Pontificia Universidad Javeriana

Modelo para la predicción de interacciones recibidas en un centro de atención telefónico

Proyecto de grado

Presentado por: Aura María Benavides Suárez, Juan Pablo Alonso Ospina, Daniel Felipe López, Juan Camilo Ortiz.

Tutores: Jairo Andrés Rendón Gamboa, Stevenson Bolívar Atuesta.

06-2022

1 Entendimiento del Negocio

1.1 Determinación de los objetivos Comerciales

1.1.1 Background

Millenium BPO es una compañía colombiana de tecnología y servicios con más de 20 años de experiencia en el sector de Contact Center y BPO, que busca estar a la vanguardia en la automatización de procesos mediante inteligencia artificial, con lo cual facilita el desarrollo de diferentes organizaciones apoyando componentes importantes de la transformación digital, como lo es el contacto con el cliente.

Análisis de mercado

Un BPO (Business Process Outsourcing) puede ser descrito como un trabajo de servicio subcontratado a un tercero, abarcando una gran gama de actividades como servicio al cliente, tratamiento de datos de entrada, trascripción, digitalización, entre otras (Mann & Graham, 2016). Para 2021, según la presidente ejecutiva de la asociación colombiana de BPO, Ana Karina Quessep, el sector de BPO generó aproximadamente 25.000 empleos más con respecto al año anterior, representa 2,8% de participación dentro del PIB nacional y contó con 605.000 trabajadores en el mismo periodo (Sectorial, 2021).

Adicionalmente, Colombia es considerado como uno de los países más atractivos para la localización de servicios globales debido a sus costos competitivos. Hacia 2021 se lograron ventas alrededor de 12.05 billones de COP, pese a la pandemia actual del COVID 19 hubo un crecimiento del sector del 15.4% y se realizaron exportaciones por 1.475 millones de USD (Asociación colombiana de BPO, 2021). Actualmente en Colombia existen más de 600 empresas dedicadas a prestar el servicio de BPO, el 65% son empresas nacionales y el restante 35% multinacionales o latinas.

En cuanto al comportamiento de las exportaciones para el sector de tercerización de servicios, las cifras del último año se ubican alrededor de US$1.475 millones, según datos de la Asociación Colombiana de BPO, siendo Estados Unidos, España, México, Perú y otros destinos en Latinoamérica los principales mercados de exportación. De acuerdo con datos de la Asociación BPO-Pro, Colombia se encuentra en el cuarto puesto entre los mercados más grandes de Latino América para el Sector BPO, después de Brasil, México y Costa Rica.

Por otro lado, el servicio de Contact Center busca facilitar la interacción que tienen las empresas con sus clientes por lo diversos canales de contacto disponibles (llamadas telefónicas, correo electrónico, SMS, redes sociales, contacto directo) atendiendo las diferentes solicitudes que se puedan generar. de esta manera ofrecen un servicio especializado en este ámbito, a un costo competitivo para el mercado, y que puede ser escalado acorde al tamaño de cada empresa.

Generalmente, una de las problemáticas características de las empresas del sector BPO es la rotación de personal, esta hace referencia a la cantidad de trabajadores que abandonan la compañía, produciendo una serie de costos adicionales, como por ejemplo el entrenamiento, capacitación y formación de la planta activa, ya que, se asigna tiempo de los trabajadores para estas actividades dada la cantidad de horas que deben dedicarse a la enseñanza de los procesos. Los altos índices de rotación de personal son factores que determinan el alcance de los objetivos de las compañías de modo tal que las áreas de Operaciones y Gestión Humana realizan un trabajo continuo parea reclutar personal con periodos de tiempo cada vez más cortos, de tal forma que contar con un sobredimensionamiento no sólo refiere al costo de la mano de obra adicional sino todos los procesos implicados para tener el recurso disponible. El sobredimensionamiento en los agentes implica mayores costos y por ende menor utilidad y perdida en el costo de oportunidad de poder usar a estos agentes en otro canal, por ejemplo. Cuando el efecto es opuesto y se genera un subdimensión de recursos, el problema está asociado al incumplimiento con el contrato a los clientes por el servicio pactado, lo que terminaría afectando la relación comercial con el cliente.

En este sentido, el principal pilar de este sector es el recurso humano cuya participación en el costo de la mano de obra directa (agentes) puede representar hasta un 80% del presupuesto operativo de la llamada (Penagos Arango,2016), por tal motivo una de las necesidades más importantes es reducir el costo de la mano de obra que permite a los centros de BPO incrementar su ahorro para inversión en tecnología o la adquisición de más recursos para atender más interacciones o las debidas interacciones de acuerdo a cada campaña.

Las interacciones se pueden dar de forma multicanal, el análisis de interacciones permite identificar procesos con mayores TMO (Tiempo Medio Operativo) y se generan a partir del momento que se tiene contacto con el cliente. Millenium mide el grado de satisfacción del cliente y de los usuarios por medio de las interacciones realizadas, una interacción son todos los procesos inbound en la empresa, es decir cuando el cliente se pone en contacto con el BPO por cualquiera de los medios de comunicación disponibles, y comienza la interacción desde el momento en el que se responde al mensaje o a la llamada entrante. Los usuarios del sistema serán atendidos, en primer lugar, por una respuesta interactiva de voz, que les presentará una serie de opciones que podrán ser elegidas a través del teclado del teléfono, una vez discriminadas las llamadas pasarán al operador correspondiente. Actualmente el BPO tiene contratos con sus clientes por interacciones y varían dependiendo de la campaña por la cual se esté ofreciendo el servicio.

Millenium cuenta con 48 campañas a 2022 pertenecientes a los sectores de financiero, retail, salud, pensiones, servicios públicos, entre otros. La campaña de estudio, perteneciente al sector financiero, representa el 13.5% total de la facturación (aproximadamente 1.500 millones de pesos mensuales) de Millenium, siendo la tercera campaña más grande con la que opera la empresa seguido de una empresa del sector de telecomunicaciones con 14.6% y una empresa de servicios públicos con 13.7%. Estas campañas (telecomunicaciones y servicios públicos) no son manejadas en su totalidad por Millenium BPO y representan entre 10% y 15% de la facturación total de las empresas, por otro lado, la campaña de créditos estudiantiles es 100% ejecutado por Millenium BPO y por lo tanto los datos están completos y disponibles para el estudio.

Una de las formas más verídicas para medir la eficiencia de las campañas en un BPO es por medio de 4 KPI: Llamadas realizadas, Cantidad de leads procesados, First Call Resolution y Tiempos de

inactividad y respuesta. Millenium BPO añade otro KPI importante a nivel interno por medio del cual se analiza eficiente y financieramente el desarrollo de la actividad, este indicador corresponde a la formula *nómina / facturación*, el cual según políticas de la empresa debe mantenerse en niveles del 65% mínimo para cada caso. Para la campaña de estudio el NF de febrero del 2022 fue del 74.1%, la compañía de servicios públicos tiene un NF del 70.4% y la de telecomunicaciones un NF del 71%.

El canal telefónico de la campaña de créditos atiende el 36% de las solicitudes presentadas por los clientes, los otros medios son, unidad gestora con el 39% (medios digitales), tutelas, IES (Instituciones de educación superior), mesa de legalización, FTE y canal online 16% y CET (Centro de atención especializados y personalizados) de manera presencial 10%.

Actualmente la campaña de estudio funciona con 250 trabajadores que hacen parte de la nómina del NF en los cuales se incluyen los agentes, las personas de mejora continua y los formadores, en febrero el costo por mantener esta nómina para la campaña de estudio fue aproximadamente de 1.120 millones de pesos mensuales, el área administrativa no hace parte de la función del NF, por lo tanto, uno de los factores fundamentales a la hora de gestionar un buen NF es la cantidad de personas precisas. La capacidad de atención de un centro de BPO depende fundamentalmente de la cantidad de personas que están atendiendo las solicitudes de los clientes por medio de los diferentes canales, de modo que la planificación del recurso humano es un medio para llegar al equilibrio entre costos y servicio. En otros estudios se ha usado el método CAT como planificador del comportamiento dinámico del personal mediante la teoría de colas (Koole y Mandelbaum, 2002). Según este modelo el número de llamadas cada hora constituye la carga del CAT (Confianza- Actitud- Tono) y es la variable más importante para determinar el número de operadores óptimo para una buena calidad en el servicio.

La cantidad óptima de personas en operación por campaña se hace indispensable para la gestión del NF y además para el cumplimiento del contrato con los clientes el cual establece lo siguiente:

En el contrato con el cliente el PROVEEDOR obliga a prestar al CLIENTE el servicio de centro de atención al usuario a través de los recursos técnicos y humanos necesarios para que los usuarios del CLIENTE puedan realizar trámites y gestionar la comunicación, uno de los deberes del proveedor es responder con calidad de trabajo-diligencia y entre sus cláusulas de permanencia se contempla lo siguiente:

“El tiempo de espera de las llamadas entrantes será de 15 segundos como máximo. Este tiempo de espera deberá cumplirse en el 80% de las llamadas. En el momento de máximo volumen de afluencia de llamadas, las que no puedan ser atendidas, serán desviadas a un contestador automático. El PROVEEDOR garantiza la atención del 93 % de las llamadas”.

Desde el punto de vista de gestión los dos principales problemas de un BPO corresponden a las falencias en la programación de personal y enrutamiento de personal, el problema de la programación de agentes se puede solucionar por medio de tres fases: pronosticar la carga de trabajo, traducir la carga de trabajo en metas y recompensas y programar agentes teniendo en cuenta el volumen por turnos de llamadas. Dentro del espectro de la inteligencia artificial podemos encontrar a la analítica predictiva como una de las más desarrolladas con la ayuda de la exploración de los datos estructurados y no estructurados, la idea de la predicción es mejorar los tiempos de los procesos y la cantidad de los agentes vinculados, mejorando la satisfacción del cliente y la percepción que este tiene sobre el servicio. Una predicción más precisa en un BPO equivale a contar con la cantidad de agentes con las

habilidades apropiadas en el momento preciso, y de esta forma lograr una experiencia exitosa en el usuario.

Dada la naturaleza del negocio, las diferentes campañas que tiene Millenium BPO presentan un comportamiento único que no se puede generalizar, es decir, el histórico de llamadas y las horas en que estas se presentan difiere en gran medida a campañas de otros sectores económicos. Por ejemplo, las campañas de servicios públicos presentan picos cuando se presentan lluvias y campañas de pagos de facturas presentan picos finalizando la jornada laboral (6:00 PM - 8:00 PM).

Propuesta de valor

Actualmente, Millenium BPO desarrolla soluciones completas de automatización conversacional, usualmente incluyen bots que varían dependiendo del canal. Dichos modelos no se limitan a la clasificación mediante procesamiento de lenguaje natural (NLP por sus siglas en inglés) sino que incluyen CRM (Plataforma de Customer Relationship Management) y RPA (Robotic Process Automatization), los canales por el cual se desarrolla la interacción y los aplicativos del cliente necesarios.

Funcionamiento de un BPO

Generalmente, los procesos de un BPO se pueden dividir en dos grandes campos. Los primeros son los procesos de “Front Office” que cuentan con servicio al cliente, venta, soporte técnico, recobro, cobranza, retención, mesa de ayuda, encuestas de satisfacción, entre otros. Por otro lado, se tiene procesos “Back Office” que pueden abarcar la gestión de recursos humanos, facturación, cartera de finanzas, contabilidad, analítica de negocio, análisis de información y CRM (Asociación colombiana de BPO, 2021).

El proceso de planeación de los recursos se basa en 3 pasos:

1. Pronostico:
   Para lograr realizar la correcta planeación de los recursos es necesario contar con un método de pronóstico con el mejor ajuste posible al comportamiento de la demanda.
2. Estimación de Recursos:
   El principal componente de un BPO corresponde al recurso humano, cuya participación en el costo de la mano de obra (agentes) representa actualmente para la compañía entre el 40 y 50%.
3. Programación de agentes:
   Para que la programación, que es la etapa final sea un éxito, se debe garantiza que el pronóstico sea lo más ajustado a la realidad de forma que los recursos sean lo más productivo posible y que en temas del indicador NF se vea reflejada la eficiencia.

Funcionamiento general de BPO-Centro de atención telefónica

Para el caso de los Centro De Atención Telefónica, la función principal es la venta de algún tipo de producto o servicio, realizar encuestas, recibir llamadas o solicitudes PQR, toma de pedidos, y actividades que permiten establecer comunicación con los clientes, proveedores o socios comerciales.

Las características principales de los Centro De Atención Telefónica son:

Ilustración 1. Principales características de los centros de atención telefónica.

Uno de los más importantes beneficios de este tipo de empresas, es que permite ahorrar la inversión que nos puede suponer comprar el equipo necesario y contratar a los agentes de venta directa como sucede en los modelos de negocio Outbound. Por otra parte, contar con agentes especializados permite atender llamadas más rápido, contar con un equipo experimentado que brinda al cliente un trato agradable y atención personalizada, lo que se traduce en satisfacción del cliente. En los últimos años, varios de los servicios de los BPO han sufrido cambios abruptos como consecuencia de la implementación de inteligencia artificial en sus procesos, un ejemplo de esto son los conocidos “chat bot” presentados como una alternativa de agente que permite interpretar conversaciones con clientes y solucionar los problemas más comunes de los usuarios para aumentar la eficiencia y rentabilidad. (Europa Press, 2019).

La metodología actual para pronosticar el número de llamadas entrantes al BPO es el promedio móvil, con tres semanas de histórico. Se utilizan datos por día y hora de la semana y se utiliza la demanda de los 7 últimos días de llamadas, con el pronóstico de agentes del tiempo disponible diario y semanal el programador elabora la programación de agentes requeridos manualmente con el uso de la herramienta Excel, donde lista los agentes con disponibilidad y demanda requerida por periodo. De esta manera, por medio de ensayo y error se asignan los agentes y se revisa constantemente no exceder con los requerimientos. Hoy en día el modelo actual tiene una precisión muy baja, lo cual termina siendo problemático al no ajustarse a las necesidades de la campaña de estudio desde que se ajusta a series de tiempo con demandas estables sin considerar tendencias o estacionalidades.

