_____Laboratorio de Diseño y Optimización de Operaciones

Modelos tradicionales vs. Modelos de Aprendizaje de Máquina para predicción de demanda de productos de telefonía celular de la empresa ZTE

Entrega Final

"Documentación"

Integrantes:

Carlos Raúl Muciño Angeles	A01104864
César Iván Trejo Cerón	A01368390
Nancy Escamilla Sánchez	A01366872

Docente:

Ana Luisa Masetto.

2 de Junio 2021

Introducción

Hoy en día el análisis de datos es muy valioso para toda empresa ya que esto permite tomar decisiones más acertadas, permite descubrir oportunidades, buscar opciones de mejora y dirigir la empresa hacia el éxito. Es por ello que en este proyecto hemos decidido utilizar ciencia de datos para poder predecir la demanda de la empresa de celulares ZTE con ayuda de herramientas de programación como RStudio y Python. Estas herramientas son demasiado útiles ya que permiten la manipulación de una gran cantidad de datos para extraer información y conocimiento. Al utilizar estas herramientas será posible obtener predicciones, las cuales son muy importantes para las empresas, porque de esta manera mantienen una buena relación con el cliente potencial, les ayuda a obtener ventaja competitiva a través de la adecuación de estrategias, obtienen conocimientos sobre sus productos, sus procesos y además les ayuda a controlar costos innecesarios e identificar riesgos y oportunidades.

Cabe destacar también que aplicar ciencia de datos permite a las empresas ubicarse en una posición única donde pueden cuantificar y determinar cuál o cuáles son las estrategias que están funcionando, las que no y el por qué.

Pero eso sí, hay que tomar en cuenta que se debe ejecutar un buen análisis de datos para ofrecer la mejor experiencia al cliente, de lo contrario o en caso de que la empresa decida dejar de lado los datos que se le indican, podría perder clientes y perjudicar al negocio. Durante el desarrollo de este trabajo podremos ver todos los pasos que se necesitan para poder realizar un análisis de datos adecuadamente y su aplicación en una empresa real la cual corresponde a ZTE.

Descripción de la situación actual

La empresa china ZTE Corporation (ZTE) se fundó en 1986 y tiene su sede en Shenzhen, provincia de Guangdong. Esta se dedica al diseño, desarrollo, producción, distribución e instalación de equipos de telecomunicaciones, sistemas de comunicaciones y soluciones de información para operadores, empresas, gobiernos y consumidores. Su cartera incluye redes inalámbricas, cableadas y centrales; sistemas, productos y servicios de software de TI y telecomunicaciones y dispositivos inteligentes tales como teléfonos, terminales móviles de datos y terminales familiares.

También ofrece servicios de diseño y consulta, así como soluciones de informatización integradas para el gobierno y las empresas a través de la aplicación de redes de comunicaciones, Internet de cosas, Big Data y tecnologías de computación en la nube. Los productos, soluciones y servicios de ZTE se utilizan en más de 160 países. En América Latina, la empresa cuenta con oficinas de ventas en Argentina, Brasil (Brasilia y São Paulo), Bolivia, Costa Rica, Chile, Colombia, Ecuador, Guatemala, Haití, Jamaica, México, Nicaragua, Panamá, Perú y Paraguay, El Salvador, Uruguay y Venezuela. En 2016, ZTE firmó un acuerdo con Telefónica para construir la primera red comercial VoLTE y FMC en América Latina.

ZTE registró un gran crecimiento gracias a la expansión que ha tenido en mercados emergentes. Podríamos decir que sus principales rivales son las marcas Huawei, Bq y Wiko, ya que todos compiten por posicionarse en el número uno. ZTE se conoce por ser una marca de teléfonos móviles ZTE baratos, ya que buscan ser competitivos en precio y ellos mismos han comunicado que quieren diferenciarse por vender teléfonos móviles al precio que realmente se merecen.

Además, estos últimos años han conseguido estar en boca de muchos gracias a sus nuevas estrategias de marketing. Ahora, estos móviles ZTE Android podemos encontrarlos en muchos más puntos de venta que antes.

Un tema importante es la ciencia de datos, ya que sin estos no podríamos conocer nada acerca de una empresa, por ejemplo: sus demandas, pronósticos, costos, ventas, entre otras cosas. Por ello es importante realizar proyectos con herramientas a las cuales hoy en día tenemos acceso como lo es RStudio ya que estas herramientas nos ayudan a mejorar la tomar decisiones con fundamentos que se llevan a cabo día con día en una empresa, tomando en cuenta que estas pueden llevar a un aumento de rentabilidad y lograr una mejora en el desempeño operacional.

Entender y describir la problemática en términos del negocio

La gran problemática que enfrenta ZTE es que tiene grandes empresas competidoras a su alrededor, por lo que ellos deben de enfocarse cada vez más en innovar nuevos productos que se acerquen a tener las características de sus competidores, satisfacer la demanda en tiempo y forma, pero considerando varios factores para poder cumplir con las especificaciones de los clientes, ofreciéndoles un producto de calidad y a un bajo costo, que satisfasga las necesidades del cliente y ofrezca ciertas características de teléfonos de la competencia a un costo más bajo y que sea funcional.

En este proyecto únicamente nos enfocaremos en la predicción de la demanda de la empresa ZTE, la cual resulta ser importante para la compañía debido a que está ayudara comprender a gran escala el negocio y adecuar ciertas decisiones para ser una empresa eficiente. Para lograr ello se utilizará la herramienta RStudio que nos permitirá realizar el procesamiento de una base de datos la cual presenta información importante de la compañía ZTE, esto con el objetivo de comprender la predicción de las demandas de ZTE.

Cabe destacar que procesar datos es un tesoro para las empresas porque con ello se puede gestionar eficientemente la variabilidad en el volumen de entrada de ventas,saber que es lo que está pasando dentro de la empresa para tomar decisiones que permitan reducir costos y generar mayores utilidades, además nos ayuda a comprender el mercado eficientemente y tomar decisiones importantes en el área de producción en este caso.

Entender y describir la problemática en términos de la Ciencia de Datos

En este punto tenemos ya la situación definida con los datos recolectados previamente, por lo que iremos desglosando cada situación que se presenta analizando los campos que se utilizaron para obtener la información de las ventas y que nos servirán para poder contestar a la pregunta planteada en nuestra problemática.

Comenzaremos por mencionar que tenemos 24 089 filas de datos que se llenaron con 14 tipos de datos, es decir, que de nuestro volumen total tenemos más de 337 mil datos. Los tipos de datos los podemos encontrar en la *Tabla 1. Datos.*

Tendremos que analizar más a fondo para identificar si se trata de errores o los datos son de calidad para trabajar con ellos durante este análisis. De cualquier manera, debemos realizar una limpieza de los datos comenzando por lo que alcanzamos a identificar ya (lista en un apartado posterior), y proceder a corroborar que los demás datos ya se encuentren listos para su análisis. Todo esto mientras esperamos que los datos sean suficientes, al parecer sí lo son para cubrir con las necesidades de nuestra problemática planteada y los objetivos que se listan en el apartado siguiente.

También podemos catalogar los datos como Estructurados, esto quiere decir que podríamos almacenarlos en una tabla, mediante Excel si los soporta o cualquier otra herramienta que permita el despliegue y visualización de ellos, para esto utilizaremos RStudio, que también nos abrirá la puerta a diversas herramientas para el tratamiento de los datos y su análisis.

Por último, las tareas a realizar son Agrupamiento para identificar patrones en las ventas que se realizaron, y posteriormente una Regresión que será nuestro análisis matemático para poder pronosticar las ventas del siguiente mes, con esto también se busca responder a la problemática de una manera congruente y certera.

La pregunta en cuestión es: ¿Cuántas unidades (y) de cada producto de ZTE, se venderán en todos los puntos de venta, al siguiente mes de registro?

Para poder dar respuesta a esta pregunta tenemos diversas variables que consideraremos como factores(x) para poder predecir una variable respuesta(y) las cuales se especifican a continuación:

Variables factores(x):

-Punto de venta, fecha,mes, año,número de ventas, sku,marca, gamma,costo promedio,zona, estado, ciudad,latitud y longitud.

