para evaluar las características de usuarios recurrentes

- Incorporación del campo prev_cancellations: en aquellos casos donde el cliente sea recurrente, evalúa la cantidad de cancelaciones observadas en las reservas previas realizadas por dicho cliente

- Incorporación del campo prev_bookings_not_cancelled: en aquellos casos donde el cliente sea recurrente, evalúa la cantidad de reservas realizadas previamente que no hubieran sido canceladas por dicho cliente.

Por último, cabe señalar que, de haber contado con herramientas para identificar las transacciones de prueba por parte de la cadena de hoteles, debería haberse eliminado las mismas del set de datos. De esta manera, el modelo podría entrenarse mejor dado que las transacciones de prueba no reflejan el comportamiento habitual del negocio.

2.3. Estadística descriptiva, análisis y selección de parámetros

El *ratio* de cancelaciones de la base bajo estudio – que es de 16,17% en el período analizado- resulta bajo en comparación con los porcentajes que muestran otros estudios sobre la temática, que señalan que pueden representar más del 20% del total de reservas (Morales, Wang, 2010), más del 40% para el caso de reservas realizadas en agencias online (D-EDGE), o hasta más del 60% en hoteles cercanos al aeropuerto o que se encuentran en la ruta (Liu, 2004). Los hoteles que forman parte de esta cadena tienen entre 3 y 5 estrellas, con diferentes capacidades totales -algunos muy exclusivos de 14 habitaciones, otros más grandes con 141 cuartos- y hasta 8 salas de reuniones. La mayoría de los hoteles tienen

restaurant, en algunos casos más de uno y todos ellos tienen excelentes puntuaciones en los buscadores más reconocidos de hoteles.

Tal como puede observarse en la Tabla 1, la cancelación de las reservas presenta alta variabilidad entre los hoteles, con ratios de cancelación de reservas que oscila entre valores del 10,6% al 48,2%.

Al tratarse de una cadena de hoteles, este trabajo pretende encontrar un modelo de predicción holístico que encaje adecuadamente en todos los hoteles. Sin embargo, para otra instancia podría analizarse si debe construirse un modelo para cada uno de los hoteles individualmente, para establecer políticas específicas en cada uno de ellos. Se presenta la información de cantidad de reservas y cancelaciones por hotel:

Tabla 1. Total de cancelaciones y reservas por hotel

Hotel ID	Ciudad	Cancelaciones	Reservas	% de Cancelaciones
1	Buenos Aires - Recoleta	755	3.184	23,71%
2	Buenos Aires - Puerto Made	2.220	10.444	21,26%
3	Buenos Aires - Centro	1.688	9.450	17,86%
4	Buenos Aires - Recoleta	1.768	8.504	20,79%
5	Buenos Aires - Centro	2.016	17.691	11,40%
6	Buenos Aires - Centro	1.561	13.869	11,26%
7	El Calafate	1.099	3.981	27,61%
8	Neuquén	1.467	8.846	16,58%
10	Iguazu	1.140	6.647	17,15%
11	Villa La Angostura	401	1.747	22,95%
12	El Chalten	596	2.244	26,56%
13	Saltos del Mocona	208	1.170	17,78%
14	Buenos Aires - Palermo	392	2.175	18,02%
15	Puerto Madryn	816	3.488	23,39%
16	Villa Traful	79	344	22,97%
18	Tigre, Buenos Aires	2.069	19.452	10,64%
19	Luján	168	986	17,04%
20	San Miguel del Monte	41	85	48,24%
Total general		**18.484**	**114.307**	**16,17%**

Fuente: Elaboración propia

A continuación, se muestran los porcentajes de cancelación por canal de comercialización de las habitaciones de los hoteles:

Tabla 2. Total de cancelaciones y reservas por canal de comercialización.

Canal de comercialización	Cancelaciones	Reservas	% de Cancelaciones
PMS	7.002	57.630	12,15%
OTA	8.169	39.019	20,94%
Hotel - call center	1.663	9.616	17,29%
Hotel - web propia	943	4.505	20,93%
GDS	707	3.537	19,99%
Total general	**18.484**	**114.307**	**16,17%**

Fuente: Elaboración propia

Tal como puede observarse en la Tabla 2, el ratio de cancelaciones es mucho mayor cuando las compras se realizan a través de las Agencias de Turismo Online (OTA) que alcanza el 20,9%, del Sistema de Distribución Global (*Global Distribution System*- GDS) que opera como mayorista llega al 20,0% y de la página web del hotel que es del 20,9%. En cambio, se observa un porcentaje menor cuando las reservas se realizan a través del Call Center de la cadena, pero principalmente cuando llegan a través del Sistema Operativo de Gestión del Hotel PMS -*Property Management System*-.

Además de los canales de comercialización se cuenta con información adicional por subcanal de comercialización donde se hicieron las reservas en la Tabla 3, se observan algunos en donde los porcentajes de cancelación resultan mucho más elevados que el promedio del canal. Para dar un ejemplo, en el caso de Booking.com las cancelaciones alcanzan el 27,6%, para Hotelbeds.com más del 34,7%, para Amadeus el 38,2% y para Despegar 6,3%.

Tabla 3. Total de cancelaciones y reservas por canal y subcanal de comercialización.