El servicio se presta de lunes a sábado, los agentes trabajan 8 horas diarias para un total de 48 horas semanales, el costo para contratar a un agente corresponde a un valor de $66.932,5 diarios con prestaciones incluidas. Por políticas corporativas no es permitido recargos por trabajos suplementarios y de acuerdo con la reglamentación Colombiana “el número de horas de trabajo diario podrá repartirse de manera variable durante la respectiva semana y podrá ser de mínimo cuatro (4) horas continuas y hasta diez (10) horas diarias sin lugar a ningún recargo por trabajo suplementario, cuando el número de horas de trabajo no exceda el promedio de cuarenta y ocho (48) horas semanales dentro de la jornada ordinaria de 6 a.m. a 10 p.m”.

En la campaña estudiada, la cantidad de agentes cambia semanalmente de forma tal que se cumplan con los pronósticos de llamadas por promedio móvil que se realiza actualmente moviendo un grupo de agentes entre canales.

1.1.2 Objetivos del negocio

Mejorar el indicador de rentabilidad basado en la relación de nómina sobre facturación (Nomina/Facturación) de la campaña, partiendo de un pronóstico semanal de llamadas que ingresarán, de forma que se pueda hacer una asignación de recursos de personal óptima y eficiente. Este análisis estará enfocado en los agentes de las campañas del sector financiero, los cuales tienen asignados el 10% de la operación de la firma.

1.1.3 Objetivos específicos

- Desarrollar un modelo que permita la programación adecuada del recurso humano y los requerimientos de la demanda.
- Asegurar que se pueda cumplir con los planes de llamadas y evitar abandonos o llamadas no efectivas en la campaña de créditos estudiantiles, impidiendo el sobre o sub-dimensionamiento de recursos.

1.1.4 Criterios de aceptación del negocio

Para el 2022 la alta gerencia propuso que todas las campañas de la compañía debían tener una relación de nómina por facturación del 65%, es decir, que entre más cercano a este porcentaje el análisis del trabajo será más efectivo.

1.2 Evaluación de la situación actual

1.2.1 Disponibilidad de recursos

Los recursos serán evaluados en tres categorías:

- Hardware y Software.
- Datos y Almacenamiento.
- Personal (recurso humano).

Recurso Hardware y Software

Para el desarrollo del proyecto se trabajará con equipos de cómputo con especificaciones mínimas como RAM de 8 gb, Procesador Inter Core I5 o Ryzen 5, los cuales se consideran básicos para procesamiento de datos masivos.

En cuanto a Software se implementarán las herramientas Python usando el IDE Anaconda y Jupyter NoteBook para la programación.

Recurso Datos y Almacenamiento

Las bases de datos son suministradas por Millenium BPO para el análisis y ejecución de los modelos. La base con todos los registros cuenta con un peso aproximado de 1.2 Mb en formato “xlsx” comprendidos desde enero 2021 a enero 2022.

Recurso de Personal

Los 4 integrantes del proyecto serán los responsables por parte de la minería de datos y el desarrollo de una propuesta de valor. De parte de la empresa Millenium BPO se cuenta con el apoyo del gerente de planeación y el gerente de datos no estructurados.

Requerimientos, supuestos y restricciones del proyecto

1.2.1.1 Requerimientos

Requerimientos del Negocio

Millenium BPO requiere definir un modelo que permita disponer los recursos eficientemente para la campaña en estudio, de forma que se pueda asegurar una generación de gasto más competente y dar cobertura completa para cada llamada recibida o realizada.

Para poder cumplir con el requerimiento del negocio es esencial cumplir con la política de seguridad de información de la compañía garantizando por medio de un contrato de confidencialidad el manejo responsable de esta.

Se debe establecer un cronograma de trabajo que expone las actividades, respectivos responsables y fechas de entrega, considerando planes de seguimiento y validación de los resultados.

Requerimientos de calidad

Dentro de las métricas o las necesidades expuestas por la compañía se está buscando un modelo que permita tener una predicción de la cantidad de interacciones recibidas lo suficientemente robusta, que se encuentre dentro de un margen de error aceptable y que permita alcanzar los objetivos del negocio.

Dicho esto, para asegurar la calidad del modelo también se debe tener en cuenta que la disponibilidad de los datos e información se genere conforme al cronograma y que se encuentre disponible para todos los miembros del proyecto.

1.2.1.2 Supuestos

Supuesto de Agentes

- La cantidad de agentes en actividad no contempla tiempos de vacaciones, ni descanso u otros escenarios esporádicos donde el trabajador no cumple con su labor el día determinado.
- Los servidores se consideran idénticos y su capacidad es la misma para todos los agentes.
- Los agentes disponibles cuentan con las habilidades necesarias para llevar a cabo de manera efectiva la resolución de la llamada, lo que nos indica que no existe una clasificación interna por tipo de habilidades para los agentes.

Supuesto Llamadas

- La duración promedio de la llamada en interacción es de 592.82 segundos
- La cantidad promedio de espera de una llamada para ser atendido es de 104.133 segundos
- No se consideran las caídas del sistema de telefonía que ocasione una segunda llamada por parte del usuario, es decir, que pueden ser interrumpidas por caídas en el sistema, mantenimiento o indisponibilidad de la línea telefónica por parte del BPO o el usuario

Supuesto Datos

- La información suministrada deberá y será tomada como datos reales de la operación, de forma que los resultados obtenidos sean válidos y puedan ser implementados con mayor facilidad para la compañía.
- Los valores nulos de la base de datos deben ser revisados y analizados, de forma tal que se les pueda aplicar un correcto tratamiento, ya sea omitiéndolos (dependiendo de la cantidad y de la variable donde se encuentren) o realizando una estimación por los valores ya conocidos por la misma base.

1.2.1.3 Restricciones

Restricción de Datos

Los datos solo serán suministrados por el área de Reportería y estos deben ser aprobados por el jefe de Planeación para asegurar protocolos de seguridad y manejo de la información, de acuerdo con el contrato de confidencialidad que entró en vigor para la ejecución del proyecto.

La base de datos para el proyecto tiene gran cantidad de información por lo que su tamaño es alto y el costo computacional que se requiere también lo es, por otro lado, el número de columnas deberán reajustarse puesto que contiene con información repetida, incompleta o innecesaria, lo cual se determinará en el momento de hacer el análisis de esta.

Restricciones legales

Dada la naturaleza de los datos, y el acuerdo de confidencialidad, la información y los resultados expuestos en este documento sólo podrá ser consultada y revisada por los miembros del proyecto, personal autorizado de Millenium BPO y docentes a cargo de liderar los proyectos de grado de la Pontificia Universidad Javeriana.

1.2.2 Riesgos y Contingencias

Se evalúan tres posibles riesgos, los cuales se exponen a continuación junto con su respectiva contingencia:

Tabla 1. Riesgos determinados.

TIPO DE RIESGO	DETALLE DEL RIESGO	CONTINGENCIA
Cronograma	Si el proyecto no se ejecuta en los plazos y términos del cronograma.	• Para evitar el incumplimiento del cronograma se debe hacer un seguimiento semanal a los compromisos acordados. • Establecer un cronograma apropiado para considerar el riesgo de extensión y así asegurar los tiempos de entrega.
Datos	Los datos suministrados pueden ser de baja calidad o sin la cobertura suficiente para el modelo.	• Dado que ahí, el tipo de la interacción es asignada por los agentes manualmente, esta será tomada como datos reales. • En caso de que los datos no sean suficientes se puede solicitar ampliar la cobertura de los datos.

Resultados	Los resultados iniciales son menos representativos de lo esperado.	• Si los resultados son menos representativos de lo esperado, se deberá asegurar la calidad del modelo y sugerir que no se realicen ajustes a la operación actual.

1.2.3 Terminología

- Se encuentra disponible en anexos 8.1 (pagina 35).

1.2.4 Costo Beneficio

El beneficio de tener una predicción acertada en la cantidad de interacciones de un BPO consiste en que al tener la cantidad de recursos suficientes y necesarios no se incurre en sub-asignación ni sobreasignación del personal, ambos son problemáticas de suficiente impacto para la campaña.

En el caso de la sub-asignación se corre el riesgo de que el personal de la campaña no esté en la capacidad de recibir la cantidad de interacciones entrantes, en consecuencia, la relación inicial de nómina sobre factura sería baja, ya que, la ocupación del personal estaría a tope, sin embargo los clientes se blindan en este tipo de circunstancias y contractualmente imponen multas al llegar a presentarse casos de interacciones no atendidas, para este caso particular se cobraría un 8% del total de lo facturado.

Por otro lado, al presentarse una sobre-aginación de recursos, el personal no presentaría una ocupación tan alta, en consecuencia, el indicador de nómina sobre factura incrementaría. En la práctica, este escenario no es frecuente debido a que históricamente la mayoría de los costos vienen por sub-asignación.

Durante enero de 2022 aproximadamente el 44% de las interacciones entrantes por el agente virtual del canal telefónico, finalizaron en abandonos por parte del usuario, debido a un mal dimensionamiento de los recursos disponibles, consecuentemente, los usuarios tienden a abandonar este tipo de llamadas debido al alto tiempo de espera. Puntualmente, en enero de 2022 se dejaron de facturar cerca de 144.000.00 COP por este concepto. Finalmente, se espera que el proyecto ayude a disminuir estas pérdidas mensuales de dinero, sabiendo que en reunión con la gerencia general se consensuó un límite máximo de pérdidas de 50.000.000 COP mensuales.

Tabla 2. Muestra de cantidad de llamadas contestadas por agente

semana	*Agentes*	*Cantidad de llamadas contestadas*	*Promedio de llamadas contestadas por agente*
1	173	5059	29
2	48	1416	30
3	175	5129	29
4	241	7040	29
5	65	1910	29

Tabla 3. Cantidades promedio de llamadas contestadas vs abandonados

Cantidad promedio de llamadas contestadas diariamente	5188
Cantidad de abandonos diario promedio	442

Actualmente, un agente contesta aproximadamente 30 llamadas diarias, teniendo en cuenta que la cantidad de abandonos diarios es del 8% de la cantidad del total de llamadas contestadas las perdidas diariamente están entre 2 y 3 millones de pesos.

El costo del salario de un agente actualmente es de 1’606.380 con prestaciones incluidas, lo que equivale un valor de $66.932,5 por día laborado. El costo por abandono de llamada es de $4.950, que se dejan de facturar, mientras que el costo por contestar una llamada es de $2.231, lo que equivale a que una llamada genera aproximadamente una ganancia del 45%.

1.3 Objetivos de la analítica

1.3.1 Objetivos de la minería de datos

Utilizando datos históricos sobre el comportamiento de las interacciones recibidas para el canal telefónico, generar un modelo de predicción del volumen de interacciones (cantidad de llamadas) que puede recibir la operación para tomar decisiones en cuanto a cómo organizar los recursos de personal.

1.3.2 Criterios de aceptación de la minería de datos

Se requiere que el modelo tenga una precisión suficiente para alcanzar el objetivo de negocio (reducir mínimo en un 1% el indicador N/F) con el fin de determinar el volumen de interacciones que llegarán al centro de atención telefónico, teniendo en cuenta que la campaña asignada, al ser el único aliado, se cuenta con la totalidad de los datos que viene por el canal telefónico.

1.4 Calendario

Tabla 4. Calendario de actividades.

			ENERO		FEBRERO				MARZO				ABRIL			
N°	Actividad	Recursos	24	31	07	14	21	28	07	14	21	28	04	11	18	25
1	Entendimiento del Problema y Datos Iniciales	Definición de objetivos Evaluación de la situación actual Definir los objetivos y alcance de la minería de datos Revisión de la data inicial														
2	Entendimiento de los Datos	Descripción de los datos Entendimiento del comportamiento de los datos Exploración de los datos y definir la calidad de la información														
3	Preparación de los datos	Seleccionar los datos que realmente aportan valor al modelo Limpiar o depurar datos que puedan afectar la estructura o el resultado del modelo Seleccionar el modelo que se va a trabajar para resolver el problema														

			MAYO				JUNIO				
N°	Actividad	Recursos	04	11	18	25	01	08	15	22	29
4	Modelación	Generar la estructura de prueba para el modelo Construir el modelo Ejecutar y revisar los parámetros del modelo para ajustar y mejorar los resultados									
5	Evaluación del Modelo	Evaluar los resultados Evaluar los resultados y compararlos con los criterios de aceptación de la minería de datos Revisar el proceso si se requiere realizar algún ajuste en el modelo Definir próximos pasados basados en los resultados									

2 Entendimiento de los datos

2.1 Data inicial

Inicialmente, se recolectaron los datos a partir de un software especializado para capturar información cuantitativa y cualitativa sobre las llamadas que usa la empresa para prestar los servicios de una de las campañas de más impacto para Millenium BPO, perteneciente al sector financiero, esta información es presentada en 4 archivos separados por rango de tiempo de 2 años, es decir:

- Primer semestre de 2020.
- Segundo semestre de 2020.
- Primer semestre de 2021.
- Segundo semestre de 2021.