Variable respuesta(y):

-Predicción de demanda de ZTE

Objetivos

- Reconocer la situación de ZTE, realizando un análisis detallado de la situación actual, problemáticas que enfrenta el negocio y comprensión de la implementación de ciencia de datos.
- Describir los datos proporcionados, y explorarlos con el fin de identificar errores de calidad en los datos proporcionados.
- Realizar una limpieza de datos correspondiente a todos los errores encontrados.
- Realizar un análisis de datos para poder comprender la situación de la empresa con datos cuantitativos
- Construir modelos de predicción y aprendizaje máquina para poder comparar la eficiencia de estos.
- Realizar gráficas para evaluar el desempeño de los modelos obtenidos
- Exponer los resultados obtenidos en el desarrollo del proyecto

Estructuración del proyecto y plan preliminar

En la *Imagen 1.1 Estructura de Proyecto ZTE* se muestra la estructura del proyecto, así como las 7 etapas que lo conforman y sus contenidos, duración de cada una de estas, nombres de responsables de cada actividad, al igual que el inicio y terminado de cada una.

Posteriormente, la *Imagen 1.2 Gantt de Proyecto ZTE* se puede observar el diagrama gantt con cada una de las actividades anteriormente mencionadas que se llevarán a cabo en el proyecto y los respectivos responsables.

Describir los datos crudos

Al analizar el archivo de datos con extensión ".csv" descubrimos que cuenta con 24089 entradas divididas en 14 variables, donde se detallan las ventas que se tuvieron durante los años 2019 y 2020 de equipos en diferentes lugares del país. Cada variable se detalla en el siguiente diccionario:

punto_de_venta: Especifica la tienda en donde se vendió el dispositivo, detallado con diferentes acrónimos o simplemente el nombre de la plaza o centro en donde la tienda se encuentra.
fecha: Variable que detalla la fecha en la que se realizó la venta con el formato de día/mes/año.
mes: Variable que detalla el mes en el que se realizó la venta.
anio: Variable que detalla el año en el que se realizó la venta.
num_ventas: Variable que detalla la cantidad de ventas registradas.
sku: Variable que detalla el número que identifica el dispositivo vendido.
marca: Variable que detalla la marca del dispositivo vendido.
gamma: Variable que detalla la categoría a la que pertenece el dispositivo vendido.
costo_promedio: Variable que detalla el precio en el que se vendió el dispositivo.
zona: Variable que detalla la parte del país en donde se vendió el dispositivo.
estado: Variable que detalla el territorio del país donde se vendió el dispositivo.
ciudad: Variable que detalla la zona urbana en donde se vendió el dispositivo.
latitud: Variable que detalla la distancia a la que se encuentra el punto de venta del ecuador.
longitud: Variable que detalla la distancia angular sobre el ecuador donde se encuentra el punto de venta.

Describir problemas de calidad

Con la información en la *Tabla 1. Datos de variables* podemos definir que todos los errores presentan un “peligro” de calidad en la información pues no permiten que se haga un análisis adecuado de los datos. Por ejemplo, puede haber ventas duplicadas pensando que se trataban de lugares, productos o marcas diferentes cuando se trata de uno solo.

Por lo tanto corremos el riesgo de obtener resultados que no proveen una imagen adecuada de la situación y consecuentemente nos darían resultados ineficientes e incluso inútiles al momento de decidir las acciones a tomar para resolver el problema planteado. Se podrían omitir registros pero esta no es la solución adecuada, ya que al omitir algunos datos nos perderemos de informaciòn valiosa para poder hacer un análisis con la exactitud adecuada. Lo mejor es analizar el dato para poder corregirlo, buscar cual fue la posible causa del error y si se trata de un dato atípico solucionarlo obteniendo su media, moda, mediana, etc de acuerdo al caso que se presente; corregirlo y poder utilizar todos los datos para realizar el procesamiento de estos adecuadamente.

Limpieza de datos
Para realizar las correcciones en cuanto a limpieza de nuestros datos utilizamos la herramienta RStudio la cual nos permitió hacer algunos cambios en datos incorrectos ya previamente encontrados, mencionaremos algunas funciones que se utilizaron para poder hacer dichos cambios:

- head: Nos permite mostrar conjuntos de datos con los registros que tenemos
- str: Para saber las clases que tenemos y el tipo de variable en las clases
- as.character: Para cambiar ciertos caracteres que estaban incorrectos en los datos
- as.factor: Nos sirvió para convertir una variables a factor
- select : Ayuda a poner filtro a columnas y poder solo observar los datos de una columna que queremos visualizar
- filter: Los filtros nos ayudaron a encontrar ciertos datos
- arrange: Nos sirvió para mostrar los datos en orden de más pequeños a más grandes
- summarise: Funciona para hacer un resumen de los datos
- unique: Para saber los datos o valores únicos que tengo en cierta columna sin que se repitan
- tolower: Para convertir ciertas palabras en minúsculas
- ster_replace: Nos ayudó a cambiar ciertas palabras mal escritas por la palabra correcta, es decir hacer reemplazos de palabras o errores ortograficos
- sepal_length: Nos sirvió para encontrar los valores exactos dentro de los conjuntos de datos
- write.csv: Nos ayudó a guardar los datos limpios

Análisis exploratorio

La interpretación de la ***Imagen 1.1*** es impactante, ya que nos habla de las ventas que se realizaban y que a lo largo de los meses de los que tenemos información estas fueron bajando, ahora podríamos decir que son mínimas a comparación del principio.

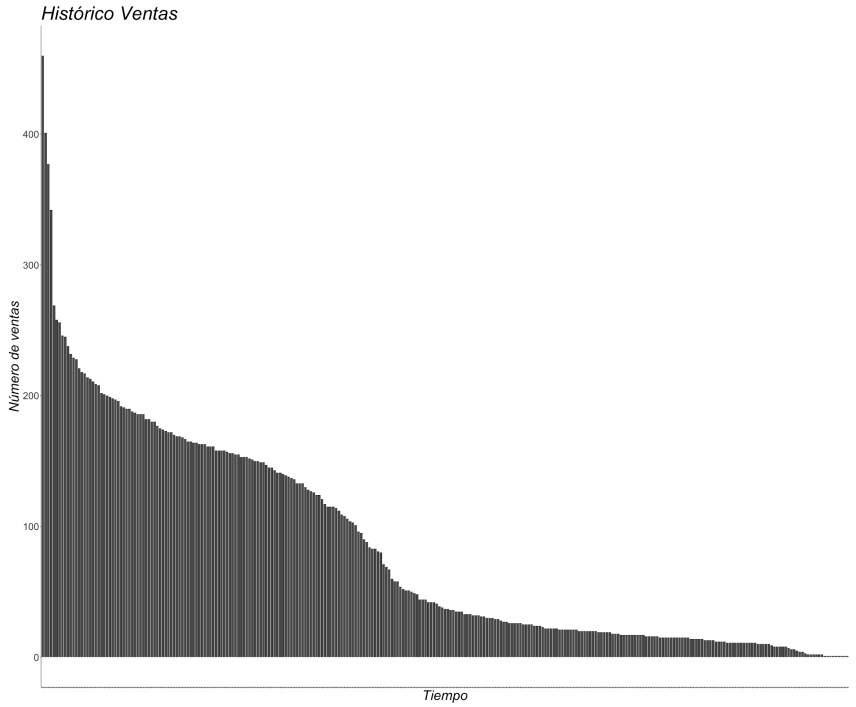

Imagen 1.1 Gráfica de barras de ventas a través del tiempo

En la ***Imagen 1.2*** se desglosan las ventas por estado y es muy visible cuáles son los estados en los que ZTE tiene mayor influencia para la venta de sus teléfonos. Teniendo entre los primeros 5 al Estado de México, Ciudad de México, Veracruz, Puebla y Nuevo León.

También se pueden observar los estados en los cuales se han obtenido las ventas más bajas, teniendo entre los últimos 5 lugares a Zacatecas, Baja California Sur, Nayarit, Oaxaca y Durango. Con esto en mente, nos hace pensar en aplicar nuevas estrategias de ventas para que estos lugares puedan aumentar las ventas los próximos meses. Se deberán revisar los factores que han impedido que se realicen ventas en comparación a los estados con mayores ventas y aplicar medidas correctivas adecuadas para lograr tener éxito en todos los estados.