Canal de comercialización	Subcanal	Cancelaciones	Reservas	% de Cancelaciones
PMS		7.002	57.630	12,15%
OTA	BOOKING.COM	5.038	18.226	27,64%
	DESPEGAR PAM 2	503	7.969	6,31%
	EXPEDIA	1.348	7.707	17,49%
	BESTDAY	299	2.373	12,60%
	HOTELBEDS	696	2.005	34,71%
	ABREU	118	327	36,09%
	SUNHOTELS	73	179	40,78%
	LOGI TRAVEL	51	122	41,80%
	CONEXTUR - OTAC	20	34	58,82%
	HRS	9	29	31,03%
	BARNEOS - OTAC	6	27	22,22%
	NITES TRAVEL	-	10	0,00%
	TIPGROUP - OTAC	6	6	100,00%
	SMALL HOTELS - OTAC	1	4	25,00%
	BOOKASSIST - OTAC	1	1	100,00%
Hotel - call center	CALL-HOTEL	1.663	9.616	17,29%
Hotel - web propia	Web propia	810	3.633	22,30%
	Google	43	350	12,29%
	MARKETPLACE	59	308	19,16%
	Tripadvisor	7	85	8,24%
	LOCALUNIVERSAL_CPA	12	58	20,69%
	Trivago	6	33	18,18%
	TRIPINSTANT	1	12	8,33%
	MAPRESULTS_CPA	2	9	22,22%
	Kayak	-	5	0,00%
	PERFMEDIA_DISPLAY_BK	3	5	60,00%
	HTLCOMB	-	3	0,00%
	SKYSCANNER	-	3	0,00%
	WEGO	-	1	0,00%
GDS	SABRE	648	3.377	19,19%
	AMADEUS	39	102	38,24%
	GALILEO	9	37	24,32%
	WORLDSPAN	11	17	64,71%
	ODD-Hotelzon	-	4	0,00%
Total general		**18.484**	**114.307**	**16,17%**

Fuente: Elaboración propia

Se realizó la estadística descriptiva de los datos para conocer mejor el comportamiento de las variables. Además, se generó la matriz de correlaciones y la de información mutua. Esta última mide cuanto nos dice un atributo sobre el otro, en donde un valor elevado en la matriz indica una importante correlación de variables y si es cero

significa que las variables son independientes. En función de ello, se observaron variables que no aportaban información adicional, adaptando la base de datos para hacerla más manejable.

Antonio, de Almeida y Nunes (2019) señalan que la selección e ingeniería de atributos son de los factores más importantes para el éxito de un proyecto de *machine learning*. El dataset que se utiliza para el armado del modelo fue creado en base a una selección de los atributos -con el objeto de reducir la dimensionalidad y buscando evitar la pérdida de información-. La selección de parámetros puede resultar beneficiosa si son removidas aquellas que son redundantes, irrelevantes o generan ruido en el armado del modelo.

De acuerdo con Domingos (2012) la selección e ingeniería de parámetros requiere no solo de conocimiento técnico sino también de creatividad, intuición y conocimiento del negocio.

En el Apéndice 1 se presentan los atributos utilizados para la realización del modelo. Cabe mencionar que se transformaron algunas variables categóricas a dicotómicas como el tipo de habitación, ciudad del hotel, canal de comercialización, entre otras. De esta manera se adaptó la base para poder utilizar diferentes algoritmos -como es el caso de redes neuronales que no soportan atributos categóricos-. El total de columnas del set de datos final asciende a 163os.

La base de datos contenía gran cantidad de valores faltantes en los siguientes atributos: *rfm_score*, *rate_category*, Canal y Subcanal. Con el objeto de mejorar la capacidad predictiva del modelo se utilizó un módulo de Azure ML Studio para reemplazar los valores faltantes, cuyo algoritmo da la posibilidad de reemplazarlos por la media, por la mediana, por la moda, por un valor específico definido o a través del Análisis de Componentes Principales. Sin embargo, en función de que los resultados no mejoraban significativamente, se decide no imputar los valores faltantes, sino que se mantienen como NA.

Por otro lado, hay que resaltar que en el set de datos las clases se encuentran desbalanceadas, con 95.823 reservas que no fueron canceladas -el 83,83% del total- y solo 18.484 que si lo fueron -16,17% del total-. Esto representa uno de los problemas que

dificultan la labor de los clasificadores y puede llevar a una disminución de la calidad de la clasificación realizada. De acuerdo con Moreno y otros (2009) ante un desbalanceo de clases -donde el número de instancias de cada clase es muy diferente- los algoritmos pueden presentar una tendencia de clasificación hacia la clase mayoritaria, minimizando de esta manera el error de clasificación y clasificando correctamente instancias de clase mayoritaria en detrimento de instancias de clase minoritaria. Algunos de los tratamientos a datasets desbalanceados son el sobremuestreo -*Oversampling*- que consiste en balancear la distribución de las clases añadiendo ejemplos de la clase minoritaria y señalan que uno de los algoritmos más representativos es el *SMOTE* (*Syntetic Minority Over-sampling Technique*), que genera artificialmente nuevas instancias de la clase minoritaria interpolando los valores de las instancias minoritarias más cercanas a una dada. El *Oversampling* tiene la ventaja de no perder información, pero puede repetir muestras con ruido además de aumentar el tiempo necesario para procesar el conjunto de datos.

Según Moreno y otros (2009) el algoritmo *SMOTE* es un algoritmo de *oversampling* que genera instancias sintéticas o artificiales para equilibrar la muestra de datos basado en la regla del vecino más cercano. La generación se realiza extrapolando nuevas instancias en lugar de duplicarlas. Para cada una de las instancias minoritarias se buscan las instancias minoritarias vecinas (más cercanas) y se crean n instancias entre la línea que une la instancia original y cada una de las vecinas. El valor de n depende del tamaño de *oversampling* deseado. Chawla y otros (2002) describen el método como un muestreo de datos donde la clase minoritaria es sobremuestreada para crear muestras sintéticas en lugar de realizar un sobremuestreo con remplazamiento.

En este trabajo, para balancear se utilizó el módulo de Azure *Machine Learning* llamado *SMOTE* para aumentar el número de casos subrepresentados del conjunto de datos que se usa para el aprendizaje automático. Se decidió realizarlo mediante este módulo, dado que se comparó con la utilización del paquete de R llamado SMOTE y este último resultaba más demandante en términos computacionales.

SMOTE es una técnica estadística de sobremuestreo de minorías sintéticas para aumentar el número de casos de un conjunto de datos de forma equilibrada. El caso específico del módulo de Azure ML Studio funciona generando nuevas instancias a partir de casos minoritarios existentes que se proporcionan como entrada. Esta implementación de

SMOTE no cambia el número de casos de mayoría por lo que queda igual la cantidad de casos de reservas no canceladas. Las instancias nuevas no son copias de los casos minoritarios, sino que el algoritmo toma muestras del espacio de características de cada clase de destino y de sus vecinos más cercanos, y a partir de ello, genera nuevos ejemplos que combinan las características del caso con características de sus vecinos. Se utiliza cuando la clase que se desea analizar está subrepresentada, el módulo devuelve un conjunto de datos que contiene los ejemplos originales y varios ejemplos de minorías sintéticas, en función del porcentaje que se especificado, en este caso fue del 200% y k=1.