Los cuatro archivos se encuentran en formato xlsx, y estos fueron suministrados directamente por la compañía, se clasificaron los datos en Demográficos y de Negocio, los cuales están distribuidos de la siguiente forma:

- Datos demográficos:
    1. Ubicación del cliente
    2. Tipo de cliente
    3. Es menor de edad
- Datos de negocio
    4. Fecha de interacción
    5. Hora de interacción
    6. Canal de atención
    7. ASA (Seg)
    8. AHT (Seg)
    9. Hold
    10. Tipo de caso
    11. Proceso inicial
    12. Tipificación inicial
    13. Subtificación inicial
    14. Tipificación de canal
    15. Subtificación de canal
    16. Estado de la llamada
    17. Cola de la llamada
    18. Que tan satisfecho se encuentra con el servicio
    19. Recomendaría la entidad Financiera
    20. Motivo de calificación buena
    21. Motivo de calificación Mala

2.2 Descripción de los Datos

Como se mencionó en el punto anterior (2.1 Data Inicial), los datos con los que se cuenta para trabajar constan de 21 columnas y al consolidar todos los archivos en una única base de datos se tiene en total 3.367.561 de registros.

- La tabla 3 que contiene la descripción y detalle de cada una de las variables, se encuentra disponible en los anexos sección 5.2.

En resumen, se tienen:

- 2 variables relacionadas con el tiempo
- 4 variables numéricas
- 15 variables categóricas

Por lo que se requiere revisar el comportamiento de estas variables y poder definir que variables son relevantes para la elaboración del análisis y modelado.

2.3 Exploración de los Datos

En el histórico de llamadas se puede ver un patrón, con picos ascendentes y descendentes que se mantiene casi de forma regular a lo largo de la serie, sin embargo se evidencia que desde 2021 el volumen de llamadas ha bajado considerablemente sobre todo al iniciar la segunda mitad de 2020, aunque se puede ver que el comportamiento de las llamadas que se reciben es similar cada 6 meses donde inicia con un pico ascendente y termina un pico descendiente y vuelve a subir radicalmente.

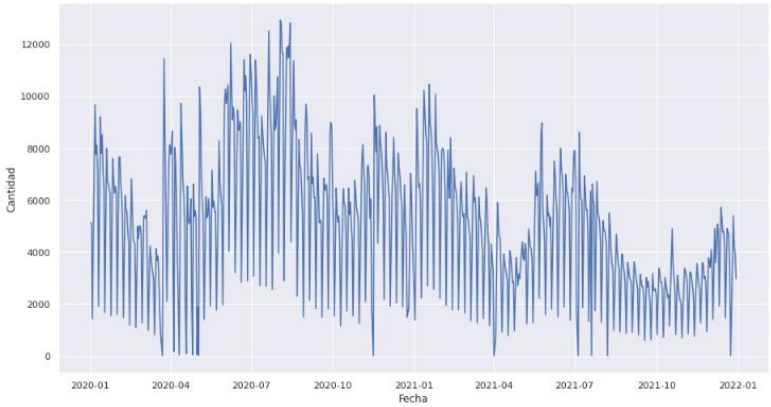


Ilustración 2. Cantidad de llamadas distribuidas en orden cronológico

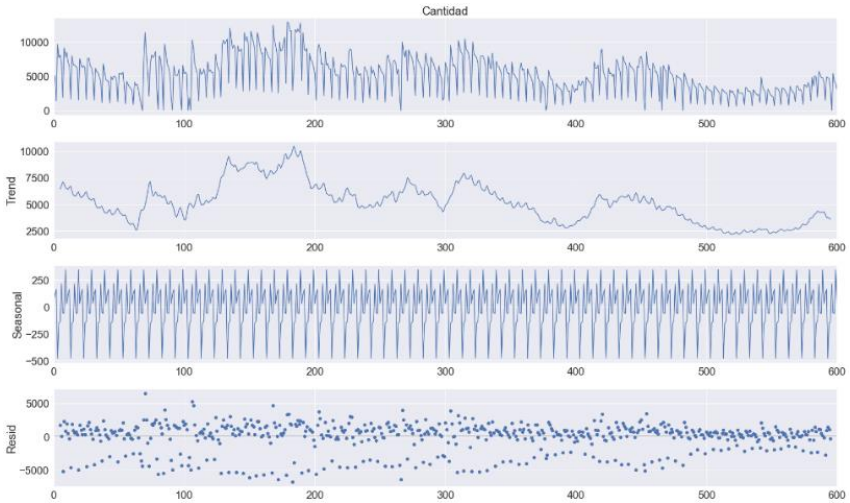


Ilustración 3. Descomposición de la serie de tiempo.

Cabe resaltar que la primera mitad de la serie de tiempo tiene un comportamiento diferente, específicamente desde el mes de marzo debido a que fue cuando se hizo efectiva la cuarentena causada por el Covid-19.

	Cantidad	Cant_Abandonadas	%Llamadas_Abandonadas
count	601.000000	601.000000	601.000000
mean	5188.925125	442.094842	0.078474
std	2777.443804	791.941296	0.150671
min	1.000000	0.000000	0.000000
25%	2981.000000	21.000000	0.006369
50%	5279.000000	77.000000	0.018685
75%	6897.000000	427.000000	0.081702
max	12926.000000	5612.000000	1.000000

Ilustración 4. Distribución de llamadas contestadas vs abandonadas por mes (izquierda). Tabla de estadística descriptiva de llamadas abandonadas (derecha).

La anterior ilustración refleja la cantidad de abandonos diarios de llamadas, donde se evidencia la correlación directa con el número de llamadas entrantes. El promedio de llamadas abandonadas es del 7.84% diariamente con una desviación estándar del 15.6%. El mínimo de llamadas abandonadas ha sido de 0, y el día que más dejaron de contestarse llamadas se reportó un abandono del 30% del total de llamadas entrantes.

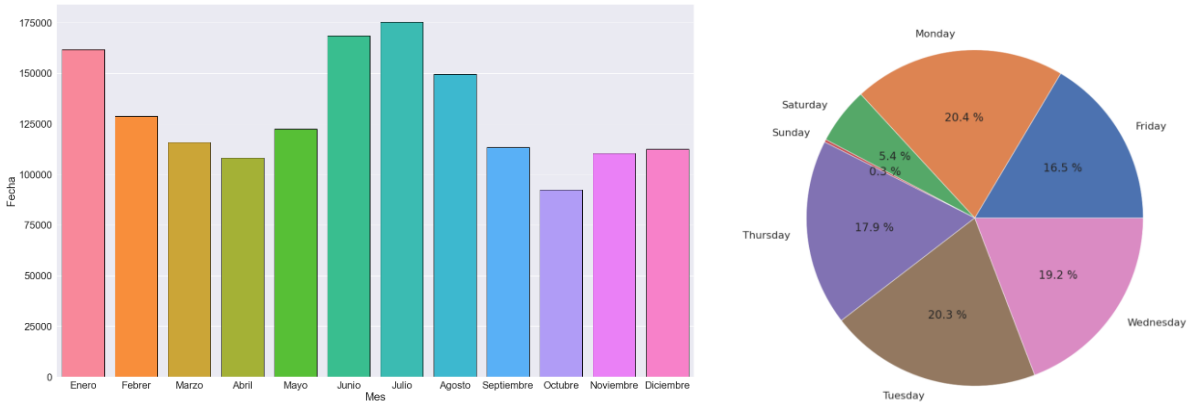


Ilustración 5. Distribución de llamadas por mes del año (izquierda). Participación de llamadas por día de la semana (derecha).

Como se mencionó anteriormente, el pico más alto de llamadas sucede cada 6 meses en promedio lo cual se aprecia en la Ilustración 5, en la cual para los meses de enero, junio y Julio son los meses con mayor volumen de llamadas, teniendo en promedio 129.939 llamadas mensuales en los dos últimos años. Actualmente el promedio de llamadas diarias es de 5.188, el cual esta principalmente concentrado en los lunes, martes y miércoles, los cuales superan las 6.000 llamadas promedio. Los domingos se tiene alrededor de 325 llamadas recibidas en los últimos dos años, las cuales se pueden asociar a un error de la muestra ya que los domingos no se presenta una jornada laboral.

De acuerdo con la Ilustración 6, las franjas horarias con mayor volumen son entre las 9 y las 11 am y entre las 2 y 3 pm.

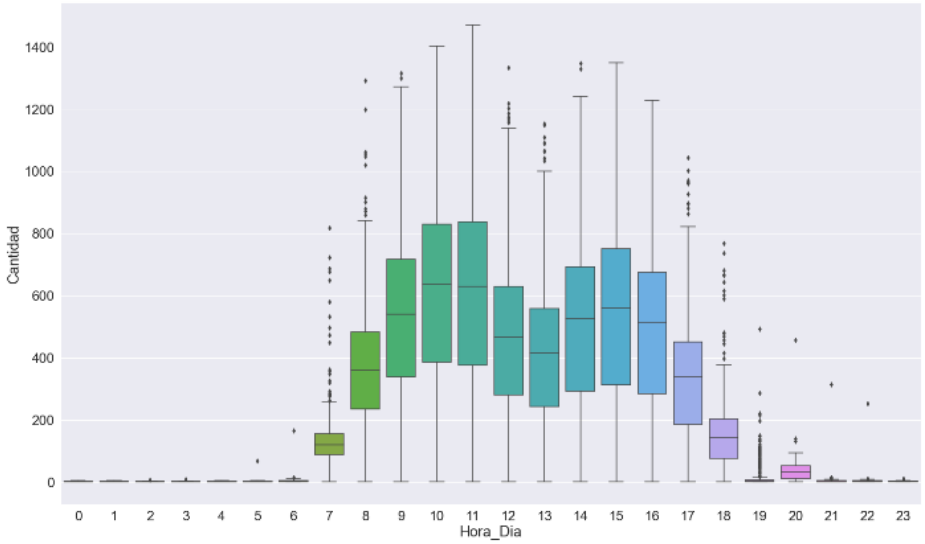


Ilustración 6. Distribución de llamadas por hora del día

2.3.1 Entendimiento de los datos por Departamento

El departamento del cual se reciben más llamadas es Bogotá D.C., y representa el 30.8% de las llamadas diarias recibidas, lo que equivale a 1.431 llamadas diarias y que anualmente son alrededor de 424.381 llamadas. Cabe resaltar que el comportamiento entre los departamentos es similar en cuanto a la distribución de llamadas en función de la media como se aprecia en los gráficos siguientes. Es importante resaltar que a pesar del volumen de llamadas de Bogotá no hay presencia de datos atípicos, contrario a lo que sucede en el caso de Arauca, Magdalena, Norte de Santander y otros más pequeños.

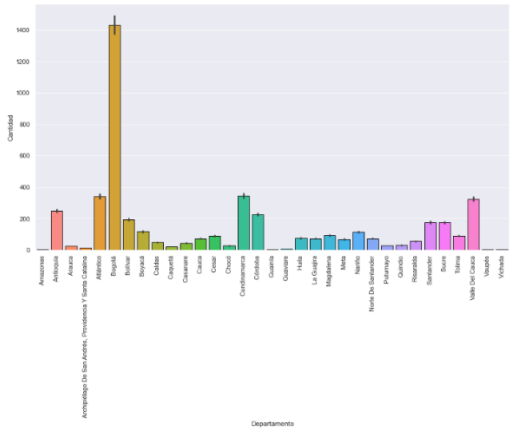


Ilustración 7. Concentración de llamadas por departamento

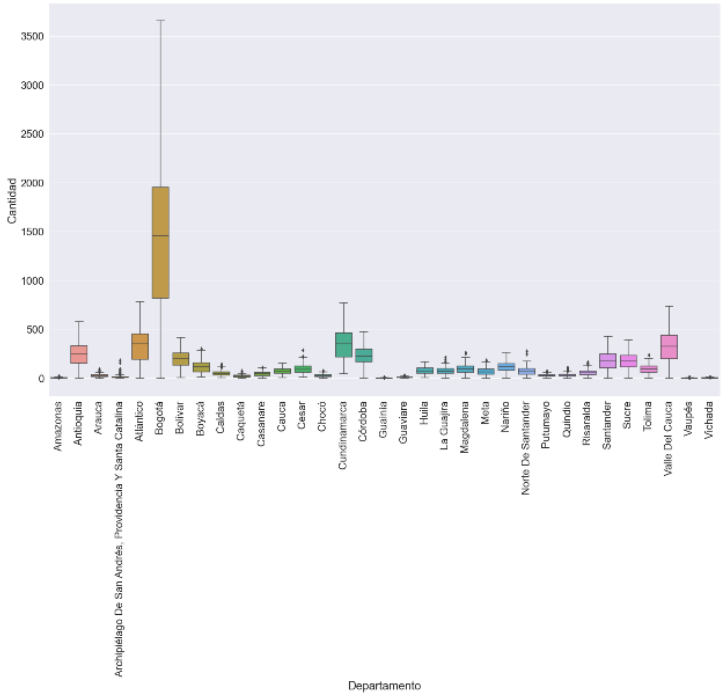


Ilustración 8. Boxplot de la cantidad de llamadas por departamento

2.3.2 Entendimiento de los datos del negocio

Se identifica tres tipos de clientes principales que realizan llamadas, los cuales están clasificados en Beneficiario, Ciudadano e Institución de Educación Superior (IES), aun así en la base de datos se encuentra registros nulos, estos campos se transformaron a cero y de esta forma se crea una nueva clasificación “Cliente no tipificado”, así se determinó que diariamente en promedio 981 de las llamadas no quedan correctamente tipificadas, representando el 18.6% de las llamadas diarias, mientras que los beneficiarios son el 58.5% de las llamadas recibidas.