La gráfica que se muestra en la ***Imagen 1.3*** expone el producto más vendido en SKU. En primer lugar se coloca el producto “N.ZV8MNG” que cuenta con al mayor número de ventas, casi el doble que los 4 productos siguientes individualmente, que serían “N.ZTA521N”, “N.ZV8MPN”, “N.ZTA521B” y “N.ZTEA6PT”.

El número de ventas del sku “N.ZV8MNG” resulta ser de 5056, en cuanto a los siguientes SKU sus números de ventas son de 2737, 2417, 2366 y 2359 respectivamente. Esta información resulta ser útil para poder conocer cuál es el producto más vendido y que características posee, ya que es el más demandado por los clientes, esto ayuda a poder aumentar la satisfacción de los clientes creando o modificando productos para aumentar las ventas totales.

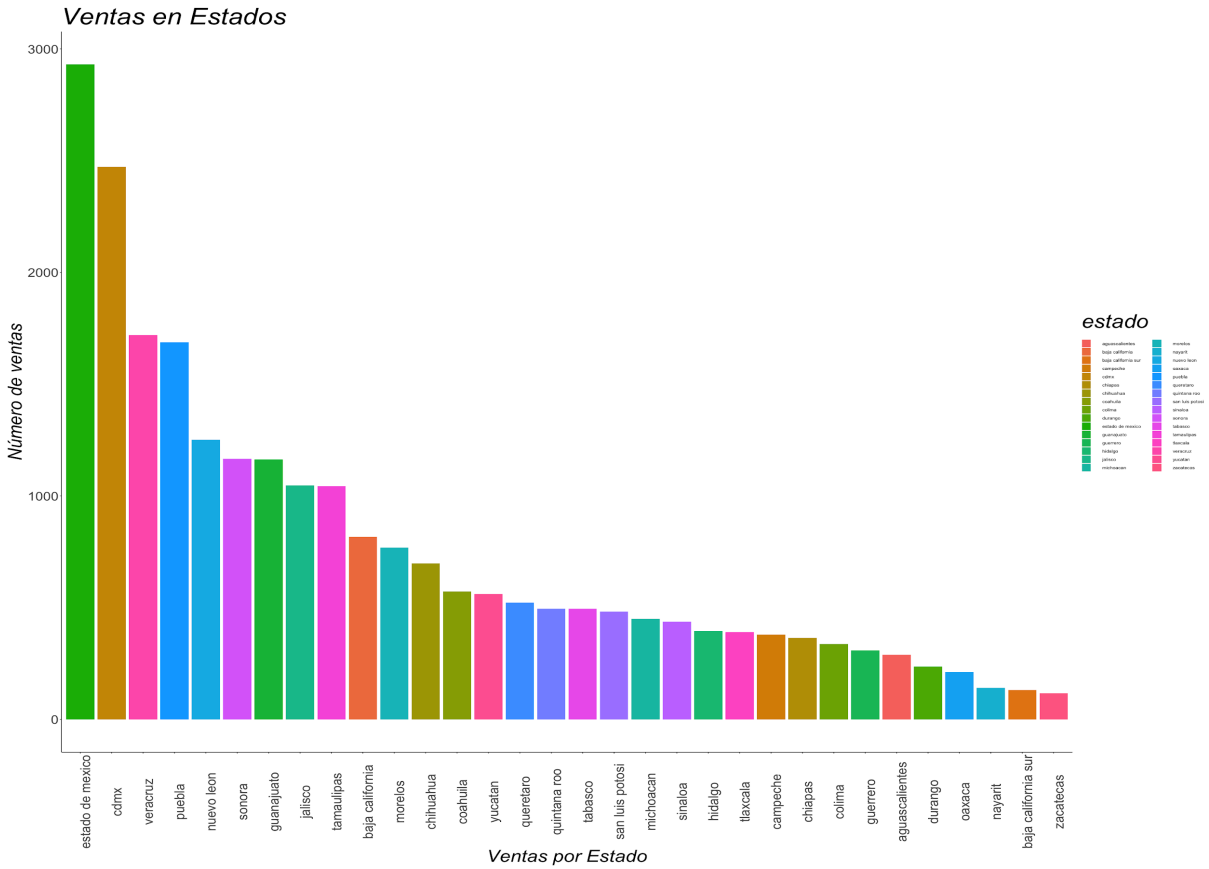

Imagen 1.2 Gráfica de barras de ventas por Estados

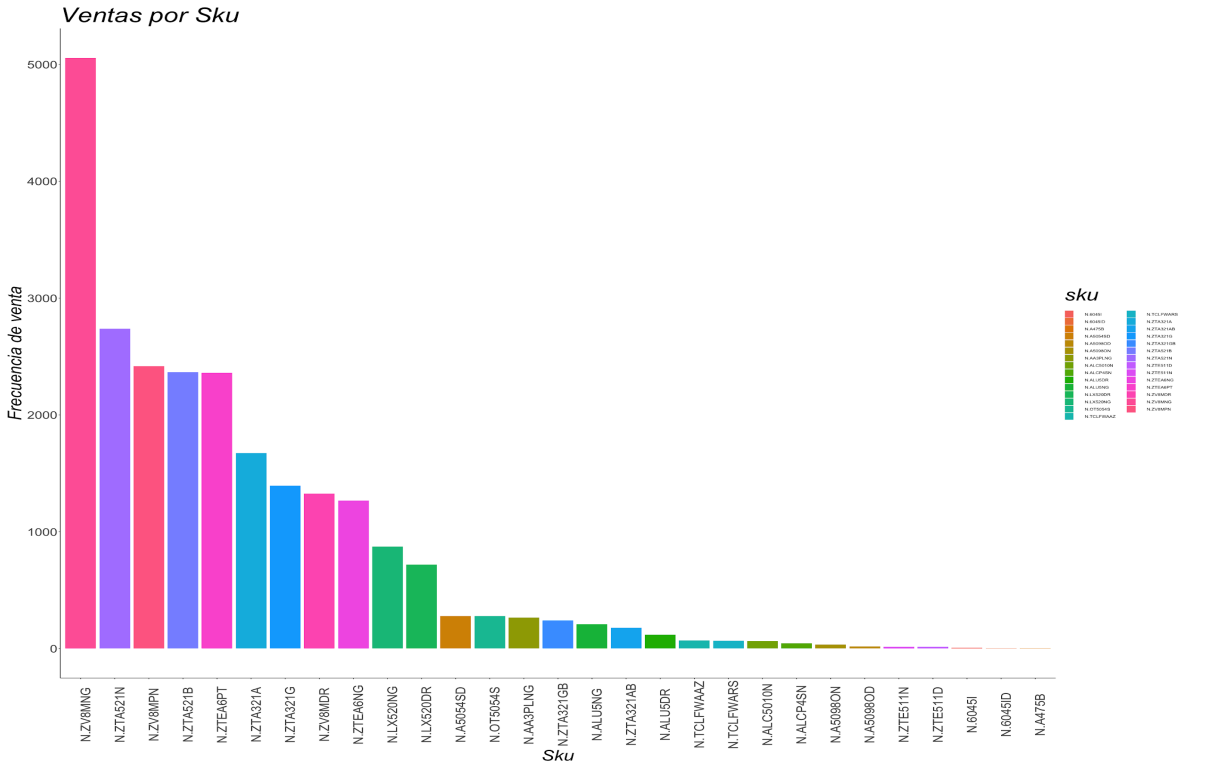

Imagen 1.3 Gráfica de barras del producto con el SKU más vendido

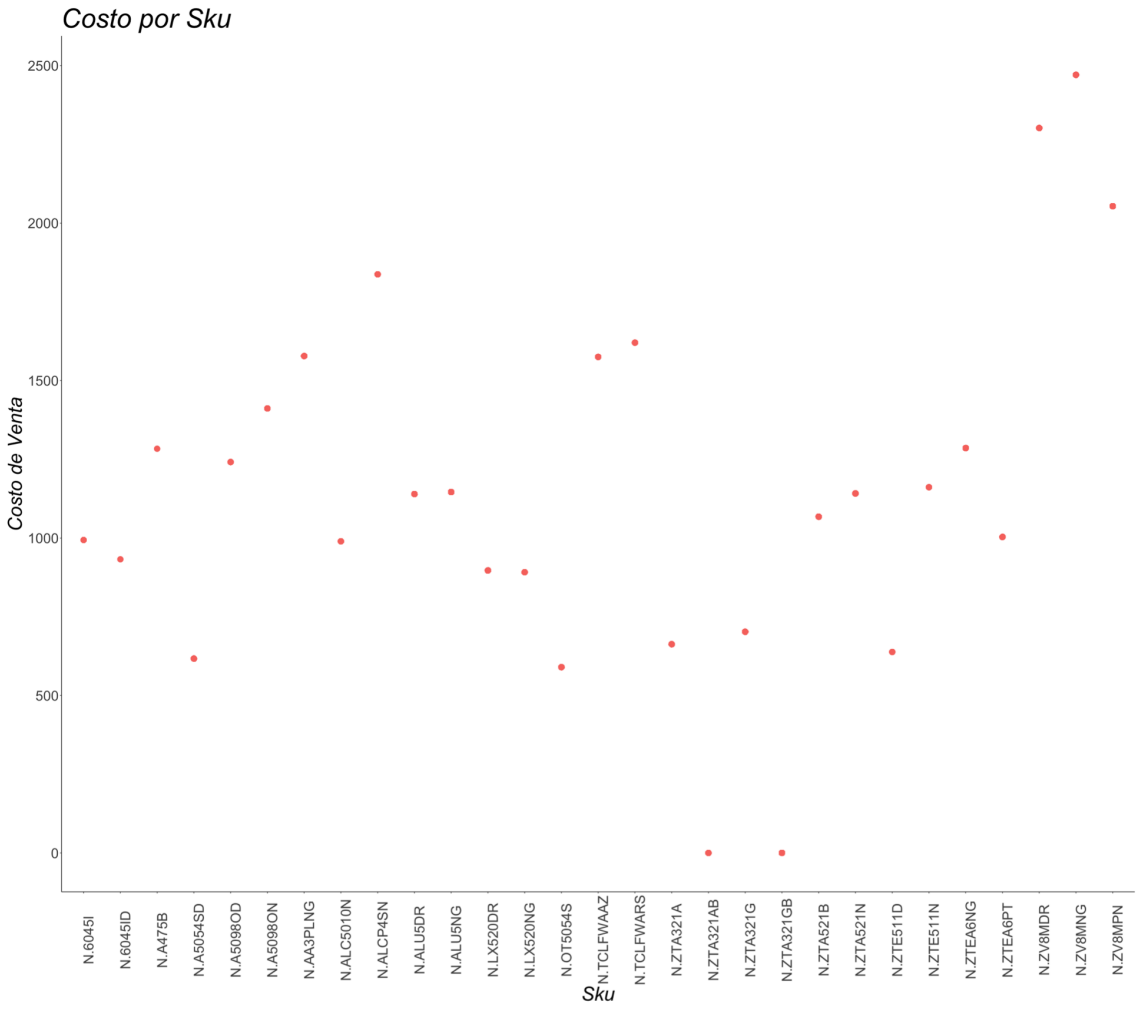

Imagen 1.4 Gráfica de dispersión de costos promedios respecto a sku