El set de datos resultante luego del sobremuestreo contiene 151.276 filas de las cuales 95.823 son casos donde las reservas no fueron canceladas -el 63,34% del total- y 55.452 son cancelaciones -36,66% del total-. Como puede observarse, la clase mayoritaria contiene la misma cantidad y se incrementó la clase minoritaria.

Además, se normalizaron algunas de las variables numéricas como el Revenue_Usd, rfm_score, ADR_USD, AVG_lead con el objetivo de cambiar los valores a una escala en común. Esto es importante para que algoritmos - como es el caso de Regresión Logística - funcionen correctamente. El módulo utilizado se llama *Normalize Data* y se seleccionó el método de transformación 'Zscore'[4].

Estas modificaciones fueron realizadas en función de lograr mejorar la performance de los modelos. Cabe aclarar que en todos los módulos que lo permiten se fijó la semilla 123, para que los resultados del trabajo puedan ser reproducidos.

Por último, se utilizó el módulo *Tune Model Hyperparameters* (Optimizar hiperparámetros del modelo) que sirve para determinar los hiperparámetros óptimos para un modelo de *Machine Learning*. El módulo compila y prueba varios modelos con diferentes combinaciones de configuraciones y compara las métricas de todos los modelos para obtener las combinaciones de valores. Durante el entrenamiento se selecciona la métrica que se

[4] Función matemática que convierte todos los valores de las columnas seleccionadas en puntuación z, de acuerdo con la siguiente fórmula, donde la media y la desviación estándar se calculan por separado para cada columna:

$$z = \frac{\mathrm{x}-\mathrm{mean(x)}}{stdev(x)}$$

quiere usar para la optimización de los modelos de clasificación, que en este caso es AUC-ROC.

3. Modelo predictivo para la cancelación de reservas

El proyecto se desarrolló con los softwares Azure Machine Learning Studio y R, a fin de efectuar cada proceso en la aplicación que se considera más útil en términos de tiempo de programación, preparación y de procesamiento y salida de resultados.

Se utilizaron tanto Azure ML Studio y R para la preparación de los datos, el reemplazo de datos faltantes, la generación de datasets con datos más balanceados, análisis exploratorio de los datos y la prueba de diferentes algoritmos.

3.1. Desarrollo de diferentes modelos a través de técnicas de aprendizaje automáticos

En este caso se busca encontrar un modelo, a través de distintos algoritmos para resolver un problema práctico, y poder a partir de esto tomar decisiones en el negocio hotelero -es en este caso relacionado a la predicción de la cancelación de reservas-. El problema se plantea como uno de clasificación.

Para Flath y Stein (2018) en el aprendizaje supervisado se busca inferir una función a partir de datos de entrenamientos presentados en formato tabular. Lo que caracteriza el aprendizaje supervisado es la predicción de una variable, que puede ser una variable continua -y se clasifica como un problema de regresión- o, como en el caso de este trabajo, de una variable categórica -que corresponde a un problema de clasificación-.

Se busca evaluar mediante distintos métodos y algoritmos cuál es que predice mejor la cancelación de las reservas, a través de técnicas utilizadas para la clasificación como Árboles de decisión, *Random Forest*, *Gradient Boosting Trees*, Redes Neuronales y Regresión Logística. Los algoritmos de aprendizaje automático exploran los datos, descubren patrones en ellos, que luego son utilizados para clasificar los nuevos casos. Además, permiten ajustar y optimizar los parámetros del clasificador elegido.

A continuación, se presentan algunos de los algoritmos que fueron utilizados.

Naïve Bayes o modelo de Bayes ingenuo está basado en una técnica de clasificación estadística llamada Teorema de Bayes y es uno de los algoritmos más usados para clasificación por ser uno de los más sencillos. El algoritmo de Azure Machine Learning Studio llamado *Two-Class Bayes Point Machine* usa una aproximación bayesiana a la clasificación lineal llamada "*Bayes Point Machine*". Este algoritmo (Herbrich, Grapel y Campbell, 2001) aproxima de manera eficiente el promedio bayesiano óptimo teórico de los clasificadores lineales (basado en su capacidad de generalizar) al elegir un clasificador "promedio", llamado el punto bayesiano. En estudios empíricos se mostró que Bayes Point Machine supera de manera consistente la performance de *Support Vector Machines*.

El caso de Regresión logística es un algoritmo útil para la clasificación de dos clases que utiliza una curva con forma de S para dividir los datos en grupos en lugar de una línea recta a diferencia de la regresión lineal. La regresión logística proporciona límites de clase lineal.

Por otro lado, las Redes Neuronales también pueden usarse para predecir o clasificar de manera binaria. De acuerdo con la documentación brindada por Azure, una red neuronal es un conjunto de capas interconectadas. Las entradas son la primera capa y se conectan a una capa de salida mediante un grafo acíclico que consta de nodos y aristas ponderadas. Entre las capas de entrada y salida puede insertar varias capas ocultas. La mayoría de las tareas de predicción pueden realizarse con solo una o varias capas ocultas. Sin embargo, las investigaciones recientes han demostrado que las redes neuronales profundas (Deep Neuronal Networks) con muchas capas pueden ser eficaces en tareas complejas, como en el reconocimiento de imágenes o de voz. En este trabajo se utilizó el módulo llamado *Two-Class Neural Network.*

Los árboles de decisión son uno de los algoritmos más sencillos y fáciles de implementar e interpretar y a su vez son de los más poderosos. De acuerdo con Barga y otros (2015) los algoritmos de árboles de decisión son técnicas jerárquicas que funciona dividiendo iterativamente el ser de dados basado en cierto criterio estadístico. El objetivo de los árboles de decisión es el de maximizar la varianza a través de los distintos nodos en el árbol, y minimizar la variancia dentro de cada uno de estos nodos. Algunos de los algoritmos de árbol de decisión más utilizados son *Iterative Dichotomizer* 3 (ID3), C4.5 y C5.0 (el

sucesor de ID3), *Automatic Interaction Detection* (AID), *Chi-Squared Automatic Interaction Detection* (CHAID), y *Classification and Regression Tree* (CART) que son utilizados para clasificación y para regresión.