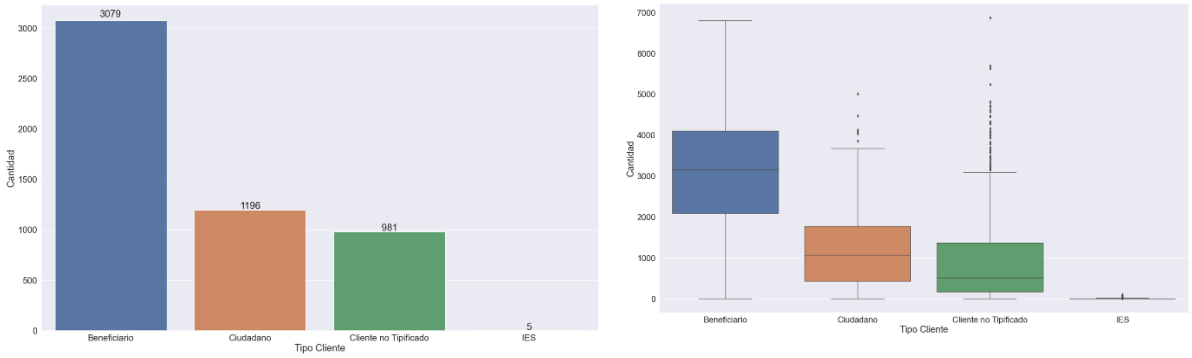


Ilustración 9.Cantidad de llamadas recibidas diarias por tipo de cliente (Izquierda) Distribución de llamadas diarias realizadas por tipo de cliente (derecha).

Es importante ver que en el caso de los clientes que no son tipificados hay mayor presencia de atípicos, lo que siguiere que hay días específicos en los cuales no se registró dicha categorización para un alto número de llamadas recibidas, sin embargo como se evidencia en la Ilustración *10*, cuando en Colombia

inicia el confinamiento como medida preventiva por la pandemia Covid-19, se dispara el número de llamadas que no fueron registradas con una clasificación de clientes, lo que sucede para el periodo de Junio y Julio, donde el confinamiento fue prolongado por más tiempo, elevando el número de llamadas recibidas sin la clasificación correcta de los cliente, lo que no ocurre con los clientes Beneficiarios.

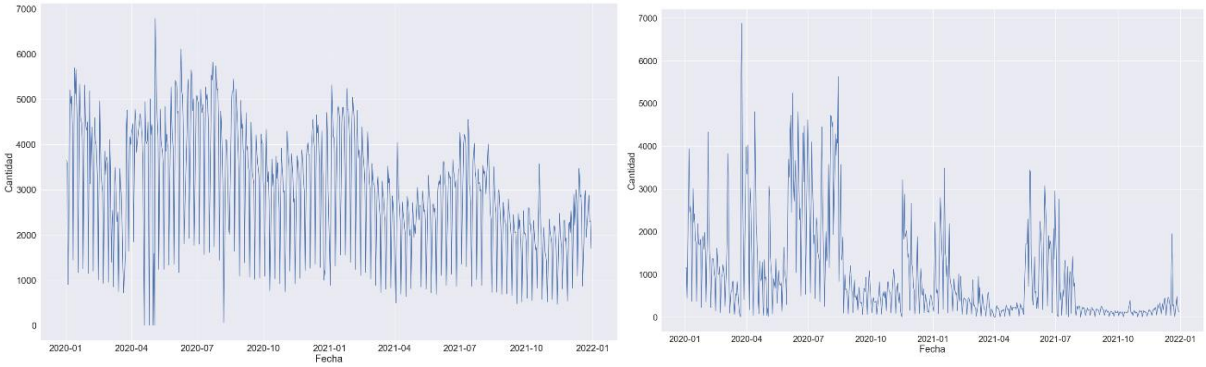


Ilustración 10. Distribución de cantidad de llamadas por día en los clientes beneficiarios (izquierda) y no tipificados (derecha).

También es importante resaltar que recibir llamadas de IES, no es muy frecuente sin embargo para el año 2021 sobre el mes de marzo y abril, se presentó significativamente un volumen dentro de esta categoría, como s evidencia en la siguiente imagen y llama principalmente por causas de “Asesoría general”, “Otorgamiento de producto” y consultar por la “etapa de estudios”.

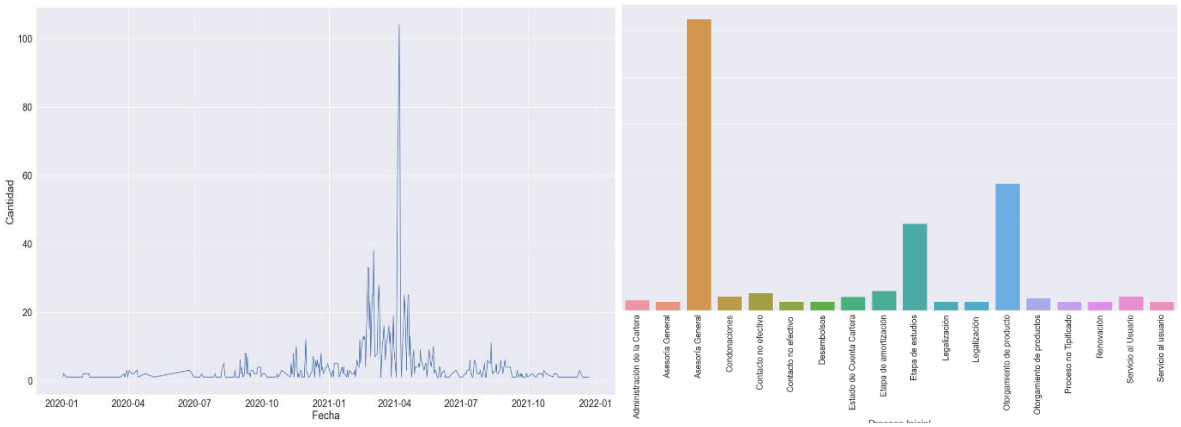


Ilustración 11. Distribución de cantidad de llamadas por día en IES (izquierda). Proceso inicial por el cual se comunican las IES (derecha).

Adicionalmente, se desarrolló un análisis de la nómina sobre factura teórica que debería tener la campaña, bajo el funcionamiento actual, y teniendo en cuenta los siguientes supuestos:

- La nómina se calcula de una manera ideal, en donde todos los agentes se desempeñan de la misma manera y con una eficiencia igual.

- De la totalidad del tiempo trabajado por los agentes, estos dedican un 60% de su tiempo a contestar llamadas (dato suministrado por el gerente de operaciones de la campaña), y resto del tiempo se considera labores operativas o tiempos muertos.
- La nómina para cada uno de los agentes es la misma, el SMLV, que puede ser consultado en la sección de costo beneficio.
- Teniendo en cuenta que la asignación de nómina se puede realizar únicamente una vez por semana (programación desde la semana anterior), se calculó el número de agentes necesarios para contestar las llamadas entrantes de la base (considerando que se lograrían contestar las llamadas abandonadas) y se tomó como base el mayor número de agentes por semana, es decir el número de agentes disponibles va a ser una constante por semana y será el suficiente para satisfacer la demanda del día con mayor flujo de llamadas.
- El valor de la nómina sobre factura calculado es un mínimo ideal, al cual se podrá llegar si se contestan todas las llamadas entrantes y la ocupación de los agentes es la esperada en todo momento. Este cálculo no contempla incapacidades, renuncias, abandonos de puesto, bajo rendimiento ni similares (que ocurren en la realidad) y solo representa el mínimo teórico al cuál sería capaz de llegar la campaña si opera perfectamente.

Ahora, en la Ilustración *12* se muestra el cálculo del N/F teórico realizado y segmentado por día, donde se ve que el mínimo se encuentra sobre los lunes, al tratarse de los días en los que más flujo de llamadas se evidencia. Adicional, se realiza el cálculo global del N/F teórico para la campaña y el resultado es un valor de 51.3%.

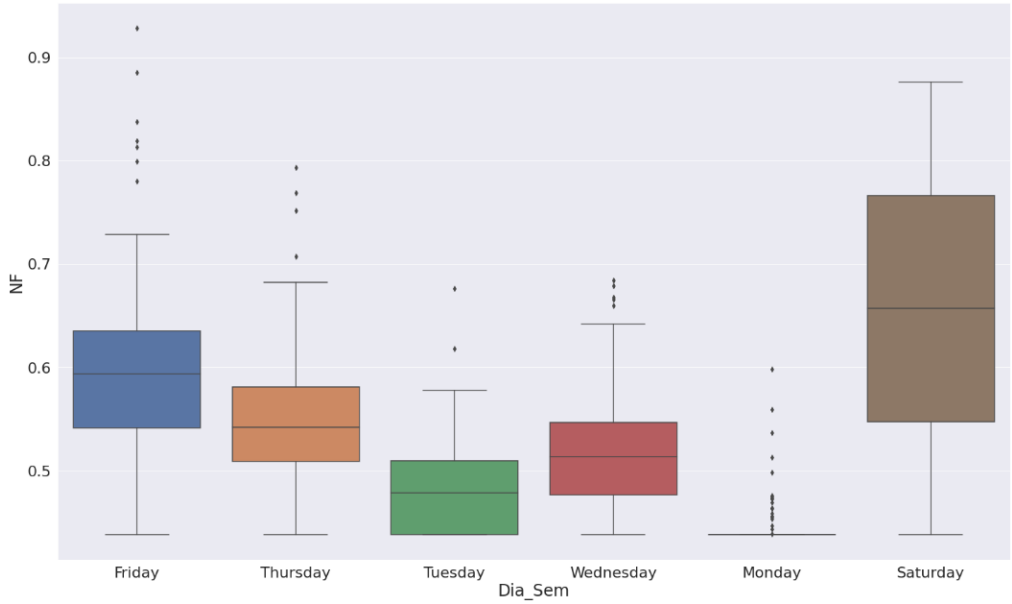


Ilustración 12. Nómina sobre facturación teórico e ideal segmentado por días.

2.4 Calidad de los Datos

Al hacer la validación principal de los datos, se encuentra que 10 de las 21 variables de los datos no cuentan con valores perdidos, las variables son:

- Fecha

- Hora
- Canal
- Hold(seg)
- Tipo cliente
- Tipo de caso
- Proceso inicial
- Tipificación Canal
- Estado Llamada
- Cola

Mientas que las variables con mayor cantidad de valores perdidos son:

- Que tan satisfecho está con el servicio (2.332.787 registros faltantes)
- Motivo Calificación Buena (2.527.881 registros faltantes)
- Motivo Calificación Mala (3.047.538 registros faltantes)
- Recomendaría la entidad financiera (2.345.492 registros faltantes)

Así también se encuentra que la variable "Tipo de Caso", presenta valores con diferencias ortográficas entre los valores únicos que puede tomar este campo, es decir, se tiene el valor "Petición" y "Peticion", los cuales son el mismo atributo, pero la diferencia radica en la tilde, por lo que para este caso se requiere normalizar la estructura de estos valores de forma que se puedan consolidar correctamente.

3 Preparación de los Datos

3.1 Selección de los datos

Para la selección de los datos, se tuvieron en cuenta los objetivos de negocio y se descartaron las siguientes variables que no son relevantes para el análisis:

- HOLD (seg)
- ASA (seg)
- AHT (seg)
- Canal
- Ubicación
- Menor de Edad
- Tipificación inicial
- Subtipificación inicial
- Subtipificación Canal
- Estado Llamada
- Que tan satisfecho de encuentra con el
- Motivo Calificación Buena
- Motivo Calificación Mala
- Recomendaría la entidad financiera

Estas variables contienen información sobre diversas tipificaciones y subtipificaciones de las llamadas que se utilizan para procesos internos de Millenium BPO. También, hay información sobre la

satisfacción del cliente con respecto a la llamada y servicio proporcionado que no aportan información que aporte a la predicción de volúmenes de llamadas asociado al objetivo de negocio.

Sin embargo, se explorará la opción de realizar el modelo únicamente tomando como variable la cantidad de interacciones, de forma que se pueda optar por una mejor solución.

3.2 Limpieza de los datos

Con fin de mejorar la calidad de los datos (analizada anteriormente) se tomaron en cuenta las siguientes acciones:

- Para las variables *Proceso Inicial, Tipo de caso y Tipo Cliente* con datos faltantes se imputaron las categorías *Proceso no tipificado, Caso no tipificado y Cliente no tipificado* respetivamente.
- Se eliminaron las fechas que tiene por día de la semana el domingo ya que por conocimiento de negocio este día no se presta el servicio.
- La variable fecha se trasformó a índice de la tabla para poder aplicar los modelos propuestos.

3.3 Construcción de los datos

Se construyeron las siguientes variables:

- *Hora_Dia:* Es una variable categórica que establece la hora del día (en un rango de 1 a 24)
- *Cantidad:* Es una variable que contiene el conteo de la cantidad de llamadas en cada día, día de la semana y mes.
- *Jornada:* Variable para determinar si la llamada registrada se realizó por la mañana, por la tarde y por la noche.

3.4 Integración de los datos

Para la integración de los datos se unieron las 4 tablas expuestas en la sección 2.1 de “data inicial”, concatenando todas las tablas, ya que estas presentan la misma estructura, pero contienen la información de diferentes periodos de tiempo. En este caso se mantuvo un orden cronológico según la variable *Fecha*.

3.5 Formato de los datos

En esta sección algunas variables se ajustaron para que la información que presentan este acorde al tipo de dato que se propone trabajar.

- Las variables de *fecha y hora* se convirtieron en formato fecha (datetime).
- Dummificación de las variables *Proceso inicial, Tipificación Canal y Tipo Cliente.*

Las demás variables mantuvieron las características originales con las que se leyó la base de datos en Jupyter Notebook.

Después de realizar el proceso de preparación de los datos para realizar el entendimiento del negocio, se realiza una matriz de correlación entre las variables categóricas que fueron dumificadas y la variable dependiente, para determinar el grado de influencia de dichas variables (ver anexo 8.3) .La única variable con correlación importante para la dependiente es la cantidad de llamadas recibidas, esto significa que el historial de cantidad de llamadas recibidas va a ser fundamental a la hora de predecir la cantidad de llamadas futuras.