En la gráfica superior (***Imagen 1.4)*** podemos observar la dispersión de los costos de cada producto en la línea de teléfonos de ZTE. Podemos decir los 3 primeros productos más costosos resultan ser el SKU “N.ZV8MNG”, después en la posición dos el SKU “N.ZV8MOR” y en la tercera posición el SKU “N.ZV8MPN”. Al revisar estos productos en la ***Imagen 1.3*** podemos deducir que dos de estos productos, a pesar de ser de los más costosos, ocupan dos lugares entre los cinco más vendidos, así que podemos asegurar que el costo es un factor que está influyendo para que los productos se vendan.

Con la gráfica de la ***Imagen 1.5*** comprobamos que la mitad de los productos vendidos están por encima de los $1,100 hasta los $2,250 aproximadamente. Ya que los teléfonos son de gama baja podemos verificar que si los precios suben, de igual manera, tendría que subir el valor que la mitad de nuestros consumidores compran y que si nos enfocamos en desarrollar productos dentro de estos rangos podríamos mantener a nuestros consumidores. Además, podríamos enfocarnos también en mantener competitivos los precios y características de los teléfonos celulares por debajo de los $1,100 para no perder al 25% de nuestros clientes.

Los costos promedios de los productos se encuentran en la ***Imagen 1.6*** y las zonas en las que estos son vendidos.