El algoritmo *Random Forest* (Breiman, 2001) está formado por un conjunto de árboles de decisión. Es un tipo de ensamble -una modificación del *Bagging*, el cual trabaja con una colección de árboles no correlacionados y los promedia (Hastie, Friedman y Tibshirani, 2001)-, donde cada árbol depende de los valores de un vector aleatorio de la muestra de manera independiente y con la misma distribución de todos los árboles en el bosque. Según Medina Merino y Ñique-Chacón (2017) el error de generalización de un bosque de árboles de clasificación depende de la fuerza de los árboles individuales en el bosque y la correlación entre ellos.

Gradient Boosted Trees (GBT), también conocido como Gradient Boosting Machine (GBM) o Gradient Boosted Regression Tree (GBRT) es definido por Chambers y Dinsmore (2015) como un tipo de algoritmo que produce diferentes modelos individuales (árboles de decisión, por ejemplo) cuyos resultados se van agregando de modo que el resultado final – el clasificador de ensamble - está formado por un modelo que es una combinación de los anteriores (clasificadores débiles), pero con una capacidad de predicción muy superior a la de los modelos individuales en los que se basa. En las sucesivas iteraciones, GBT aprende y minimiza los errores de los modelos anteriores y ajusta los árboles de decisión a los residuos o errores con el fin de ir actualizando y minimizando los mismos. De acuerdo con Vargas, Carmona (2018) una de las características más significativas de los algoritmos basados en *boosting* es que aprenden de los errores de los múltiples modelos a medida que los va generando, además GBT permite la existencia de valores extremos, correlaciones altas entre las variables, relaciones no lineales, la presencia de valores perdidos y admite el uso de variables categóricas como independientes.

Chen y Guestring (2016) reconocen que, entre los métodos de aprendizaje automático utilizados en la práctica, GBT es una técnica que sobresale en muchas aplicaciones. Los autores muestran que el uso del algoritmo XGBoost es muy efectivo para el armado de modelos de clasificación. Por eso es que en este caso se usa para predecir si cada una de las reservas va a ser cancelada o no. Antonio, de Almeida y Nunes (2019) señalan que el

XGBoost es un ensamble de árboles de decisión que es reconocido como uno de los más efectivos y rápidos entre los algoritmos de clasificación y también de regresión.

La efectividad del XGBoost, en lo que respecta particularmente en el control de *overfitting*, es alcanzada porque tiene un set de parámetros que permiten calibrar el modelo -parámetros para hacer al entrenamiento más robusto frente a la existencia de ruidos en los datos-.

En este trabajo se utilizaron los algoritmos contenidos en los diferentes módulos de Azure ML Studio: *two-class Decision Forest*, two-class *Decision Jungle*, two-class Boosted Decision Tree, two-class *Bayes Point Machine*, two-class Neural Network y two-class *Logistic Regression*.

3.2. Resultados

Entender la importancia de la elección de la métrica a utilizar para la evaluación del modelo es fundamental para el éxito o fracaso de cualquier proyecto de *Data Science* según Flath y Stein (2018). La selección del criterio de evaluación adecuado es más complicada en los casos de clasificación que de regresión según los autores. La mejor métrica puede variar según el problema de negocio del que se trate y el costo relacionado de los falsos positivos y los falsos negativos.

Las medidas de performance utilizadas para el caso de problemas de clasificación derivan de la matriz de confusión, que cuenta la cantidad de datos que fueron clasificados de manera errónea -Falsos Positivos y Falsos Negativos- y de manera correcta -Verdaderos Positivos y Verdaderos Negativos -. Sin embargo, la selección de cuál medida utilizar tiene que tener en cuenta el contexto de aplicación de lo que se quiere predecir. Por ejemplo, si se trata de predecir una clase desbalanceada, una medida de evaluación como la *Accuracy* falla dado que el modelo puede tener un *Accuracy* alto pero un pobre poder predictivo.

En términos generales, en lo que refiere a técnicas a utilizar para evaluar la capacidad predictiva de los algoritmos de clasificación, se utiliza el criterio de exactitud (*Accuracy*) o el AUC (*Area Under the Curve)* ROC (*Receiver Operating Characteristics)*. En este trabajo

se utiliza como métrica principal el AUC-ROC, y se muestran también otras métricas alternativas.

La curva ROC muestra la tasa de positivos verdaderos frente a la de falsos. Se busca maximizar el Área Bajo la Curva -AUC- correspondiente, dado que cuanto más se acerque dicha curva a la esquina superior izquierda, mejor será el rendimiento del clasificador al maximizar la tasa de positivos verdaderos a la vez que se minimiza la tasa de falsos positivos. El AUC brinda más información que la del criterio de exactitud, permitiendo realizar rankings de las predicciones, por ejemplo, de las reservas con más probabilidad de ser canceladas. Esto permite tomar acciones relacionadas a gestión de inventario de habitaciones o medidas de marketing sobre los clientes para evitar las cancelaciones.

Por otro lado, la validación cruzada es una técnica importante que se usa en *machine learning* para evaluar la variabilidad de un conjunto de datos y la confiabilidad del modelo entrenado con esos datos. Existen varias opciones para validar modelos: *Leave and Out - Cross Validation* y *k-fold Cross Validation*. La validación cruzada -*Cross Validation*- divide aleatoriamente los datos de entrenamiento en varias particiones. En el caso del algoritmo disponible en Azure se establece de forma predeterminada en 10 particiones -*folders*- si no se ha particionado previamente el conjunto de datos. Sin embargo, en caso de querer cambiar el número de particiones que vine predeterminado, se puede indicar el número k de particiones que se quiere realizar mediante el uso de otro módulo que se llama *Partition and Sample*.

El módulo de Azure funciona reservando los datos de una partición para la validación y usa el resto para entrenar un modelo. En caso del k=10, se generan diez modelos durante la validación cruzada, con cada uno de los modelos entrenado con 9/10 de los datos y probado con el 1/10 restante. Durante las pruebas del modelo para cada *folder*, se evalúan varias estadísticas de precisión, para este caso las correspondientes para evaluar los modelos de clasificación. Cuando el proceso de compilación y evaluación se completa para todas las particiones, el modelo de validación cruzada genera un conjunto de métricas de rendimiento y resultados puntuados de todos los datos.