4 Modelado

Los pronósticos son métodos de estimación utilizados para la predicción de un acontecimiento, este tipo de pronósticos permiten mejorar la toma de decisiones en las empresas siempre que entre las variables que conforman el estudio se tenga en cuenta el comportamiento histórico de las mismas, las cuales puede suponer la continuidad de una tendencia o algún tipo de cambio debido a una contingencia.

Al tener en cuenta que posiblemente no existe un modelo determinístico que logre predecir con exactitud la variable de interés dada la dependencia que se debe considerar sobre el tiempo, al considerar este fenómeno sobre el pronóstico es posible obtener un modelo que permita delimitar un valor futuro muy cercano al real conocido como un modelo estocástico, es decir, un modelo que matemáticamente reproduzca el comportamiento de una serie de tiempo con el que se puedan efectuar previsiones utilizando el periodo de tiempo anterior correspondiente.

Para cumplir con el objetivo de analítica se utilizaron cuatro métodos de machine learning descritos en las siguientes secciones.

4.1.1 Series de tiempo

Una serie de tiempo es una secuencia cronológica de observaciones de una variable determinada. El primer paso para desarrollar el correcto análisis de una serie es graficarla y esto permitirá descubrir patrones históricos útiles en la predicción. Para ello se deben analizar los siguientes componentes: tendencia, que representa el crecimiento o la disminución de la serie y sus cambios en la media; el ciclo, que es el movimiento ondulatorio hacia arriba y abajo alrededor de la tendencia; la variación estacional, determinada por el patrón periódico que ocurre y se repite en cierto determinado periodo de tiempo y un componente irregular que se refiere a aquella sección de la serie que no sigue un patrón determinado ni reconocible. Existen cuatro componentes que contribuyen a que los cambios observados en una serie de tiempo pierdan valor en sus predicciones, para comprender una serie de tiempo es necesario realizar el análisis de estos componentes:

1. Tendencia secular: La tendencia secular de una serie es el resultado de factores que se producen a largo plazo. En términos generales, la tendencia de una serie de tiempo se considera consecuencia de fuerzas persistentes que afectan el crecimiento o la reducción de esta tales como cambios en la población, características demográficas, cambios en los ingresos.
2. Variación estacional: Este componente representa la variabilidad de los datos dada la influencia de las estaciones, es decir las variaciones que ocurren año tras año en los mismos meses o trimestres con la misma intensidad.
3. Variación Cíclica: Con frecuencia las series de tiempo presentan secuencias alternas de puntos bajos y altos de la línea de tendencia que duran más de un año, esta variación se mantiene después de haberse eliminado las tendencias estacionales o irregulares.
4. Variación Irregular: Este tipo de variaciones corresponden a factores de corto plazo, imprevisibles y no recurrentes que afectan a la serie de tiempo, es el componente que explica el factor aleatorio de la serie y no se puede predecir su impacto.

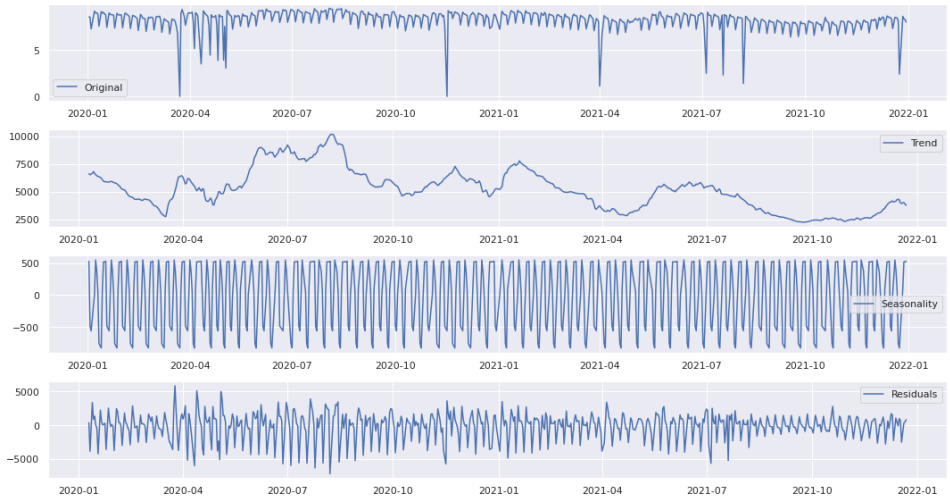


Ilustración 13. Componentes de la serie de tiempo.

El comportamiento de la serie de tiempo para el presente caso la tendencia observada es polinomial grado 3, es decir que los datos fluctúan, el orden 3 suele tener una o dos colinas o valles. La estacionalidad observada es evidente en la tercera línea del grafico donde se observan los picos y los puntos bajos de formas periódicas y regular.

4.1.2 Regresión lineal

La regresión lineal es un campo de estudio que enfatiza la relación estadística entre dos variables continuas (predicción y respuesta). La variable predictora se denota con mayor frecuencia como X y también se conoce como variable independiente. La variable de respuesta se denota con mayor frecuencia como Y que sería la variable dependiente. En caso de que no se pueda aplicar regresión a un estudio significa que las variables no tienen correlación alguna entre ellas. El modelo se expresa de la siguiente manera:

$$Y = Bo + B1X1 + \cdots + B, Xm + E$$

- *Y* es la variable dependiente
- *X1, X2, Xm* son las variables independientes
- *Bo, B1, Bm* son los parámetros del modelo, miden la influencia que las variables explicativas tienen sobre las independientes

Cabe aclarar que para el caso de este trabajo, la regresión lineal estará en función de los periodos anteriores, que funcionaran como la variable predictora.

4.1.3 Random Forest Regression

La regresión Random Forest es un algoritmo de aprendizaje supervisado que utiliza el método de aprendizaje conjunto para la regresión. El método de aprendizaje por conjuntos es una técnica que combina predicciones de múltiples algoritmos de aprendizaje automático para realizar una predicción más precisa que un solo modelo.

Las ventajas que tiene este algoritmo son las siguientes:

- Puede resolver ambos tipos de problemas, es decir, clasificación y regresión y realiza una estimación de ambos frentes.
- Puede manejar grandes cantidades de datos con gran dimensionalidad por lo que se considera un buen método de reducción de dimensionalidad.
- Tiene un método efectivo para estimar datos faltantes y mantiene la precisión cuando falta una gran proporción de los datos.

Esta herramienta crea n cantidad de árboles, conocidos como un conjunto de árboles de decisiones, para crear un modelo que puede emplearse para la predicción. Cada árbol de decisión se crea mediante partes generadas aleatoriamente de los datos (de entrenamiento) originales. Cada árbol genera su propia predicción y vota sobre el resultado. El modelo de bosque considera los votos de todos los árboles de decisiones para predecir o clasificar el resultado de una muestra desconocida. Se trata de algo importante, ya que los árboles individuales pueden tener problemas de exceso de ajuste en un modelo; sin embargo, combinar varios árboles de un bosque para la predicción aborda el problema de exceso de ajuste asociado con un único árbol. (Rodrigo, 2020)

4.1.4 Modelo GradientBoosting

El Gradient Boosting es uno de los algoritmos más populares del machine learning, y es utilizado comúnmente, debido a su alta precisión para encontrar relaciones no lineales entre una variable objetivo y sus predictoras. Este algoritmo es un ensamble donde se crean una serie de modelos base y se combinan con el fin de generar un modelo final fuerte (Masui, 2021). El ensamble se realiza a partir de árboles de decisión, generalmente, no tiene en cuenta residuos sino errores en la predicción.

En el contexto de la predicción numérica, durante el entrenamiento:

- Se asigna un peso a cada registro del conjunto de entrenamiento.
- Mediante cada interacción se aprende un modelo que busca minimizar la suma de los pesos de los ejemplos con mayor error.
- Los errores se utilizan para actualizar los pesos: se aumenta el peso de los registros con mayores errores y se reduce el peso de los que tiene menor error.

Y durante las pruebas:

- Se da más relevancia a los modelos que tienen mejor comportamiento.

4.1.5 Modelo autorregresivo integrado de promedio móvil ARIMA

Existen metodologías desarrolladas por Box y Jenkins que permite predecir los valores futuros de una serie de tiempo usando como base valores pasados de las variables que influyen en el modelo.

$$YT = F(YT{-}1\ ,\ YT{-}2)$$

Este tipo de metodología requiere que la serie sea estacionaria en donde la media, varianza y covarianza permanecen constantes sin importar el momento en el cual se mide. Si se debe diferenciar una serie de tiempo *d* veces para hacerla estacionaria y luego aplicar el modelo ARMA (*p, q*), se dice que la serie de tiempo original sigue un proceso autorregresivo integrado de promedio móvil o ARIMA (*p,q,d*) donde *p* denota el número de términos autorregresivos, del número de veces que la serie debe ser diferenciada para hacerse estacionaria y *q* el número de términos de promedio móvil (Gujarati,1997).

Los procesos ARIMA son suficientes para explicar los procesos como tendencia, pero insuficiente para representar procesos de estacionalidad y por ello se hace necesaria la transformación de la serie.

Se han encontrado algunas extensiones del modelo ARIMA que contempla estacionalidad de la serie de tiempo como el Modelo Autorregresivo Integrado de promedio móvil estacional (SAMIRA), Modelo autorregresivo integrado de promedio móvil con variables exógenas (ARIMAX), Modelo autorregresivo integrado de promedio móvil estacional con variables exógenas (SARIMAX).

Este modelo estocástico ARIMA o metodología Box-Jenkins corresponden a una combinación lineal de valores históricos de una variable independiente y su ajuste es evaluado siempre que los residuales sean valores pequeños

4.2 Diseño de prueba

Tomando la base de datos original, se procede a dividir en dos bases, entrenamiento y prueba, y teniendo en cuenta que se está trabajando un modelo de series tiempo, esta partición se realiza de forma continua, garantizando que la base de entrenamiento tendrá la información de los primeros registros hasta el 27 de junio de 2021 y la de prueba estaría comprendida desde esta fecha hasta el 30 de diciembre de 2021, manteniendo una relación 70% - 30%.

Las métricas para evaluar los modelos construidos para la campaña serán RMSE (Error cuadrático medio) y MAPE (Mean Absolite Percentage Error) como indicadores del desempeño del pronóstico de demanda.

4.3 Construcción del Modelo

Para la generación del modelo, se utilizaron tres atributos, una función de cambio, un promedio móvil, y una característica polinomial. Específicamente, se utilizaron medias móviles con valores de 3 y 7, función de cambio con valores de 1 y 7 y una función polinomial de grado 3. Una vez creadas estas características, se define un horizonte de tiempo que corresponda al caso de negocio, en este caso se utilizan 7 días, ya que, las predicciones deben realizarse semanalmente, lo que se aplicara para los modelos de Regresión Linela, Random Forest Regressor y Gradient Boosting

4.3.1 Regresión Lineal

4.3.1.1 Parámetros y Definición del modelo

Para la regresión lineal tenemos los siguientes hiperparámetros:

- fit_intercept: para dejar o quitar la constante β_0. Si se elimina, la recta pasará obligatoriamente por el punto 0 del eje de abscisas.
- normalize: para normalizar los datos o no, normalmente la regresión lineal suele funcionar mejor con datos normalizados, para que todas las variables estén a la misma escala. Por lo que se asegura el modelo dejando este parámetro como positivo (true)

Se construye el modelo tomando como base la regresión lineal, la cual se realiza sobre la base de entrenamiento descrita en el numeral 4.2 y que, a partir de los coeficientes calculados por el modelo, se realiza el pronóstico de la base de prueba, para evaluar sobre esta la eficiencia del modelo.

4.3.2 Random Forest Regressor

4.3.2.1 Parámetros y definición del modelo

La clase RandomForestRegressor del módulo permite entrenar modelos random forest para problemas de regresión. Los parámetros e hiperparámetros empleados por defecto y que destacan:

- N_estimators; número de árboles incluidos en el modelo.
- Max_depth: profundidad máxima que pueden alcanzar los árboles.
- Max_features: número de predictores considerados a en cada división.
- N_jobs: número de cores empleados para el entrenamiento. En random forest los árboles se ajustan de forma independiente, por lo que la paralelización reduce notablemente el tiempo de entrenamiento. Con -1 se utilizan todos los cores disponibles.

Se desarrolla un cross-valideiton de 5 para determinar los mejores parámetros a implementar en el modelo, por lo que para efectos del modelo se utilizaron los siguientes parámetros:

Tabla 5 Parámetros del modelo de Random Forest Regressor

Parámetro	***Valor***
N_estimators	5
Max_depth	None
Max_features	Auto
N_jobs	-1

4.3.3 Construcción Modelo GradientBoosting

4.3.3.1 Parámetros

Para la construcción del modelo se utiliza un regresor de gradient boosting con los parámetros presentes en la Tabla *6*.

Tabla 6. Parámetros del modelo de gradient boosting utilizado.

Parámetro	***Valor***
Pérdida	Lad
Tasa de aprendizaje	0.1
No estimadores	500
No máximo de características	sqrt

Acto seguido, se ajustó el modelo de gradient boosting utilizando la serie temporal, empleando un modelo GARFF (modelo de auto regresión generalizado). Este modelo es compatible con scikit-learn y permite el ajuste a cualquier regresor para realizar predicciones a series temporales, en este caso concreto se utilizó gradient boosting.

4.3.3.2 Definición del modelo

Con el fin de realizar el modelado, se utilizó la librería de Python “Giotto-tda”, se modela la tendencia subyacente y la estacionalidad de la serie temporal, lo cual se logra con un modelo polinomial o exponencial, enseguida se elimina la tendencia de la serie de tiempo. En Ilustración *14* se puede

evidenciar la transformación que se hace sobre la cantidad de llamadas entrantes al canal telefónico de la campaña, a lo largo del tiempo, con la reducción de dimensión propuesta para la aplicación del modelo. Por otro lado, en la Ilustración *15* se expone el ajuste de la onda más cercana a los datos de entrenamiento transformados.