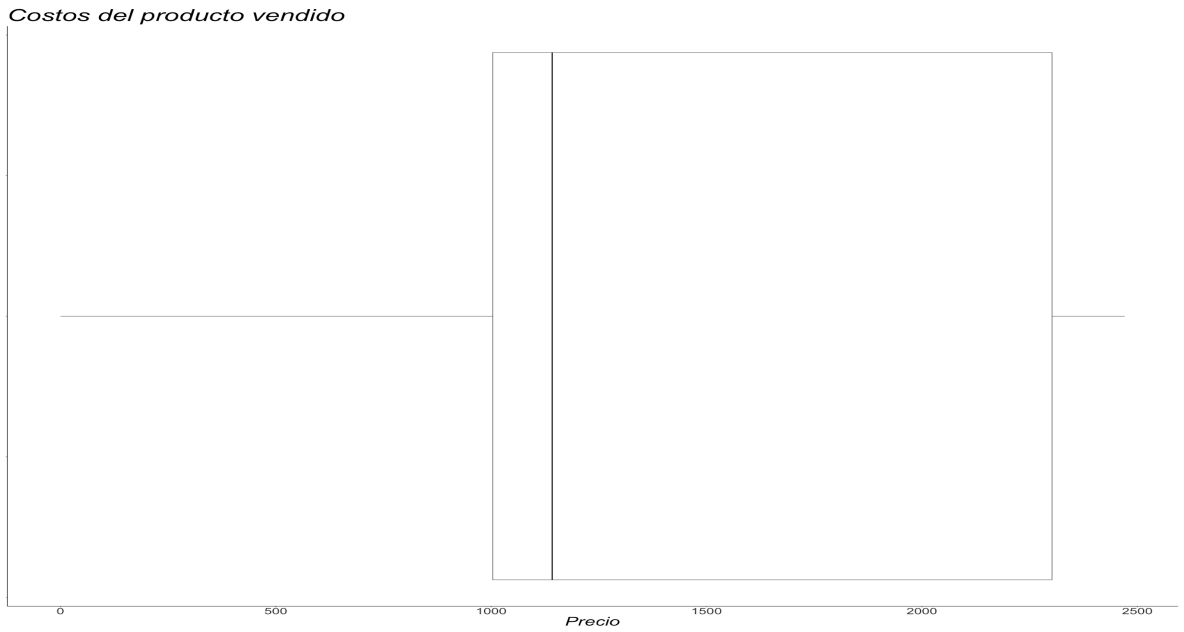

Imagen 1.5 Gráfica de costo del producto vendido

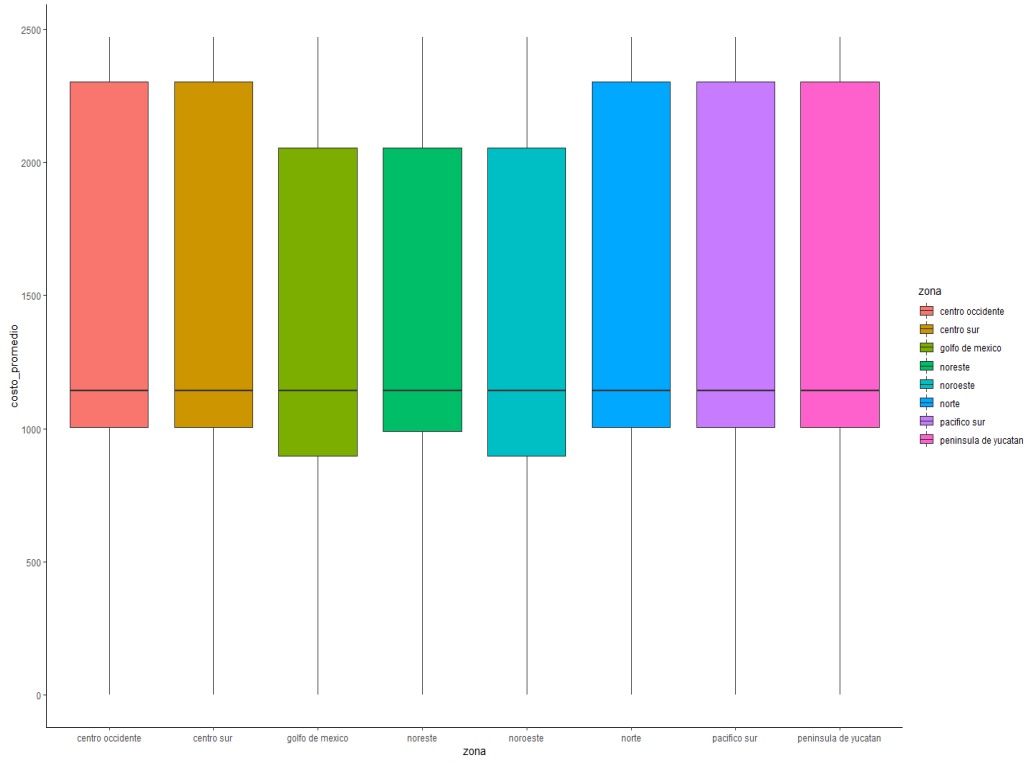

Imagen 1.6 Gráfica de caja para comportamiento de diferentes zonas respecto a las ventas promedio

Con esta gráfica podemos concluir que no hay variaciones significativas en los costos promedios, ya que todas las cajas a simple vista muestran la misma mediana, pero hay algunas zonas que tiene una ligera variabilidad en los costos promedios mínimos, como lo son el Golfo de México, noreste y noroeste. También podemos decir que el 75% de los costos promedios de las zonas centro occidente, centro sur, norte, pacifico sur y península de yucatán se encuentran alrededor de $2300.

Con todos estos datos, resaltamos que en general ZTE no está pasando por el mejor momento en cuanto a ventas se refiere. Estas han ido en declive desde que tenemos los registros de las ventas, hemos logrado identificar que muchos clientes los buscan por los precios bajos, sin embargo, esto no es suficiente para mantener el mismo nivel de ventas y ha impactado en los ingresos de la compañía.

Es muy probable que estén yendo en picada por la entrada de nuevos competidores al mercado como Huawei, que tiene excelentes productos a bajos precios y que han desplazado a ZTE como uno de los referentes en el mercado de celulares de gama baja. Por ello, recomendamos que se reestructure la compañía de manera que pueda seguir fabricando teléfonos celulares a bajo costo destinando los recursos que tengan a sus mejores modelos y los que ofrecen más ventas que serán los que marcan la pauta de lo que el cliente está buscando. Se necesita dejar de producir ciertos modelos que no aportan demasiado a la empresa y con este dinero realizar investigación y mejoramiento de procesos para que su producción mantenga costos bajos, siempre tomando en cuenta los requerimientos de los clientes y los beneficios que ello aportará a la compañía.

Ingeniería de características
Se utilizó como herramienta de ayuda RStudio con la finalidad de crear variables que ayuden a la situación y que aporten valor a nuestro modelo.

- **Lectura de datos:** Para poder hacer una correcta lectura de datos utilizamos la función *read. csv* para poder leer los datos del equipo y para visualizarlos utilizamos *head*. También utilizamos diversas funciones para ver las variables que tenemos (str) y ver características estadísticas (*summary*).
- **Índices:** En el cual tuvimos que crear índices por separado para algunas variables como punto_de_venta, mes, sku.
- **Agregar columnas con índices:** En este paso usamos la función *left join* para juntar la información de las tablas que previamente ya habíamos creado
- **Agrupar:** Agrupamos las columnas de los *id* que creamos, todo ello con el fin de tener codificadas las variables y poder hacer pronósticos que nos ayuden en la recabación de datos que aporten valor a nuestro modelo, usamos *group_by* que agrupa las ventas que se hicieron en el mismo punto de venta, el mismo mes del mismo producto *summarise* junta en una nueva columna (ventas_totales) el num_ventas si se cumplen las condiciones de arriba.
- **Completar series de tiempo:** Usamos la función *merge* para combinar las variables creadas con ventas_totales y también usamos is.na para reemplazar todos los NA que tenga en mis datos de ventas_totales por ceros.
- **Variable respuesta:** Ventas del siguiente mes, en la cual tengo que usar la función lead que me ayuda a subir o ir moviendo los números de mis variables.
- **Crear nuevas características: Conteos, promedios y rezagos**
    - **Paso1:** Tenemos que crear características de ventas de promedios por mes, tienda, producto y ventas totales; con las cuales se crean las características que necesitamos de manera rezagada más adelante. Primero se hace un conteo por duplas de características, usando las funciones *group by*, *summarise*, sum y *mean*. Al final de esto podremos saber las ventas totales y el promedio de ventas en todos los meses y en los puntos de ventas.
    - **Paso 2:** Incluir variables en datos completos. Usamos la variable *left_join* para poder realizar un traslado de las columnas ventas_totales_en-tienda_de_cada_mes y ventas_totales_en-tienda_de_cada_sku
    - **Paso 3:** Crear rezagos. Creamos los rezagos para ventas totales considerando promedios de 2 meses, 3 meses, etc.
    - **Paso 4:** NA de rezagos. Lo que tenemos que hacer es eliminar los NA que tenemos en los datos con la función *na.locf* y al final guardar los datos con *write.csv*.

Modelado

Se construyen modelos que permitan solucionar el problema. Como ingenieros industriales, una de las herramientas utilizadas con más frecuencia es “Promedios Móviles”, pero como científicos de datos hay más modelos que se pueden emplear.

El modelo de promedios móviles (*Moving Averages*) consiste en calcular una media de datos anteriores previamente determinados para pronosticar un periodo siguiente, generalmente funciona mejor cuando se realiza con series de datos sin tendencia.

Algunas consideraciones y limitaciones a tener en cuenta del modelo son:

- Si existe mucho cambio en los datos, entre más datos se tomen en cuenta para pronosticar más error puede existir ya que son lentas para reflejar un cambio en la tendencia. Lo contrario a si se utilizan pocos datos ya que puede aumentar el error si de la misma manera existe un cambio considerable en los datos.
- El peso asignado a cada dato tomado es el mismo.
- Pronostica únicamente un periodo o se tiene que hacer modificación a la fórmula utilizada.
- Tenemos que calcular el error de los pronósticos para ver cuál pronóstico es el que tiene el mejor comportamiento y poder compararlos. El error que se utiliza para comparar varios modelos es el llamado MAE, el cual significa “Error absoluto medio”.

¿Cómo realizamos el modelo?

Utilizando la herramienta de Python vista en la clase pudimos realizar fácilmente el modelado de promedios móviles simple utilizando algunos códigos que nos permiten mover nuestros datos y aplicar respectivamente el promedio para los datos dependiendo del promedio móvil que estemos solicitando a Python.

Con este modelo lo que vamos a hacer es calcular tres promedios móviles de los siguientes tres meses que tengo en mis datos. Vamos a comparar tres modelos para ver que tan bien se comporta si pronosticamos con el promedio de un mes, dos meses y tres meses. Posteriormente, obtenemos los errores MAE de cada uno de los modelos y graficarlos para buscar el mejor modelo, es decir, el que tenga menor error y un comportamiento estable.