La validación cruzada mide la precisión de un modelo y muestra, además, algún indicio sobre si el conjunto de datos es representativo y el grado de sensibilidad del modelo

a las variaciones en los datos. Por otro lado, la validación cruzada es mucho más demandante a nivel computacional y tarda mucho más tiempo que la validación con una división aleatoria, esto se debe a que se entrena y valida el modelo varias veces con un conjunto de datos mayor.

A continuación, se presentan los resultados de los modelos realizados:

Tabla 4. Resultados

Modelo	Accuracy	Precision	Recall	F-Score	AUC
Decision Jungle	0,86	0,82	0,14	0,24	0,77
Decision Forest	0,87	0,76	0,25	0,37	0,80
Binary Neural Network	0,86	0,62	0,56	0,59	0,85
Binary Bayes Point Machine	0,89	0,79	0,50	0,61	0,88
Logistic Regression	0,88	0,78	0,47	0,59	0,89
FastTree (Boosted Trees) Classification	0,91	0,81	0,68	0,74	0,93
Con Smote y Normalizado					
Decision Jungle	0,76	0,74	0,51	0,60	0,80
Decision Forest	0,80	0,80	0,62	0,70	0,87
Binary Neural Network	0,85	0,84	0,81	0,83	0,92
Binary Bayes Point Machine	0,84	0,87	0,74	0,80	0,91
Logistic Regression	0,85	0,87	0,75	0,80	0,91
FastTree (Boosted Trees) Classification	0,92	0,92	0,88	0,90	0,96
FastTree (Boosted Trees) Classification	0,92	0,93	0,88	0,91	0,97

Fuente: Elaboración Propia

Como se puede observar en la Tabla 4, los mejores resultados se obtuvieron a partir de la utilización del algoritmo *Boosted Trees*. Asimismo, se obtuvo una buena performance del modelo con Regresión Logística, el algoritmo *Binary Bayes Point Machine* y el algoritmo *Binary Neural Network.*

Cabe mencionar que al utilizar el módulo de *Tune Model Hyperparameters* de Azure para optimizar los hiperparámetros del algoritmo *Boosted Trees*, el resultado que arroja es un AUC-ROC de 0,97.

Para el caso del algoritmo de Regresión Logística se puede además obtener información relacionada a la importancia de cada atributo para explicar el modelo. De los resultados obtenidos, se desprende que para el caso de Regresión logística los mayores pesos de los parámetros se corresponden con datos de la base de reservas de la cadena, es decir, que son los más relevantes para explicar la cancelación o no de una reserva.

Algunos de los atributos más importantes son:

- El tiempo de anticipación con el que se realizó la reserva -Avg_Lead_Time- donde a mayor tiempo entre la reserva y el *check-in*, las reservas son más plausibles de ser canceladas;
- La tarifa media por noche -ADR_usd del inglés *Average Daily Rate*-, donde a mayor precio medio, las reservas son más plausibles de ser canceladas;
- Si el hotel cuenta o no con información del e-mail del cliente, las reservas más plausibles de ser canceladas son aquellas donde el hotel no cuenta con el e-mail del cliente;
- La cantidad de cancelaciones previas que tuvo ese cliente - atributo que fue creado en base a las reservas de los hoteles y la información de sus clientes-, ante mayor cantidad de cancelaciones previas, las reservas son más plausibles de ser canceladas;
- Las reservas cuyo *check-in* será realizado durante el fin de semana son menos plausibles de ser canceladas;
- El RFM_score del cliente, ante mayor RFM_score del cliente, las reservas son menos plausibles de ser canceladas;
- Si el cliente pertenece al programa de fidelización de la cadena de hoteles, las reservas son menos plausibles a ser canceladas;
- Entre otros.

Para encontrar la importancia de las variables en el armado del modelo también se usó el módulo *Permutation Feauture Importance.* Para el caso de *Boosted Trees*, se muestra que la importancia de las variables coincide en muchos casos con el caso de Regresión Logística, figurando los siguientes atributos: RFM_score, si el hotel tiene información del email del huésped, si el cliente pertenece al programa de fidelización, la tarifa en dólares -Revenue_USD-. Para el caso de Regresión Logística usando dicho módulo, los atributos más importantes son también el RFM_score, si el cliente brindó información del email o no, si es del programa de fidelización, si el check-in se realiza el fin de semana o no, entre otros. Esto refleja que los atributos con mayor peso representativo en el armado de los distintos modelos son muy similares.

De los atributos más importantes del modelo, la gran mayoría son datos provenientes directamente de la base de reservas del hotel. Esto facilitaría el armado del modelo, ya que

en la cadena todos los hoteles utilizan el mismo sistema, con algunos agregados que fueron creados para este trabajo -a través de ingeniería de atributos -, como si el cliente ya canceló en ocasiones anteriores o si el *check-in* se realizará durante el fin de semana. El agregado y la ingeniería de estos atributos podrían agregarse sin mayores dificultades a la base de datos para la implementación del modelo en la práctica.

Los resultados muestran que se puede obtener un modelo con buena capacidad predictiva a través de algoritmos más complejos como *Boosted Trees* o Redes Neuronales y también con Regresión Logística -que son de los más utilizados y conocidos-. Cabe señalar que existen *trade-offs* asociados a la elección del algoritmo, dado que los algoritmos más complejos aportan mayor poder de clasificación, pero muchas veces resulta difícil la interpretabilidad, la comunicación y la explicación del modelo. Además, se muestra en este caso la conveniencia de que el set de datos se encuentre más balanceado, dado que mejora los resultados en todos los casos.

3.3. Implementación del modelo de aprendizaje automático

Un modelo de aprendizaje automático busca mejorar los resultados o performance del negocio y lograr una ventaja competitiva frente a la competencia. Para poder alcanzar estos objetivos, es importante poder una exitosa implementación del modelo. Es por esta razón que resulta de importancia la elaboración de un marco para poder definir como el modelo debería ser implementado.