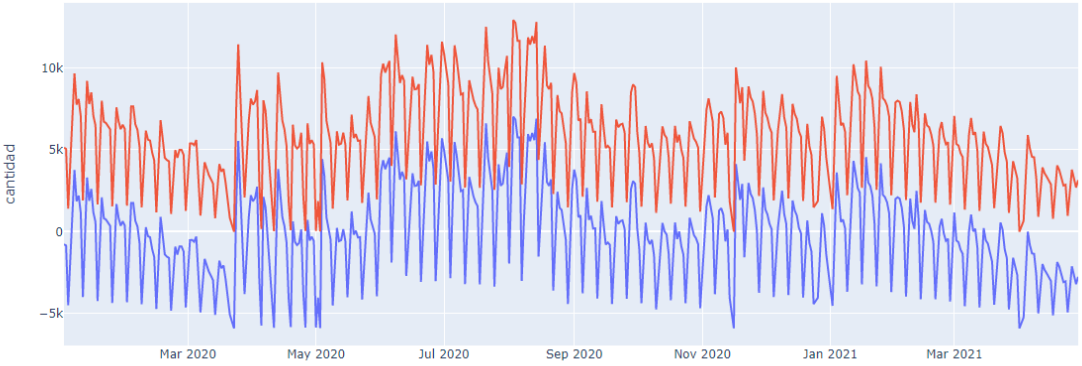


Ilustración 14. Cantidad de llamadas ingresadas del 1 de enero de 2020 al 30 de abril de 2021 en escala normal (color rojo) y trasladadas con eje igual 0 (color azul).

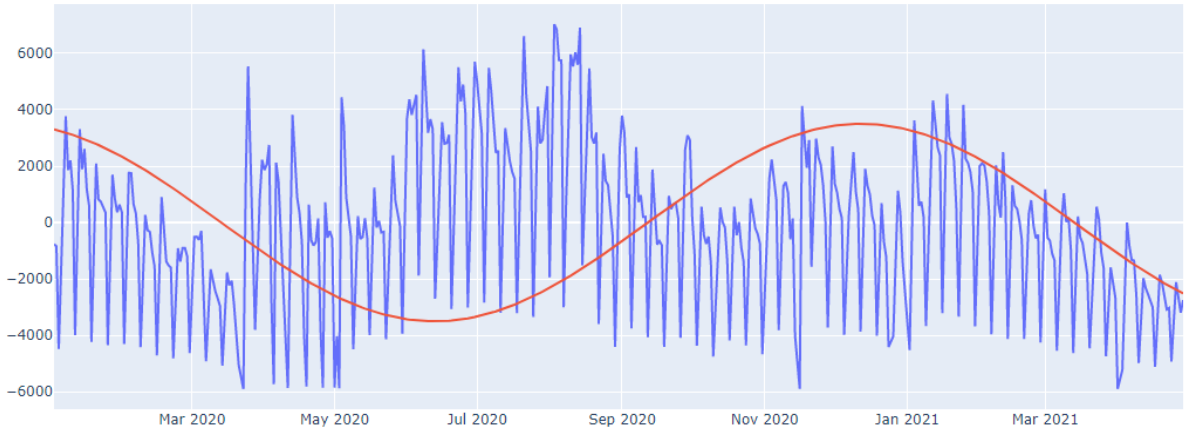


Ilustración 15. Cantidad de llamadas ingresadas del 1 de enero de 2020 al 30 de abril de 2021 trasladadas con eje igual 0, con el ajuste de la onda más adecuada.

4.3.4 Construcción modelo ARIMA

4.3.4.1 Parámetros

Se ejecutó un modelo ARIMA parámetros se encuentran en la Tabla *7* Parámetros del Modelo ARIMA, donde se captó un periodo estacional completo mediante la parte autorregresiva (182 días, que representan la mitad de un año), se realiza una sola diferenciación para eliminar la tendencia de la serie y se tuvieron en cuenta los errores de predicción de los 7 periodos anteriores (para este caso sería la semana inmediatamente anterior).

Tabla 7 Parámetros del Modelo ARIMA

Parámetro	***Valor***
p	182
d	1
q	7

4.3.4.2 Definición del modelo

Para el segundo modelo se utilizó un modelo ARIMA (*p, d, q*) donde se realizó un análisis de la autocorrelación y correlación parcial Ilustración *16*. Gráficos de autocorrelación simple y autocorrelación parcial, para la serie diaria con fin de determinar los hiperparametros de cada uno de los modelos, en términos de cuánta influencia tiene los periodos anteriores y ver si existe alguna tendencia dentro de la serie, y si hay estacionalidad.

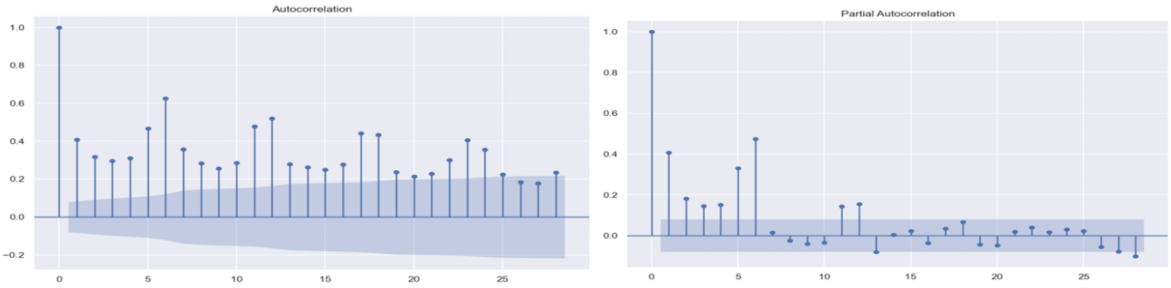


Ilustración 16. Gráficos de autocorrelación simple y autocorrelación parcial

En el correlograma de autocorrelación se puede evidenciar que existe estacionalidad cada 6 periodos y estos son significativos hasta los 24 periodos. Para el correlograma de autocorrelación parcial se logra identificar que los primero 6 desfases son significativos.

En este punto se tomó las series de tiempo que contiene la cantidad de llamadas diarias a lo largo de 2020 y 2021 para realizar la separación de los datos en entrenamiento y prueba, ver Ilustración *17*. Partición de la serie de tiempo en entrenamiento y prueba (entrenamiento - azul, prueba - rojo). Esta separación se dio en función de capturar un periodo estacional completo de forma que para el set de prueba se logre comparar con el comportamiento que se evidencia cada 6 meses a lo largo de los 2 años.

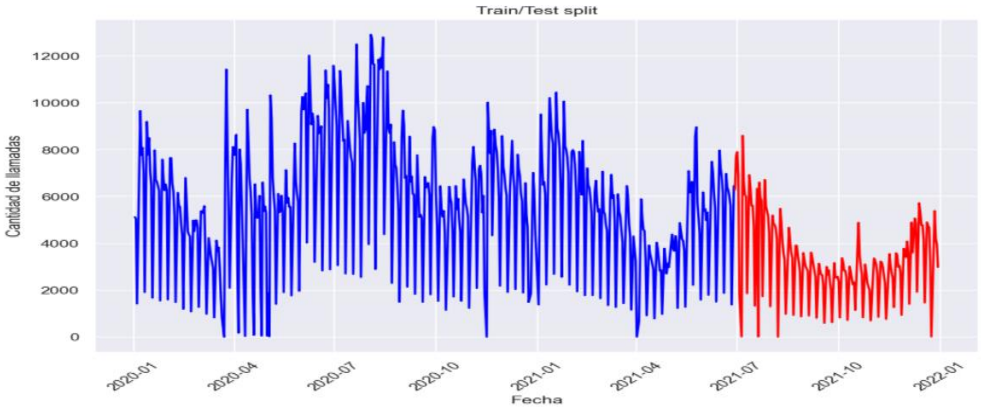


Ilustración 17. Partición de la serie de tiempo en entrenamiento y prueba

4.4 Evaluación del Modelo

4.4.1 Evaluación del modelo Regresión Lineal

Ejecutado el modelo de regresión lineal y comparándolo con los datos de prueba se puede ver que el modelo esta sobre estimado para los casos en que el volumen de llamadas es menor a la media de los datos reales. Y se tiene un 15% de error en el pronóstico (MAPE) como se evidencia en la tabla de las métricas utilizadas

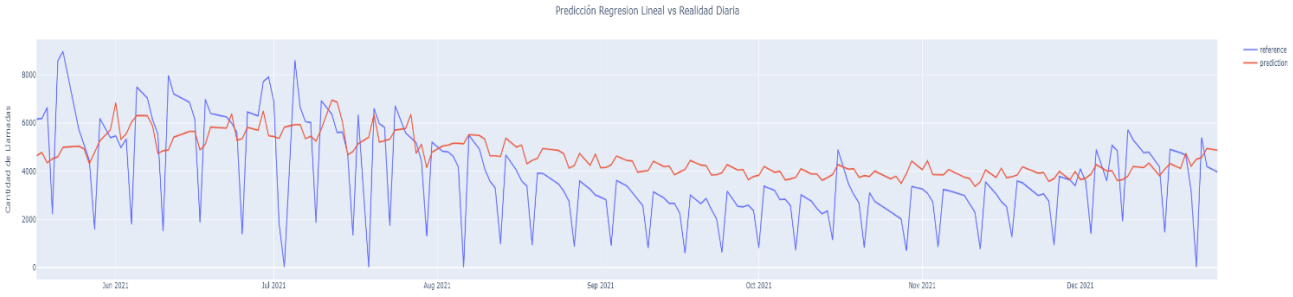


Ilustración 18. Comparación de los valores pronosticados y los valores reales modelo Regresión Lineal

A continuación, se muestra la tabla con los resultados obtenidos.

Tabla 8. Métricas modelo de regresión lineal

Métrica	Resultado
RMSE	1834
MAPE	15.38%

4.4.2 Evaluación del modelo Random Forest Regressor

De acuerdo con los hiperparámetros estimados de acuerdo a la validación cruzada y apalancado en las iteraciones realizadas, se tiene un modelo que esta mejor estimado sobre el valor real de los datos, como se evidencia en la imagen, sin embargo, a diferencia del modelo de regresión lineal se pude

evidenciar que este modelo podría estar sub estimado, debido a que en los picos más altos de los valores reales, no se está contemplando el total de llamadas que podrían estar llegando, a pesar de tener un mejor comportamiento de la métrica MAPE en relación al modelo de regresión lineal, ya que el RMSE es mayor que el modelo anterior.

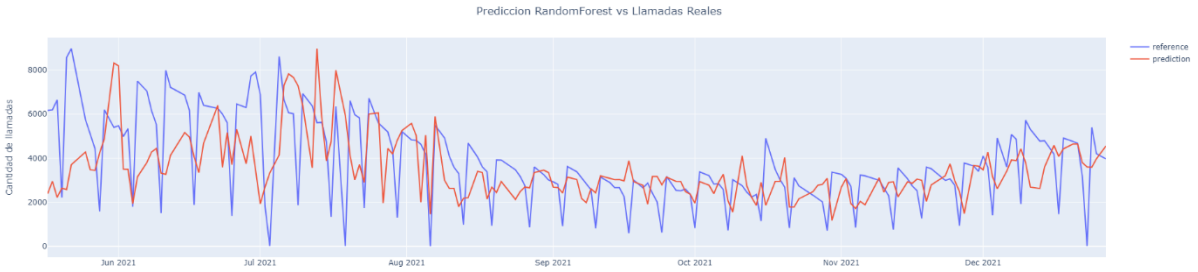


Ilustración 19. Comparación de los valores pronosticados y los valores reales modelo Random Forest Regressor

A continuación, se muestra la tabla con los resultados obtenidos.

Tabla 9. Métricas modelo de Random Forest Regressor

Métrica	Resultado
RMSE	1921
MAPE	8.93%

4.4.3 Evaluación modelo GradientBoosting

Finalmente, se desarrolló la predicción de la cantidad de llamadas recibidas desde junio de 2021 a final de año en el mismo año, con el modelo propuesto en la Tabla *6*. En la Ilustración *20***¡Error! No se encuentra el origen de la referencia.** se expone gráficamente la comparación entre los datos resultantes de la predicción y los datos reales de testeo, por otro lado, en la Tabla *10*. Métricas modelo de Gradient , se muestran las métricas correspondientes al modelo indicado.

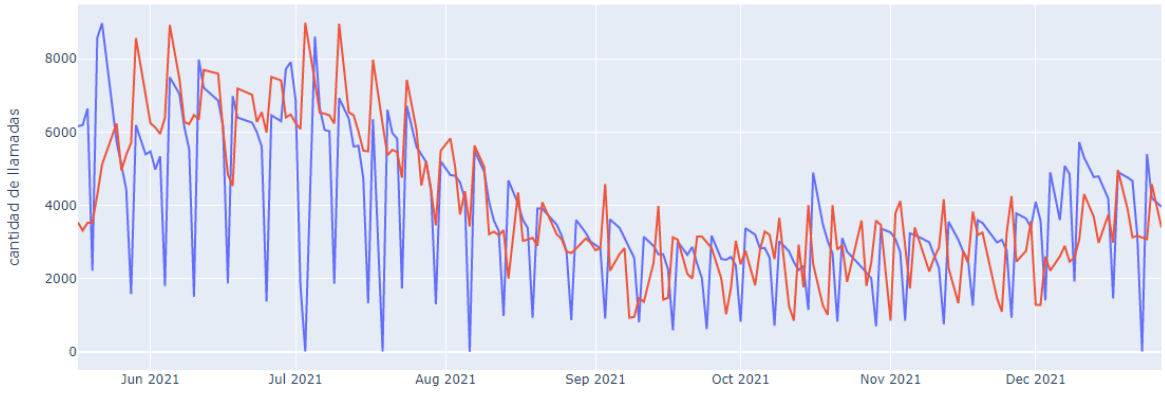


Ilustración 20. Cantidad de llamadas ingresadas del 1 de junio de 2021 al 31 de diciembre de 2021 reales (color azul) y con predicción del modelo de Gradient Boosting propuesto (color rojo).