Construcción del modelo para promedios móviles

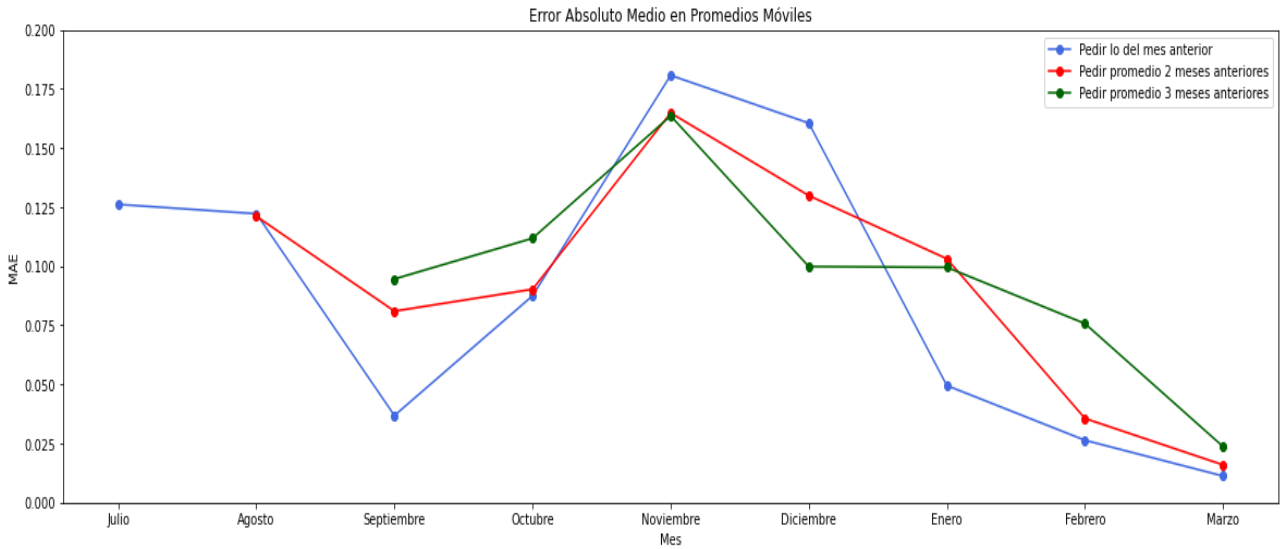

Imagen 2.1 Error Absoluto Medio en Modelo: Promedios Móviles

Resultados de los modelos de promedios móviles

En esta gráfica podemos observar los tres resultados de los modelos realizados con los promedios móviles hechos con un, dos y tres meses anteriores. Con esta gráfica concluimos que existe el mayor error en el mes de noviembre para los tres modelos.

El modelo del mes anterior (gráfica azul) se mantiene constante en los meses de julio y agosto, después las predicciones mejoran para el mes de septiembre porque el error disminuye repentinamente. Para los meses de octubre y noviembre el error aumenta radicalmente, continuando al mes de diciembre se observa que el error disminuye un poco y mantiene el mismo comportamiento en gran medida para los meses de enero, febrero y marzo, confirmando que las buenas predicciones son buenas para esos meses.

El modelo de dos meses anteriores (gráfica roja) comparado con el de un mes anterior (gráfica azul) nos muestra que ambas tienen el mismo comportamiento en el mes de agosto y octubre aunque en el mes de septiembre su error es más grande. En los meses de noviembre y diciembre los errores son menores pero para los meses siguientes (enero, febrero y marzo) los errores son mayores.

Finalmente, observamos que aunque el modelo de promedio móvil para tres meses anteriores (gráfica verde) comparado con el de dos meses anteriores (gráfica roja) tienen un comportamiento similar hay diferencias importantes. Durante los meses de septiembre y octubre los errores son mayores, en el mes de noviembre el error se posiciona en el mismo lugar pero parece mantener errores más grandes para los siguientes meses a excepción de diciembre donde el pronóstico fue más acertado.

Considerando la evidencia basada en la gráfica, creemos que el mejor modelo de promedios móviles resulta ser el de tres meses anteriores porque es el que vemos que se mantiene más estable en comparación a los otros dos. Podemos observar que el mayor error que presenta es en el mes de noviembre y en comparación con el modelo de dos meses anteriores su error se encuentra ligeramente más bajo, además de que el error en el mes de diciembre también es más bajo en comparación a los otros dos modelos.

Construcción de modelo de aprendizaje de máquina

El modelo que vamos a construir para poder realizar predicciones para la demanda de la compañía ZTE es el modelo de aprendizaje máquina llamado: Árboles de Decisión. Un árbol de decisión es un modelo predictivo formado por reglas binarias con las que se consigue repartir las observaciones en función de sus atributos y predecir así el valor de la variable respuesta.

Algunas consideraciones sobre los árboles de decisión son:

- Se utiliza para problemas de clasificación y regresión.
- Son fáciles de construir, interpretar y visualizar.
- Mientras más información se tenga, mejores son los resultados.
- No siempre se hace uso de todos los predictores.
- No es necesario que se cumplan los supuestos de la regresión lineal (linealidad, normalidad, homogeneidad).

Algunas limitaciones sobre los árboles de decisión son:

- La máquina se ajusta a aprender casos particulares que le enseñamos y es incapaz de reconocer nuevos datos de entrada (puntos atípicos).
- Se crean árboles sesgados si una de las clases es más numerosa que otra.
- Se pierde información cuando se utiliza para categorizar variables numéricas continuas.

Construcción del modelo de aprendizaje máquina para “árboles de decisión”

Utilizando la herramienta de Python vista en la clase pudimos realizar fácilmente el modelado de los árboles de decisión utilizando código. Realizamos un árbol de decisión para regresión con una profundidad de uno (es decir que solo tenga una división) para obtener una predicción. Lo mismo hicimos con una profundidad de cinco con el objetivo de ver cuál es su comportamiento y elegir el mejor entre ellos.

Con validación cruzada, generamos las particiones para entrenamiento y prueba. Entrenamos el modelo con las Xs y Ys de la primera partición y generamos una predicción respecto al entrenamiento y prueba; posteriormente se evalúa ambas etapas.
Nuestra referencia será el error de entrenamiento, es aquel que muestra la situación optimista mientras que el error de prueba es el que muestra cómo se va a comportar en el mundo real, con datos que desconoce y de los cuales no tiene una referencia más allá de lo que pueda reconocer del entrenamiento.

Por lo tanto, en el entrenamiento aglutinamos todos los conjuntos de datos y observamos los resultados que obtuvimos en las predicciones comparando la columna de “y_ ventas_siguiente_mes” contra la columna de “ventas_por_mes_pred”, se obtuvo ambas columnas el error, cuyo resultado nos indicará que tan ha acertado es nuestro modelo en sus predicciones.

Continuamos proporcionando datos al modelo de árboles de decisiones para conocer cuál será su comportamiento. Al igual que con el entrenamiento, ejecutamos la prueba para las ocho particiones y posteriormente calculamos el error de cada una.

Finalmente, comparamos las métricas de error correspondientes al entrenamiento y a la prueba de los tres modelos ejecutados.

Resultados del modelo de árboles de decisión

Utilizamos uno y cinco nodos de profundidad, y los resultados son mostrados en la siguiente tabla:

metrica	conjunto	mes	modelo_base	dt_1_profundidad	dt_5_profundidad
mae	entrenamiento	julio		107.901	99.659
mae	entrenamiento	agosto		75.194	69.865
mae	entrenamiento	septiembre		54.874	51.519
mae	entrenamiento	octubre		61.221	58.604
mae	entrenamiento	noviembre		7.994	80.802
mae	entrenamiento	diciembre		72.656	83.637
mae	entrenamiento	enero		65.369	73.817
mae	entrenamiento	febrero		58.583	66.035
mae	prueba	agosto		5.665	76.142
mae	prueba	septiembre	94.599	14.234	21.078
mae	prueba	octubre	111.816	80.262	80.002
mae	prueba	noviembre	163.604	154.817	153.633
mae	prueba	diciembre	99.872	60.203	169.335
mae	prueba	enero	99.541	21.646	2.662
mae	prueba	febrero	75.731	11.084	11.534
mae	prueba	marzo	23.699	1.303	1.539

Imagen 3.1 Tabla de resultados de los modelos

Nota: No se entrenan a los datos en estadística tradicional, es por ello que en el modelo base, el cual corresponde al modelo de promedios móviles, no tenemos datos de entrenamiento.