A pesar de que la implementación del modelo en un ambiente de producción no está bajo el alcance de este trabajo, resulta relevante destacar que la parte correspondiente a implementación es crítica para que el modelo sea exitoso.

Antonio, de Almeida y Nunes (2019) señalan que al implementar el modelo en un espacio de producción en tiempo real se encontraron con que tendía a sobreajustar los datos: los modelos no predecían bien para reservas nuevas, es decir, reservas desconocidas para una fecha futura que no habían sido incluidas en el desarrollo del modelo (en el entrenamiento de este). De acuerdo con Domingos (2012) este es un problema común en los modelos de aprendizaje automático.

Los autores revelan que dos cuestiones influyeron considerablemente en la baja de la performance en el entorno de producción: *data leakage* -que ya fuera mencionado anteriormente- y el problema de *dataset shift*, que se refiere al problema que se da cuando los datasets de entrenamiento y prueba tienen diferentes distribuciones de probabilidad. Quiñonero-Candela y otros (2009) señalan que la distribución conjunta de inputs y outputs difieren entre la etapa de entrenamiento y de prueba. Para el caso que pusieron en producción Antonio, de Almeida y Nunes (2019) las razones del cambio de distribución fueron: la estrategia de división de conjuntos de datos estratificados para la creación de los conjuntos de datos de *training* y *testing* no garantizaba una distribución comparable entre los conjuntos de datos por la velocidad a la que cambia el negocio de los hoteles y el rápido crecimiento de la industria del turismo en los últimos años y la creciente demanda anual provoca un rápido aumento en los precios -Tarifa diaria promedio- y que contribuyen a las diferencias en la distribución de entradas y salidas a lo largo del tiempo.

Por otro lado, agregan además otro factor que influye en este problema, que se relaciona con la gran cantidad de cambios en la industria hotelera y el incremento de operaciones que hace que se incorporen nuevos jugadores en el negocio -como ocurre con las Agencias de viajes en línea *OTA*s- y la desaparición relativa de otros jugadores -como las agencias de viajes tradicionales-. Estas constantes transformaciones en el negocio contribuyen a un cambio en el peso de los distintos canales de venta en la operación del hotel y pueden afectar la distribución de ciertas características o parámetros a lo largo del tiempo.

En función de la experiencia de los trabajos realizados por los autores, hay que tener en cuenta estas cuestiones en el armado del modelo -en la construcción y división del dataset e ingeniería de los parámetros- para que luego no afecte la implementación. Como en este caso el mejor modelo fue obtenido con *Boosted Trees* y teniendo en cuenta los continuos cambios que pueden ocurrir en el negocio y por lo tanto en el set de datos, debe tenerse en cuenta que en caso de llevar a producción el mismo debería calibrarse periódicamente.

Conclusión

El presente trabajo realizado en base a datos de más de un año de reservas en hoteles de una cadena que se encuentran ubicados en distintas ciudades de Argentina reveló la

importancia que tiene para la organización contar con datos de buena calidad de las reservas y de los clientes, y el hecho de tener un buen manejo de bases de datos para poder generar modelos predictivos que permitan accionar para impulsar mejoras en los niveles totales de ventas, ocupación y reputación hotelera.

Durante la etapa de análisis se exhibió la capacidad predictiva de los modelos, teniendo presente que para la fase de implementación y puesta en producción del mismo deberían realizarse algunas mejoras, ya sea ampliando el rango de fechas de reservas para mejorar la capacidad de aprendizaje del modelo y poder entrenarlo con datos que permitan una mejor generalización, como así también la potencialidad de realizar análisis pormenorizado en cada uno de los hoteles, de manera de identificar patrones que permitan mejorar la performance predictiva de los modelos.

A partir del trabajo se mostró que pueden predecirse con una AUC-ROC del 0,97 las cancelaciones de las reservas con un modelo genérico para los hoteles analizados. Sin embargo, al observarse gran heterogeneidad de datos y características entre los hoteles, podría elaborarse un modelo para cada uno de ellos en pos de mejorar aún más la performance de estos de manera de obtener datos más precisos de las cancelaciones y realizar mejores acciones de manejo de la demanda e ingresos por parte de la organización. La construcción de un modelo de clasificación entre las reservas más propensas a ser canceladas y las que no, es una herramienta que puede usar la cadena para disminuir la incertidumbre en sus proyecciones de demanda, proyecciones financieras, mejorar la ocupación de las habitaciones del hotel tratando de captar esas reservas para que se concreten o reconociendo aquellas que podrían ser fraudulentas o altamente plausible de cancelación para poner directamente esa plaza como disponible. Además, para el sistema de *Revenue Management*, podría calcularse la demanda neta por día a partir de dichas predicciones, y de este modo realizar una programación más dinámica de precios y asignación de habitaciones.

Para el caso particular analizado en el presente trabajo, el impacto de la cancelación de reservas desde agosto de 2018 a octubre de 2019 para esta cadena de hoteles habría significado aproximadamente 4.856.727 dólares estadounidenses -teniendo en cuenta que, en el transcurso de los meses analizados, mientras se cancelaron reservas también se efectuaron nuevas-. Es por eso que el monto sirve como referencia del problema y no como la monetización de las pérdidas por cancelaciones.

Por lo tanto, estos modelos permitirían a los administradores no sólo tomar decisiones sobre la política de cancelaciones de la empresa, sino también mejorar las finanzas al hacer proyecciones de demanda más precisas. Asimismo, la utilización de este tipo de modelos permitiría disminuir la incertidumbre en la asignación de stock (de las habitaciones), es decir, poder realizar un mejor manejo de inventario o mejorar la estrategia de marketing para evitar que se produzca dicha cancelación.

Futuros estudios

De la elaboración del modelo de predicción de cancelación de reservas, se desprenden algunos interrogantes que podrían convertirse en futuras investigaciones.

En función de los resultados obtenidos en este trabajo, resulta importante transmitir a la organización el valor que tienen los datos para resolver problemas de gestión, y la necesidad de mejorar la calidad y gobernanza de los datos con los que cuenta la organización para realizar un prototipo de modelo que pueda ser efectiva y exitosamente implementado.