Tabla 10. Métricas modelo de Gradient Boosting

Métrica	Resultado
RMSE	1879
MAPE	14.13%

4.4.4 Evaluación modelo ARIMA

Teniendo el modelo entrenado y la serie de tiempo divida en entrenamiento y prueba se realiza la predicción sobre los datos de prueba, se obtiene las métricas para evaluar el desempeño del modelo y para poder compararlo con los modelos anterioriores

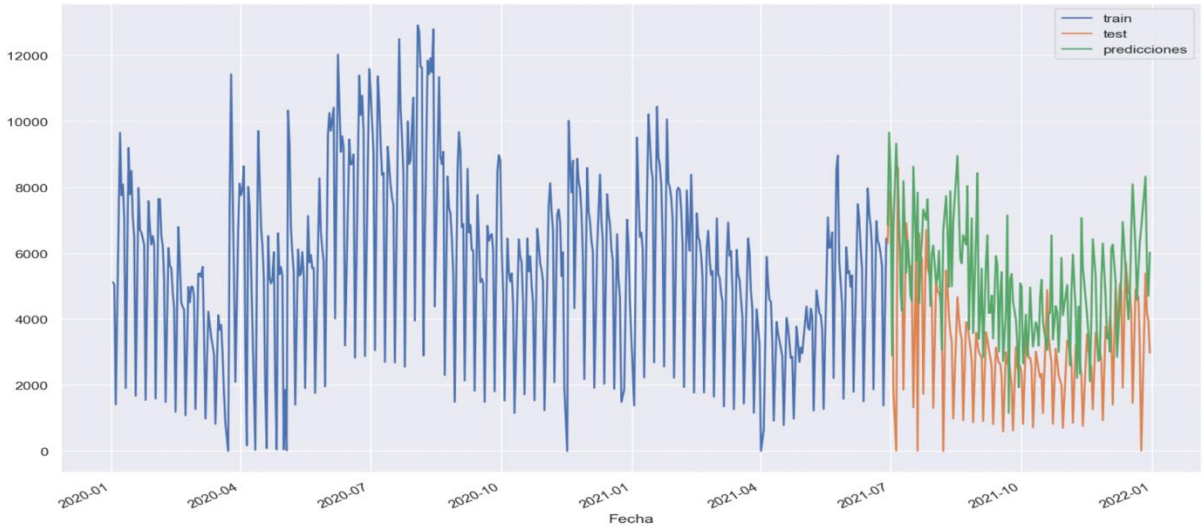


Ilustración 21. Gráfico del pronóstico obtenido frente a la serie de prueba.

Teniendo en cuenta que lo resultados se asemejan al comportamiento del set de prueba, se logra evidenciar que existe una sobreestimación en la cantidad de llamadas. Posiblemente por errores de periodos anteriores y porque en la estacionalidad de la serie para el entrenamiento del modelo se tiene en cuenta los datos correspondientes al primer semestre de 2020 que presenta gran volatilidad por el contexto de pandemia.

Tabla 11. Métricas resultantes del modelo 2.

Métrica	Resultado
RMSE	2837
MAPE	26.16%

5 Evaluación

5.1 Evaluación de Resultados

Ahora bien, teniendo en cuenta el contexto del negocio se realiza la comparación de las diferencias de la nómina y la facturación teóricas de la cantidad de llamadas reales en el periodo de prueba (de junio 2021 a diciembre 2021) versus la predicción en la misma fecha.

En la Ilustración *22* se expone la diferencia de la nómina teórica calculada a la cantidad de llamadas reales y la nómina teórica calculada a la cantidad de llamadas predichas con el modelo GradientBoosting, dado que, de acuerdo a las métricas, se considera como el de mejor desempeño. En donde los valores positivos nos expresan sub-asignación de recursos en una fecha determinada y valores negativos nos indican sobre asignación de los mismo. Se evidencia que el modelo seleccionado tiende más a la sobre asignación que la sobre asignación. Teniendo en cuenta que la nómina en una semana determinada se considera constante.

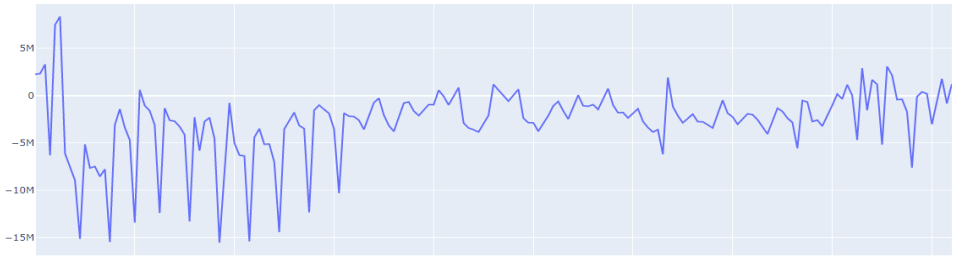


Ilustración 22. Diferencia de nómina teórica en base a la cantidad de llamadas recibidas y nómina teórica en base a la cantidad de llamadas predichas por el modelo 1.

Por otro lado, se realiza el mismo ejercicio con la facturación teórica, el cual está expuesto en la Ilustración *23*, en donde los valores positivos representan dinero que se contemplaba que ingresaría a la campaña, pero por sub-asignación de recursos no se facturó, sin embargo, se lograría facturar la totalidad de las llamadas predichas ingresadas. Por otra parte, los valores negativos corresponden a que en realidad entraron menos llamadas de las predichas, por lo tanto, se facturó menos de lo esperado, se trataría de una sobre asignación de recursos.

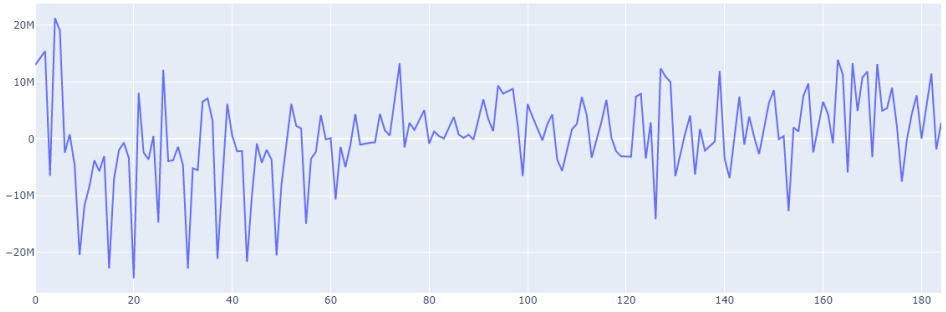


Ilustración 23. Diferencia de factura teórica en base a la cantidad de llamadas recibidas y factura teórica en base a la cantidad de llamadas predichas por el modelo 1.

Finalmente, cuando se realiza la comparación teórica global del indicador N/F este tiene un valor ideal y mínimo de 43.8%, partiendo de los resultados del modelo de Gradient Boosting, de acuerdo con esto tendríamos una mejora en la asignación de recursos, ya que, teóricamente el indicador estaría

aproximadamente 7 puntos por debajo de lo calculado con el funcionamiento del teórico de la asignación actual (51.3%). Ahora bien, hay que tener en cuenta que el indicador no solo depende de asignación y que factores externos como ausentismos, permisos remunerados, e incapacidades, entre otras, pueden afectar negativamente el indicador. Sin embargo, al tener mínimo ideal más bajo se aumentan la probabilidad de generar más valor en la empresa.

5.2 Pasos a Seguir

Conforme a lo revisado y experimentado en el desarrollo de este ejercicio, se proponen los siguientes puntos, para que la organización los tome en consideración y determine la importancia de cada uno con el fin de dar continuidad al estudio:

- Se sugiere ampliar el análisis, realizando un modelo tomando a detalle rangos de horas de forma que se tome la cantidad de llamadas por hora o cuartos de hora, de forma que se pueda simplificar el proceso de asignación de recursos y se corrija la variabilidad que pueda existir de acuerdo con la metodología hoy aplicada en la organización.
- Extender en el análisis teniendo en cuenta la duración de las llamadas y su influencia en el abandono, identificando si están relacionadas directamente con el volumen de llamadas como se concluye en el ejercicio realizado, o en su defecto se atribuye a otro factor interno dentro del proceso que por los datos no se puede identificar, como agentes reiterativos con llamadas abandonadas.
- Validar la funcionalidad de este modelo con otras campañas de otros clientes de forma que se pueda aumentar el impacto en el indicador de N/F.
- Como se evidenció, existe un gran número de datos atípicos en las bases suministradas, por lo que de igual forma se sugiere tener un mejor control en la calidad de los datos, asegurando que se pueda unificar la nomenclatura de las diferentes categorías o tipologías en marcadas en las variables.

6 Manejo Responsable de la Información

En Colombia rige la ley 1581 de 2012, la cual tiene por objeto desarrollar el derecho constitucional que tienen todas las personas a conocer, actualizar y rectificar la información que se haya recogido sobre ellas en bases de datos o archivos, y los demás derechos, libertades y garantías constitucionales a que se refiere el artículo 15 de la Constitución Política. Dentro de la ley se exponen unos principios de protección de datos que deberán ser aplicables a todas las bases de datos, por lo que este capítulo evaluará su aplicabilidad dentro del marco del desarrollo de esta propuesta:

- **Principio de legalidad en materia de Tratamiento de datos:** las bases usadas son propias de la empresa Millenium BPO, cumpliendo con la política interna y con lo establecido en la ley para este principio.
- **Principio de Finalidad:** Para esta propuesta se busca usar estas bases de datos para proponer e identificar una oportunidad de mejora en los procesos de atención de servicio al usuario y mejorar la rentabilidad del negocio.
- **Principio de Libertad:** En el momento que un usuario inicia una interacción se le solicita la autorización o no del tratamiento de datos, dado que de acuerdo con este principio los datos personales no podrán ser obtenidos o divulgados sin previa autorización.

- **Principio de veracidad o calidad:** las bases de datos son suministradas por Millenium BPO, de forma que, para la presente propuesta, se asume que las bases son veraces y los datos reflejan la realidad del negocio.
- **Principio de transparencia:** para el caso de esta propuesta, no se detalla información personal de los clientes de forma que se puedan ser relacionados directamente con el estudio, sin embargo, este principio se cumple a través de la política de la empresa.
- **Principio de acceso y circulación restringida:** en el marco de la propuesta este principio no se ve afectado de forma que las bases suministradas no contienen ninguna información personal y de acuerdo con este principio, el tratamiento se sujeta a los límites que se derivan de la naturaleza de los datos personales, de las disposiciones de la presente ley y la Constitución. Sin embargo, en cuanto al marco de la empresa el principio se cumple de forma que los datos personales adquiridos no se encuentran disponibles en Internet u otros medios de divulgación o comunicación masiva.
- **Principio de seguridad**: la empresa cuenta fuertes políticas de seguridad que involucran medidas técnicas, humanas y administrativas para evitar adulteración, pérdida, consulta, uso o acceso no autorizado o fraudulento a la información.
- **Principio de confidencialidad**: para efectos de este trabajo y poder acceder a la información recolectada y almacenada previamente por Millenium BPO, se estableció un acuerdo de confidencialidad, de forma que se garantice este principio.

En cuanto a la Política para el Tratamiento de Datos Personales que tiene Millenium BPO, se tienen unas consideraciones adicionales que cabe resaltar, conforme al tratamiento de las bases de datos que se suministraron:

- En las relaciones contractuales, la política incluye en los procesos de contratación cláusulas con el fin de autorizar de manera previa y general el tratamiento de datos personales relacionados con la ejecución del contrato, lo que incluye la autorización para recolectar, modificar o corregir datos personales del titular en momentos futuros. También incluirá la autorización de que algunos de los datos personales, en caso dado, puedan ser entregados a terceros con los cuales la organización tenga contratos de prestación de servicios, para la realización de tareas tercerizadas.
- Cuando la organización contrata a terceros para la realización de tareas complementarias, y el contratado requiera de acceso a datos personales, la organización les suministrará acceso a estos datos siempre y cuando exista una autorización previa y expresa del titular para esta transmisión. Se incluirá una cláusula que prohíba una entrega posterior a otros terceros, así como el uso comercial de los datos personales entregados.

7 Bibliografía

Asociación colombiana de BPO. (2021). *El sector BPO en Colombia*. Obtenido de https://www.bpro.org/que-es-el-sector-bpo

CCM. (29 de 05 de 2019). Obtenido de ¿Qué es un Contact Center y cómo impacta al crecimiento de tu negocio?: https://www.callcentermexico.com.mx/blog/que-es-un-contact-center-y-como-impacta-al-crecimiento-de-tu-negocio

Europa Press. (2019). Robots, IA y redes neuronales: así es la tecnología detrás de las empresas BPO. *El Confidencial*. Obtenido de https://www.elconfidencial.com/tecnologia/2019-12-26/business-process-outsourcing-bpo-konecta-inteligencia-artificial-empresas_2390583/

Giotto-tda. (2021). *Topology of time series*. Obtenido de https://giotto-ai.github.io/gtda-docs/latest/notebooks/topology_time_series.html

IBM. (2022). *Introducción al CRISP DM*. Obtenido de https://www.ibm.com/docs/es/spss-modeler/SaaS?topic=situation-risks-contingencies

Mann, L., & Graham, M. (2016). The domestic turn: Business Process Outsourcing and the Growing Automatitation of Kenyan Organisations. *The Journal of Development Studies*, Vol. 52, No. 4, 530–548. Obtenido de http://dx.doi.org/10.1080/00220388.2015.1126251

Masui, T. (01 de 2021). *All You Need to Know about Gradient Boosting Algorithm − Part 1. Regression*. Obtenido de Toward Data Science: https://towardsdatascience.com/all-you-need-to-know-about-gradient-boosting-algorithm-part-1-regression-2520a34a502

Millenium BPO. (2022). Obtenido de Contact Center | BPO - ML+AI | RPA/RDA: https://www.linkedin.com/company/milleniumbpo/about/

Palma, F., Kalin, C., & Tunstall, L. (20 de 02 de 2020). *Getting started with giotto-time*. Obtenido de Towards Data Science: https://towardsdatascience.com/getting-started-with-giotto-time-d9b2088d60ca

Rodrigo, J. A. (10 de 2020). *Random Forest con Python*. Obtenido de https://www.cienciadedatos.net/documentos/py08_random_forest_python.html

Sectorial. (21 de 10 de 2021). *Se Espera 5% en Aumento de Ingresos Para el Sector BPO en Colombia en 2021*. Obtenido de sectorial.co/informativa-contact-center-y-bpo/item/460606-se-espera-5-en-aumento-de-ingresos-para-el-sector-bpo-en-colombia-a-2021

Hanke, J. E., & Reitsch, A. G. (1996). Pronosticos en los Negocios. Production Supervision.