Comparación de métrica con gráficas de desempeño

Con ayuda de Python realizamos las gráficas con los errores obtenidos en los modelos de aprendizaje máquina de árboles de decisión y el modelo de promedios móviles.
Paso 1. Filtrar errores de entrenamiento y prueba
Paso 2. Graficar

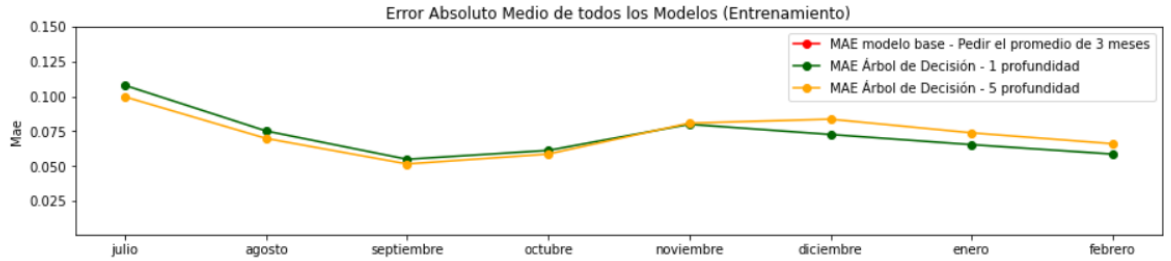

Imagen 4.1 **Gráfica de MAE de los "Entrenamiento"**

Con la ***Imagen 4.1*** Solo podemos comparar los errores absolutos medios de los árboles de decisión para uno y cinco nodos de profundidad ya que no existen datos de entrenamiento para el modelo de promedios móviles. Podemos decir que el mejor modelo resulta ser el árbol de decisión con profundidad de 5 ya que presenta un menor error en los meses desde julio hasta octubre. En noviembre se aprecian los dos modelos similares y en diciembre, enero y febrero se observa que el error es mayor en comparación con el árbol de decisión de 1 de profundidad pero en realidad tenemos la mayoría de puntos de error por debajo de la gráfica verde, por lo que concluimos que el árbol de decisión con cinco de profundidad es el que tiene menor error y es el mejor modelo para los datos de entrenamiento.

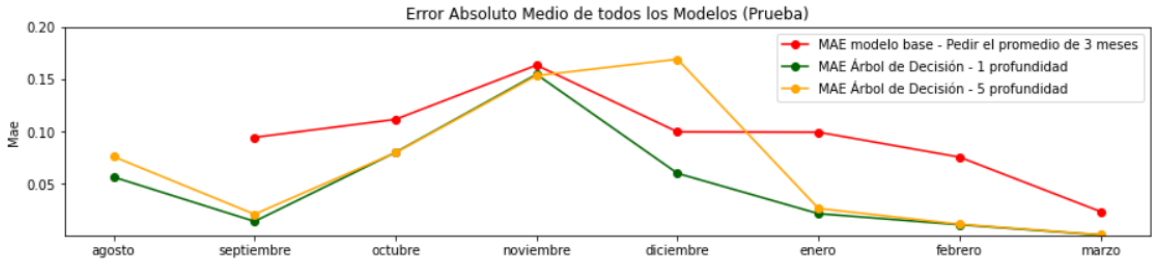

Imagen 4.2 **Gráfica de MAE de las "Prueba"**

En la ***Imagen 4.2*** logramos observar el MAE del modelo base (pedir el promedio de tres meses) y es muy claro que tiene un error más grande en todos los meses comparado a los otros dos modelos utilizados, excepto en el mes de diciembre donde su error se encuentra aproximadamente en un punto medio comparado con aquel de los árboles de decisión.

En la representación correspondiente al árbol de decisión de cinco de profundidad (gráfica amarilla) podemos observar que sus errores son, exceptuando diciembre donde es el mayor de los tres modelos, menores comparado con el modelo de promedios móviles para tres meses anteriores.

Y por último, la línea verde del error absoluto medio que corresponde al árbol de decisión de uno de profundidad observamos que los errores en comparación con los otros dos modelos resultan ser iguales o en varios casos menores que el de los otros dos modelos analizados.

Con todo lo anterior descrito, concluimos que el mejor modelo para la predicción de demanda de la empresa ZTE resulta ser el modelo de aprendizaje máquina “Árbol de decisión de profundidad uno”. Sería eficiente y conveniente que ZTE comenzara a aplicar dicho modelo de predicción ya que tiene errores pequeños en casi todo los meses, aunque es importante comentar que se recomienda revisar qué es lo que sucede especialmente en el mes de noviembre ya que es el mes en el que presenta un mayor error y esto podría generar a la empresa ciertos problemas como costos, inventarios, logística, sobreproducción, etc.

Resultados

La compañía ZTE encargó contestar la pregunta siguiente: **¿Cuántas unidades de cada producto de ZTE, se van a vender en todos los puntos de venta, al siguiente mes de registro?**

Con esta pregunta en mente es que este proyecto fue elaborado y después del análisis realizado, utilizando el árbol de decisión con profundidad de nodo 1 logramos responder de la manera siguiente:

Punto de Venta	Producto	Ventas totales
40	4	1
204	12	1
847	12	1
857	12	1
906	1	1
906	3	1
949	16	1
1014	16	1
1071	10	1
1128	3	1
1169	23	1
1255	15	1
1297	10	1
1297	11	1
1362	8	1
1362	11	1
1363	23	1
1454	4	1

1454	5	1

La compañía puede esperar tener 19 ventas en total, de las cuales, el producto con más ventas será el 12 con un número de 3 ventas en total.

Sku	Ventas Totales
1	**1**
3	**2**
4	**2**
5	**1**
8	**1**
10	**2**
11	**2**
12	**3**
15	**1**
16	**2**
23	**2**

A reserva del error del modelo, lo que podemos comunicar es que prácticamente sus ventas a día de hoy son **nulas**, y el pronóstico no parece muy alentador. Tomando en cuenta eso nuestra recomendación se reduce a enfocar sus esfuerzos en impulsar los once SKU que tendrán por lo menos una venta en este mes, si eso no es posible nuestra otra recomendación sería dejar de producir cualquiera de estos 29 productos que se analizaron ya que con 99% de seguridad no están cubriendo ni siquiera los gastos de fabricación. Se deben recortar los gastos, debe haber una reestructuración de la empresa y realizar un análisis a profundidad para conocer las causas que llevaron a que ZTE fuera desplazada del mercado de telefonía.

Recomendamos utilizar también la metodología “Design for Lean Six Sigma” para la elaboración de un nuevo producto innovador, que los clientes quieran comprar y que esté enfocado en las necesidades de los mismos por el poco o nulo mercado que aún se pueda rescatar. Además, por ser un mercado extra competido también recomendamos realizar una evaluación del proyecto antes de iniciarlo para conocer si será viable o si la empresa finalmente será declarada en bancarrota.

Conclusiones finales

En general podemos concluir que es importante la utilización de modelos predictivos, ya que con esto una empresa podría ser más competitiva y mejorar su procesos para poder satisfacer las demandas de sus clientes con éxito. En el caso de ZTE, que navega en un sector muy competido, la importancia de la aplicación de varios modelos de predicción es inmensa. El saber cual tiene el mejor desempeño y por lo tanto las predicciones más acertadas le brindaría una confianza más grande al momento de tomar decisiones y poder mantener su parte de mercado e incluso pensar en adelantar a su competencia.

Más allá del caso mismo, en este proyecto pudimos apreciar las aplicaciones de herramientas de apoyo para una gran cantidad de datos como lo fue RStudio y Python. Para el procesamiento de muchos datos resulta ser más fácil poder aplicar comandos de apoyo para realizar fácilmente la lectura, limpieza, modelado, etc. Nos dimos cuenta, especialmente en la limpieza de datos, que con la aplicación de filtros en RStudio nos ahorrarnos mucho tiempo en comparación de lo que pudo ser el realizar este procedimiento manualmente buscando todos los errores uno por uno o con otra herramienta menos capaz.

También apreciamos que la aplicación de varios modelos de predicción resulta ser de vital importancia ya que podemos encontrar cuál de ellos tiene el mejor desempeño, que pueda brindar a una empresa resultados más certeros con el mínimo margen de error y permitir que la empresa pueda tomar decisiones acertadas.

Una vez que conoces y entiendes el problema la aplicación de modelos resulta ser una herramienta fácil de utilizar pero sobretodo de gran impacto pues los mismo modelos nos permiten hacer mejoras en un proyecto, además brindan una respuesta barata, eficiente, con precisión y nos permiten entender de forma más clara y sencilla problemas muy complejos.