Por otro lado, sería importante incorporar un volumen mayor de información y hacer extensivo el modelo a los otros hoteles de la cadena que se encuentran ubicados en países como Colombia o Estados Unidos, de manera de evaluar los resultados que se obtienen, y verificar si para los mismos se obtiene una buena capacidad predictiva en general al igual que para el caso de Argentina.

Asimismo, podrían agregarse más fuentes de datos para mejorar la capacidad predictiva del modelo, como es el caso de recolección de datos no estructurados relacionados con los hoteles, incorporar los datos del clima, calendario de feriados y fechas importantes (nacionales y de los turistas que más reciba el hotel), de eventos que se realicen en la ciudad donde está ubicado el hotel, comentarios de las redes sociales, precios de los competidores, entre tantas otras fuentes de datos.

El Observatorio de Turismo de la Ciudad de Buenos Aires se encuentra desarrollando un Sistema de Inteligencia Turística, plataforma digital que permite visualizar de forma dinámica los principales datos del turismo a partir de múltiples fuentes de información, que podría explotarse de alguna manera, sobre todo para el caso de los hoteles ubicados en esta

ciudad. Además, podrían incorporarse datos de la Encuesta de Turismo Internacional del Instituto Nacional de Estadísticas y Censos (INDEC) y la información del turismo receptivo y emisivo en el Aeropuerto Internacional de Ezeiza y Aeroparque Jorge Newbery que se encuentra en la página de Datos Abiertos. En base a esto, se podría determinar cómo influye el flujo de turistas en la ocupación y cancelación de los hoteles bajo análisis, y ver si mejora la performance del modelo de predicción.

Con el modelo de predicción de las reservas plausibles a ser canceladas para toda la cadena hoteles, se busca crear una herramienta que le permita al hotel visualizar diariamente proyecciones más certeras de demanda neta y, en consecuencia, ajustar la política de precios de la empresa para un mejor manejo de la demanda.

Por último, podría plantearse la cancelación de reservas como un problema de detección de anomalías y compararse con los resultados obtenidos en este trabajo.

Referencias bibliográficas

Antonio, N., de Almeida, A., & Nunes, L. (2019). An automated machine learning based decision support system to predict hotel booking cancellations. *An automated machine learning based decision support system to predict hotel booking cancellations*, (1), 1-20.

Antonio, N., de Almeida, A., & Nunes, L. (2017). Predicting hotel booking cancellations to decrease uncertainty and increase revenue. *Tourism & Management Studies*, *13*(2), 25-39.

Antonio, N., de Almeida, A., & Nunes, L. (2017). Predicting hotel bookings cancellation with a machine learning classification model. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 1049-1054). IEEE.

Barga, R., Fontama, V., Tok, W. H., & Cabrera-Cordon, L. (2015). *Predictive analytics with Microsoft Azure machine learning*. Berkely, CA: Apress.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Chambers, M. y Dinsmore T. W. (2015). Advanced Analytics Methodologies: Driving Business Value with Analytics, Pearson Education, New Jersey.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794). ACM.

Chen, C. C., Schwartz, Z., & Vargas, P. (2011). The search for the best deal: How hotel cancellation policies affect the search and booking decisions of deal-seeking customers. International Journal of Hospitality Management, 30(1), 129-135.

Domingos, P. M. (2012). A few useful things to know about machine learning. *Commun. acm*, *55*(10), 78-87.

Egan, David y Haynes, Natalie (2018). Manager perceptions of big data reliability in revenue management. International Journal of Quality & Reliability Management.

Fosso Wamba, S., Akter, S., Edwards, A., Chopin, G., y Gnanzou, D. (2015). How 'Big Data' Can Make Big Impact: Findings from a Systematic Review and a Longitudinal Case Study. International Journal of Production Economics, 165, 234-246.

Flath, C. M., & Stein, N. (2018). Towards a data science toolbox for industrial analytics applications. *Computers in Industry*, *94*, 16-25.

Foster Provost y Tom Fawcett (2013). Data Science and its Relationship to Big Data and Data-Driven Decision Making. https://doi.org/10.1089/big.2013.1508

Giudici, P., Applied Data Mining, John Wiley & Sons Inc, 2003.

Gupta Krishna, Gauba Tanay, Jain Saransh (2017). Big Data in Hospitality Industry: A survey. International Research Journal of Engineering and Technology (IRJET), 2395-0056 Volume: 04 Issue: 11 | Nov -2017

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Hand, D.J.; Mannila, H. & Smyth, P. (2000). Principles of Data Mining. The MIT Press. USA.

Herbrich, R., Graepel, T., & Campbell, C. (2001). Bayes point machines. *Journal of Machine Learning Research*, *1*(Aug), 245-279.

Holsapple, C, Lee-Post, A and Pakath, R. (2014). A unified foundation for business analytics. *Decision Support Systems, 64: 130–141*.

Ivanov, S. (2014). Hotel revenue management: From theory to practice. Zangador.

Kim, G. H., Trimi, S., & Chung, J. H. (2014). Big data applications in the government sector. *Communications of the ACM*, *57*(3), 78-85.

Kimes, S. E., & Wirtz, J. (2003). Has revenue management become acceptable? Findings from an International study on the perceived fairness of rate fences. Journal of Service Research, 6(2), 125–135.

Kusnetzky, D. (2010). What is "Big Data"? ZDNet.

Li, J., Xu, L., Tang, L., Wang, S., & Li, L. (2018). Big data in tourism research: A literature review. Tourism Management, 68, 301-323.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012). Big data: the management revolution. Harvard business review, 90(10), 60-68.

Marr, M. (2016). How Big Data and Analytics Are Changing Hotels and The Hospitality Industry. Forbes, 01/2016.

Medina-Merino, R., & Ñique-Chacón, C. (2017). Bosques aleatorios como extensión de los árboles de clasificación con los programas R y Python. Interfases, 0(010), 165-189. doi:http://dx.doi.org/10.26439/interfases2017.n10.1775

Mehrotra, R., & Ruttley, J. (2006). Revenue management (second ed.). Washington, DC, USA: American Hotel & Lodging Association (AHLA).

Miele, S., & Shockley, R. (2013). Analytics: The real-world use of big data. *Retrieved from IBM Institute for Business Value, Said Business School*.