Damodar N Gujarati (1997). Econometría. Dawn C Porter.

8 Anexos

8.1 Terminología

- Usuario: El usuario es el cliente final que nos está llamando o estamos llamando.
- Cliente: Empresa contratante de servicios de la campaña que realiza la venta o da el servicio al cliente.
- Canal: El medio por el cual el usuario nos contacta o lo contactamos. Ejemplos: chat en página web, telefonía, WhatsApp, correo electrónico.
- Call Center: centro de atención telefónica personalizada al usuario.
- Recurso Sobredimensionado: Asignación de personas a una campaña por arriba del ideal para cumplir con las funciones.
- Formato “xlsx”: Formato de archivo basado en XML y habilitado para macros de Excel 2007 a 2013

- Python: lenguaje de programación interpretado cuya filosofía hace hincapié en la legibilidad de su código.
- IDE: Un entorno de desarrollo integrado o entorno de desarrollo interactivo, en inglés Integrated Development Environment, es una aplicación informática que proporciona servicios integrales para facilitarle al desarrollador o programador el desarrollo de software
- Hardware: El hardware, equipo o soporte físico en informática se refiere a las partes físicas, tangibles, de un sistema informático, sus componentes eléctricos, electrónicos, electromecánicos
- Software: Se conoce como software, logicial o soporte lógico al sistema formal de un sistema informático, que comprende el conjunto de los componentes lógicos necesarios que hace posible la realización de tareas específicas, en contraposición a los componentes físicos que son llamados hardware.
- TMO: Tiempo de duración de una interacción.
- BPO: Sector de tercerización de procesos de negocio
- Front Office: el concepto abarca aquellas estructuras empresariales encargadas de interactuar directamente con el cliente.
- Out Bound: es el conjunto de acciones de marketing que tienen el objetivo de captar consumidores mediante métodos directos y unidireccionales.
- PQR: siglas del conjunto de palabras Preguntas, Quejas y Reclamos.
- Chat box: también conocido como un asistente digital, es un programa o herramienta la cual sirve como un simulador de conversaciones humanas.
- Estacionalidad: Es la repetición sistemática de determinadas variaciones no necesariamente regulares en alguna variable cada cierto periodo de tiempo normalmente igual o menor a un año que afectan las decisiones de los agentes y técnicas de producción disponibles en cada ámbito.
- Autorización: Consentimiento previo, expreso e informado del Titular para llevar a cabo el Tratamiento de datos personales
- Base de Datos: Conjunto organizado de datos personales que sea objeto de Tratamiento
- Dato personal: Cualquier información vinculada o que pueda asociarse a una o varias personas naturales determinadas o determinables
- Encargado del Tratamiento: Persona natural o jurídica, pública o privada, que por sí misma o en asocio con otros, realice el Tratamiento de datos personales por cuenta del responsable del Tratamiento.
- Responsable del Tratamiento: Persona natural o jurídica, pública o privada, que por sí misma o en asocio con otros, decida sobre la base de datos y/o el Tratamiento de los datos
- Titular: Persona natural cuyos datos personales sean objeto de Tratamiento
- Tratamiento: Cualquier operación o conjunto de operaciones sobre datos personales, tales como la recolección, almacenamiento, uso, circulación o supresión.

8.2 Descripción de los datos

Tabla 12. Descripción de los datos.

N°	Nombre del Campo	Tipo de Dato	Descripción del Dato
1	Fecha	Datetime	Representa la fecha en que se realizó la llamada
2	Hora	Datetime	Hora en que entre la llamada
3	Canal	String	Canal por el cual el usuario hace contacto con el asesor
4	ASA(Seg)	Float	Tiempo promedio en que se demora un agente en contestar una llamada
5	AHT (Seg)	Float	Tiempo promedio de duración de una llamada desde que entra hasta que finaliza
6	HOLD (Seg)	Float	Tiempo de espera dentro de una llamada
7	Ubicación	String	Ciudad/Departamento desde el cual se realiza la llamada
8	Tipo de Cliente	String	Clasificación de cliente (Beneficiario, Ciudadano tipo cliente o Instituto de Educación Superior)
9	Menor de Edad	String	Establece si la persona que realiza la llamada es mayor o menor de edad
10	Tipo de caso	String	Hace referencia al motivo general de la llamada
11	Proceso Inicial	String	Tipificación sobre la variable “Tipo de caso”
12	Tipificación Inicial	String	Tipificación sobre la variable “Proceso inicial”
13	Subtificación Inicial	String	Tipificación sobre la variable “Tipificación inicial”
14	Tipificación Canal	String	Tipificación sobre la variable “Tipificación Canal”
15	Subtipificación Canal	String	Tipificación sobre la variable “Canal”
16	Que tan satisfecho se encuentra con el servicio	Float	Hace parte de la encuesta de satisfacción al finalizar la llamada – se representa en una escala numérica
17	Motivo de la calificación Buena	String	Hace parte de la encuesta de satisfacción al finalizar la llamada
18	Motivo de la calificación Mala	String	Hace parte de la encuesta de satisfacción al finalizar la llamada
19	Recomendaría la Entidad Financiera	String	Hace parte de la encuesta de satisfacción al finalizar la llamada
20	Estado de la llamada	String	Determina si la llamada fue efectiva o abandonada
21	Cola de la llamada	String	Marquilla para determinar la entrada de la llamada (utilizada en procesos internos)

8.3 Matriz de Correlación

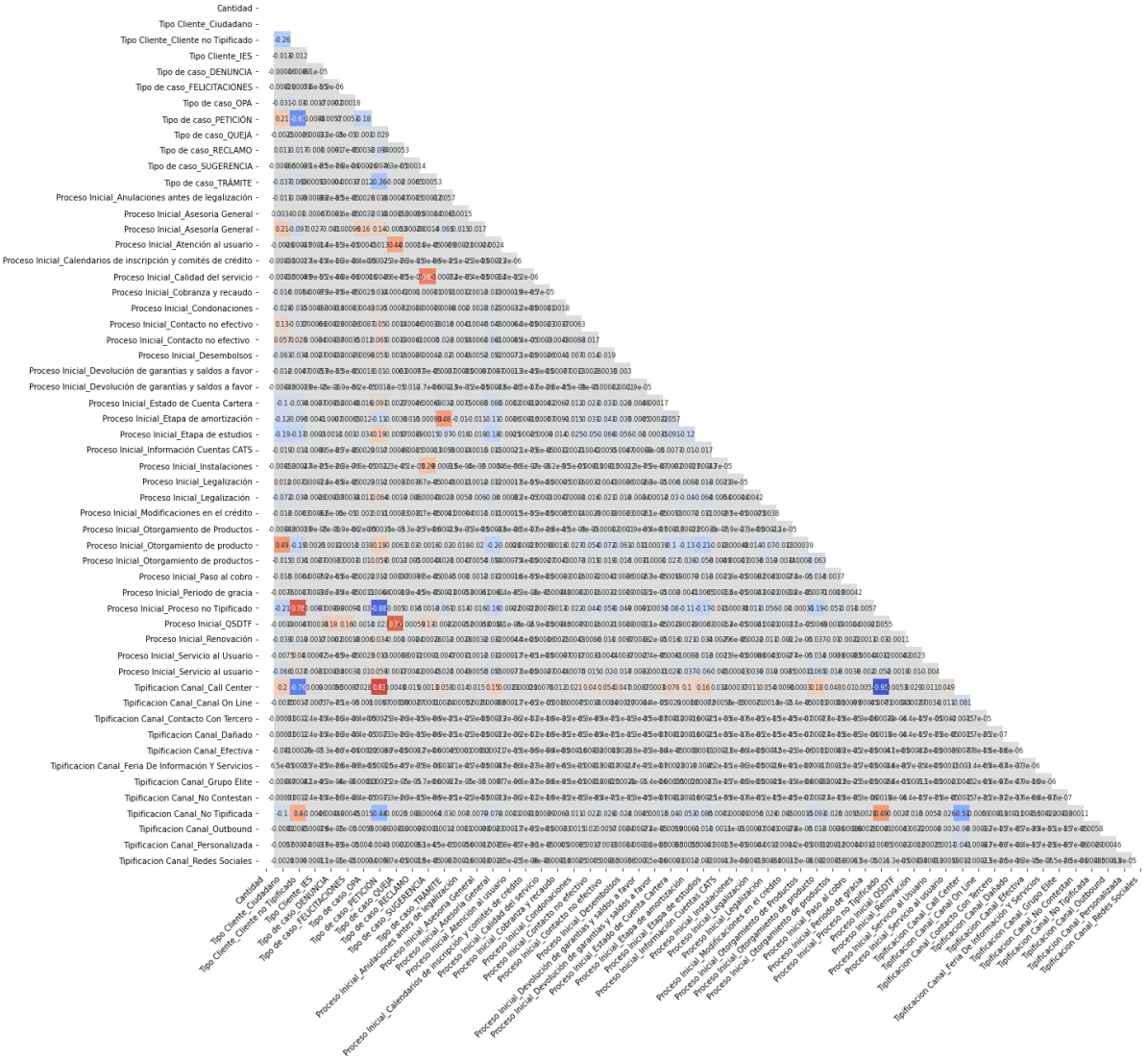
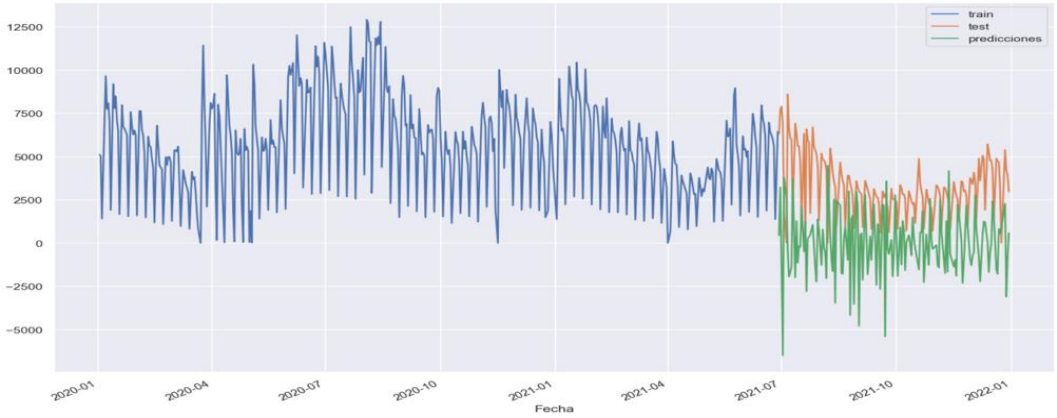


Ilustración 24. Matriz de Correlación

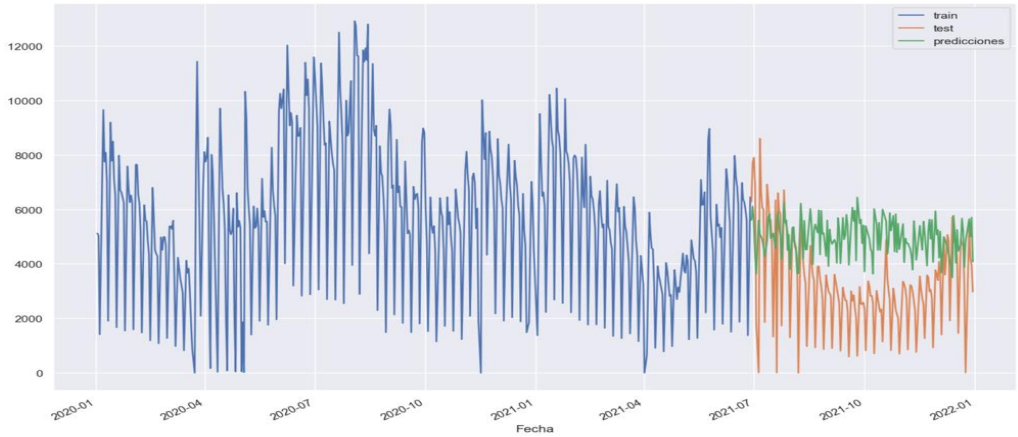
8.4 Modelos ARIMA evaluados

El siguiente modelo es un ARIMA sobre la serie diferenciada (lag=1) con fin de eliminar el factor tendencial.



Parámetro	***Valor***
p	182
d	1
q	7

El siguiente modelo es un SARIMA donde se utiliza la diferenciación estacional con fin de capturar los picos y valles de la serie (periodicidad=26).



Parámetro	***Valor***
p	12
d	2
q	1
P	1
D	1
Q	6