Bibliografía

[1] pr.noticias. (2021). El presidente de ZTE, Xu Ziyang: “En esta nueva etapa, el 5G se encontrará tanto con retos tecnológicos como con incertidumbres comerciales que debemos solucionar juntos”.. 01/03/2021, de pr.noticias.com Sitio web: https://prnoticias.com/2021/02/23/el-presidente-de-zte-xu-ziyang-en-esta-nueva-etapa-el-5g-se-encontrara-tanto-con-retos-tecnologicos-como-con-incertidumbres-comerciales-que-debemos-solucionar-juntos/

[2] El español. (2021). Noticias sobre ZTE. 01/03/2021, de El Androide Libre S.L. Sitio web: https://elandroidelibre.elespanol.com/tag/zte

Anexo

Nombre del tipo de dato	*Clasificación*	*# de datos*	*Problema de calidad*
Punto de Venta	Carácter	1 462 niveles	Hay registros únicamente con mayúsculas y nombres repetidos con pequeñas diferencias
Fecha	Carácter	289 niveles	No existen problemas
Mes	Carácter	15 niveles	Hay registros con letra en lugar de números
Año	Entero	3 niveles	Hay números incompletos
Número de ventas	Entero	1 nivel	No existen problemas
SKU (Producto vendido)	Carácter	29 niveles	No existen problemas
Marca	Carácter	6 niveles	No hay homogeneización de ZTE
Gamma (Nivel del producto vendido)	Carácter	1 nivel	No existen problemas
Costo promedio	Número	28 niveles	Hay muchos registros con solo un 0
Zona	Carácter	9 niveles	Hay registros con mayúsculas y no está claro cómo se definen las zonas
Estado	Carácter	35 niveles	Hay registros que contienen estados inexistentes
Ciudad	Carácter	208 niveles	Hay errores de ortografía y ciudades inexistentes
Latitud	Número	1437 niveles	Hay registros a los que les falta un punto decimal
Longitud	Número	1410 niveles	Hay registros a los que les falta un punto decimal

Tabla 1. Datos de variables

		Nombre	Duracion	Inicio	Terminado	Nombres del Recurso
1		**Proyecto para predicción de demanda de productos de telefonia ZTE**	68.083 day...	**3/03/21 08:00 AM**	**7/06/21 08:40 AM**	
2		**Etapa 1**	3 days?	**3/03/21 08:00 AM**	**5/03/21 05:00 PM**	
3		Describir la situación actual	1 day?	3/03/21 08:00 AM	3/03/21 05:00 PM	Nancy Escamilla
4		Entender y describir problemática en términos de ZTE	1 day?	3/03/21 08:00 AM	3/03/21 05:00 PM	Carlos Muciño
5		Entender y describir problemática en términos de ciencia de datos	1 day?	4/03/21 08:00 AM	4/03/21 05:00 PM	Cesar Trejo
6		Plasmar los objetivos	1 day?	5/03/21 08:00 AM	5/03/21 05:00 PM	Nancy Escamilla
7		Estructurar el proyecto y hacer un plan preliminar	0.167 days?	5/03/21 08:00 AM	5/03/21 09:20 AM	Nancy Escamilla;Carlos Muciño;Cesar Trejo
8		**Etapa 2**	0.333 days?	**8/03/21 08:00 AM**	**8/03/21 10:40 AM**	
9		Describir los datos crudos	0.333 days?	8/03/21 08:00 AM	8/03/21 10:40 AM	Nancy Escamilla;Carlos Muciño;Cesar Trejo
10		Detectar problemas de calidad	0.333 days?	8/03/21 08:00 AM	8/03/21 10:40 AM	Nancy Escamilla;Carlos Muciño;Cesar Trejo
11		**Etapa 3**	26.333 day...	**22/03/21 08:00 AM**	**27/04/21 10:40 AM**	
12		Limpiar los datos	8.333 days?	22/03/21 08:00 AM	1/04/21 10:40 AM	Nancy Escamilla;Carlos Muciño;Cesar Trejo
13		Análisis Exploratorio	0.5 days?	26/04/21 08:00 AM	26/04/21 01:00 PM	Nancy Escamilla;Carlos Muciño
14		Seleccionar y construir variables para la etapa de modelado	0.333 days?	27/04/21 06:00 AM	27/04/21 10:40 AM	Nancy Escamilla;Carlos Muciño;Cesar Trejo
15		**Etapa 4**	3 days?	**27/04/21 08:00 AM**	**29/04/21 05:00 PM**	
16		Construir modelo de pronóstico	0.667 days?	27/04/21 08:00 AM	27/04/21 02:20 PM	Nancy Escamilla;Carlos Muciño;Cesar Trejo
17		Construir modelo de aprendizaje maquina	1 day?	29/04/21 08:00 AM	29/04/21 05:00 PM	Nancy Escamilla;Carlos Muciño;Cesar Trejo
18		**Etapa 5**	9.333 days?	**5/05/21 08:00 AM**	**18/05/21 10:40 AM**	
19		Construir gráficas para contrastar desempeño de modelos	2.667 days?	5/05/21 08:00 AM	7/05/21 02:20 PM	Nancy Escamilla;Carlos Muciño;Cesar Trejo
20		Construir conclusiones de resultados	1.333 days?	15/05/21 08:00 AM	18/05/21 10:40 AM	Nancy Escamilla;Carlos Muciño;Cesar Trejo
21		**Etapa 6**	4 days?	**21/05/21 08:00 AM**	**26/05/21 05:00 PM**	
22		Crear reporte final del proyecto	1 day?	21/05/21 08:00 AM	21/05/21 05:00 PM	Nancy Escamilla;Carlos Muciño;Cesar Trejo
23		Crear presentación ejecutiva	1 day?	26/05/21 08:00 AM	26/05/21 05:00 PM	Nancy Escamilla;Carlos Muciño;Cesar Trejo
24		**Etapa 7**	0.083 days?	**7/06/21 08:00 AM**	**7/06/21 08:40 AM**	
25		Presentación del proyecto	0.083 days?	7/06/21 08:00 AM	7/06/21 08:40 AM	Nancy Escamilla;Carlos Muciño;Cesar Trejo

Imagen 1.1 Estructura de Proyecto ZTE

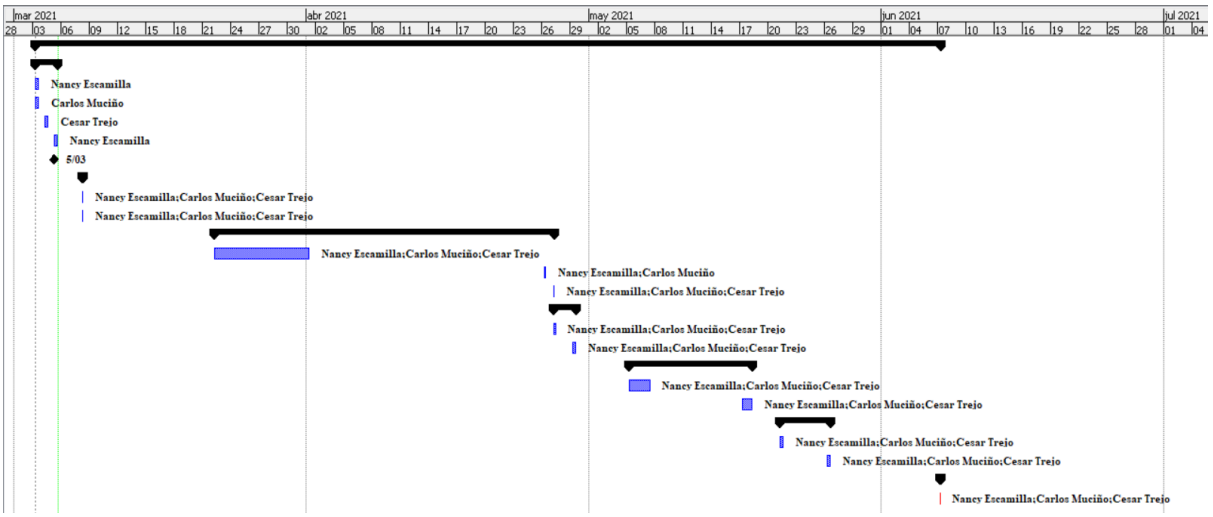

Imagen 1.2 Gantt de Proyecto ZTE