Morales, DR and Wang, J. (2010). Forecasting cancellation rates for services booking revenue management using data mining. *European Journal of Operational Research, 202(2): 554–562.*

Moreno, J., Rodríguez, D., Sicilia, M. A., Riquelme, J. C., & Ruiz, R. (2009). SMOTE-I: mejora del algoritmo SMOTE para balanceo de clases minoritarias. *Actas de los Talleres de las Jornadas de Ingeniería del Software y Bases de Datos*, *3*(1).

Nisbet, R., Elder, J., & Miner, G. (2009). *Handbook of statistical analysis and data mining applications*. Academic Press.

OBS Business School (2018). El salto del Big Data al Huge Data. Situación del Big Data en 2018-2019.

Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. D. (2009). *Dataset shift in machine learning*. The MIT Press.

Riasi, A. (2018). Overbooking practices in the hotel industry and their impact on hotel's financial performance *(Doctoral dissertation, University of Delaware*).

Sanz-Magallón Delhaize Gonzalo (2018). Oportunidades del uso de Big Data en empresas del sector de alojamiento. Implicaciones para la política de competitividad del sector turístico a partir del caso de la Unión Europea.

Schmarzo, B. (2013). Big Data: Understanding how data powers big business. *John Wiley & Sons.*

Schroeck, M., Shockley, R., Smart, J., Romero-Morales, D., & Tufano, P. (2012). Analytics: The real-world use of big data. *IBM Global Business Services*, *12*(2012), 1-20.

Smith, S. J., Parsa, H. G., Bujisic, M., & van der Rest, J. P. (2015). Hotel cancelation policies, distributive and procedural fairness, and consumer patronage: A study of the lodging industry. *Journal of Travel & Tourism Marketing, 32(7), 886-906.*

Steinmann, M., Matei, S. A., & Collmann, J. (2016). A theoretical framework for ethical reflection in big data research. In *Ethical Reasoning in Big Data* (pp. 11-27). Springer, Cham.

Van Ryzin, G. J., & Talluri, K. T. (2005). An introduction to revenue management. *In Emerging Theory, Methods, and Applications (pp. 142-194). Informs.*

Vargas, Carmona (2018). Análisis de la utilidad del algoritmo Gradient Boosting Machine (GBM) en la predicción del fracaso empresarial.

Anexo 1

Acuerdo con la organización para el manejo y uso responsable de los datos.

El presente acuerdo es celebrado por:

Lissette Bolinaga Otero,

DNI [redacted]

Y por otra parte:

[redacted]

Que para fines del presente serán definidos como "las partes", quienes a su vez se reconocen para obligarse y cumplir los siguientes puntos:

Objetivo del acuerdo

El presente acuerdo se refiere a la información confidencial que comparten las partes, con el objetivo de la realización de un proyecto académico.

Ambas partes acuerdan que todos los datos que se intercambien entre ellas quedará amparada por la obligación de confidencialidad.

Estipulaciones

PRIMERA. Autoriza al uso y procesamiento de los datos correspondientes a Lissette Bolinaga Otero de acuerdo con las condiciones de confidencialidad entre las partes.

SEGUNDA. Entiende que los datos y la información será utilizada exclusivamente para fines académicos en el marco del Trabajo de la Especialización en Métodos Cuantitativos para la Gestión y Análisis de Datos en Organizaciones, de la Escuela de Estudios de Posgrado de la Facultad de Ciencias Económicas de la Universidad de Buenos Aires.

Para constancia de lo anterior, se firma en Buenos Aires [redacted]

Firma: [redacted]

Nombre y Apellido: [redacted]

Apéndice 1

Estructura del set de datos y descripción de los parámetros.

Set de datos final utilizado para el armado del modelo

Atributo	Tipo	Descripción
Reserva_tipo	Categórica	Tipo de reserva
Cliente_tipo	Categórica	tipo de cliente según categorías del hotel
Cod_desc	Dicotómica	Si posee algún descuento
Revenue_usd	Numérica	Tarifa en dólares
ADR _USD	Numérica	Tarifa promedio diaria en dólares.
Room_Nights	Numérica	Cantidad de noches de la reserva
Avg_Lead_Time	Numérica	Cantidad de días de anticipación con los que se hizo la reserva
Rate_categoría	Categórica	Categoría de la tarifa
Canal	Categórica	Canal de comercialización
Subcanal	Categórica	subcanal de comercialización. Especifica web
Hotel_id	Numérica	Hotel id
Agente	Numérica	Id del agente que vendio a través del Call Center
Room_code	Categórica	Tipo de habitación asignado en la reserva
hotel_city	Categórica	Ciudad donde se encuentra ubicado el hotel
loyalty_rewards	Dicotómica	Si el cliente pertenece o no al programa de fidelización
Rewards_status	Categórica	Categoría del cliente en el programa de fidelización.
rfm_score	Numérica	Recency, Frequency and Monetary Modeling. Score basado en tres ejes: datos de actividades recientes, frecuencia y datos
checkin_finde	Dicotómica	Si el Chek-in será es en fin de semana
hotel_stars	Numérica	cantidad de estrellas del hotel
hotel_qualification_tripadvisor	Numérica	calificación de 0 a 10
hotel_qualification_booking	Numérica	calificación de 0 a 10
hotel_qualification_google	Numérica	calificación de 0 a 10
adults	Numérica	cantidad de adultos
children	Numérica	cantidad de niños
rooms	Numérica	cantidad de habitaciones de la reserva
is_repeated_guest	Dicotómica	si es cliente reservo más de una vez
prev_cancellations	Numérica	cantidad de cancelaciones
prev_bookings_not_canceled	Numérica	cantidad de reservas no canceladas
booking_week	Numérica	semana del año del check-in
check-in_week	Numérica	semana del año de la reserva
booking_weekday	Categórica	día de la reserva. Ej. Lunes, Martes, etc.
checkin_weekday	Categórica	día del chekin. Ej. Lunes, Martes, etc.
tiene_email	Dicotómica	si proveyó su mail personal
is_cancelled	Dicotómica	si la reserva fue cancelada 1, sino 0. Es la columna Label

Fuente: Elaboración propia