Tecnológico de Monterrey

LABORATORIO DE DISEÑO Y OPTIMIZACIÓN DE OPERACIONES

Ana Luisa Masetto Herrera

Modelos tradicionales vs. Modelos de Aprendizaje de Máquina para predicción de demanda de productos de “zte”

Integrantes:

Alejandra Santillan Palos

A01365626

Eda Eraldí Martínez Vázquez

A0365784

Justine Gutiérrez Mendoza

A01365878

José Luis Estrada Bastida

A01365811

Guillermo Martínez Ávila

A01365593

Fecha de entrega:

19 de noviembre, 2021

Etapa 1

1.1 Descripción de la situación actual.

ZTE es una marca china que busca posicionarse en el top 3 en el mercado en Estados Unidos y México, busca posicionarse de gran forma en México ya que sabe que así podrá tener una gran entrada en el mercado latino americano.

Se caracterizan por vender celulares a un precio justo, quieren ofrecer productos innovadores y con una excelente calidad, un gran diferenciador respecto a sus competidores son los bajos precios que maneja, sus precios empiezan desde los $1,199 hasta los $18,999, es una de las mejores marcas a precio económico que utilizan el sistema operativo de android, su mayor competidor es la marca Huawei ya que incluso son marcas que nacieron en el mismo país. Una gran estrategia de marketing que utiliza ZTE es ser patrocinador de los equipos de la NBA (los Rockets de Houston y los Timberwolves de Minnesota), así como del equipo de fútbol de la liga española Sevilla F.C y del equipo de fútbol de primera división de los Pumas de la UNAM.

De acuerdo con CIU en 2016 la marca contaba con el 9% del mercado mexicano en ventas colocando en la cuarta posición, actualmente las 4 marcas que tienen más ventas son Samsung, Motorola, Huawei y Apple.

Es muy importante realizar este tipo de proyectos, ya que por ejemplo en este tema la marca Apple es la que cuenta con una mejor cadena de suministro y utilizan la metodología de justo a tiempo, para poder utilizar esta metodología de forma correcta es necesario tener un buen sistema de predicción para saber cuántas unidades se van a vender de cada producto, es necesario aplicar herramientas y conceptos de ciencias de datos para poder obtener la información más relevante y poder aplicar este tipo de metodologías de mejor forma.

1.2 Entender y describir la problemática (en términos del negocio).

El problema al que se enfrenta esta empresa es el de construir un eficiente portafolio de productos, mediante una predicción de la demanda. Los costos logísticos de una empresa son el segundo costo más grande sólo después de los costos de materias primas, estos costos pueden ir desde un 4 hasta un 40% de los costos totales de la empresa. De estos costos logísticos, los rubros más importantes son el manejo de inventarios y el transporte, de ahí la necesidad de tener buenos pronósticos de demanda. Este pronóstico de la demanda afecta directamente la planificación estratégica del negocio ya que involucra la administración de la cadena de suministro, la compra de materias primas y la distribución de pedidos. Dentro de los problemas que nos podemos encontrar si no tenemos un eficiente pronóstico de la demanda están:

➔ **Disminución de la satisfacción del cliente:** Si los plazos de entrega se incrementan y no hay una distribución eficiente, el cliente no recibe sus productos a tiempo por lo que disminuye su satisfacción. Para evitar esto es necesario estimar una demanda real, para producir con tiempo los productos necesarios y el cliente los reciba en tiempo y forma.

➔ **Mala planeación de materias primas y recursos:** La planeación de manufactura y ensamble se tiene que hacer con meses de anticipación para ordenar los materiales que se van a requerir, si la demanda estimada es incorrecta, no se tendrán algunos componentes para ciertos modelos y se tendrá exceso de componentes de otros porque el ticket de producción tendrá que estar en constante cambio.

➔ **Altos costos por inventarios:** Al tener un mal pronóstico, se tendrá una bajo demanda para ciertos productos y una sobre demanda de otros. La baja demanda causará un exceso de inventarios ya que son productos que el cliente no está comprando, en un escenario optimista, a pesar de los altos costos de inventarios se quedará almacenado hasta que haya una compra, pero en un escenario pesimista si se queda más de ciertos meses en inventarios deberá de ser desechado porque se volverá obsoleto al salir nuevos modelos al mercado.

➔ **Deficiente planificación de la estrategia de ventas:** La estrategia de ventas incluye la fijación de precios, promoción de productos y decisiones de compra, si no se tiene un pronóstico de venta adecuado, la estrategia de ventas tampoco se llevará a cabo de manera eficiente.

➔ **Falta de comunicación con proveedores y departamento de compras en cuanto a ticket de producción:** Al tener cambios drásticos en la demanda por su mal pronóstico, habrá también cambios drásticos en el ticket de producción por lo que los proveedores y el departamento de compras tendrán que correr con los cambios en las órdenes y esto podrá traer más problemas en los plazos de entrega y en la comunicación.

En base a los puntos expuestos, es por lo que a lo largo de este proyecto se trabajará en armar una buena predicción de la demanda para la empresa telefónica “zte”.

1.3 Entender y describir la problemática (en términos de ciencia de datos, tipo de tarea, tipo de datos, etc.).

➔ **Datos:** La base de datos tiene registros de variables numéricas (enteras y flotantes), fechas y de carácteres. Tenemos dos tipos de variables, de entrada y de salida. Las variables de entrada o “x”, son las variables que nosotros conocemos y por lo tanto ingresamos al modelo para aprender y adquirir conocimiento de ellas (Ventas totales, puntos de venta, SKU, fecha, estado). La variable de salida “y”, es la información que yo quiero obtener con base en las variables de entrada y para el proyecto en cuestión, son las ventas del siguiente mes por tienda (abril).

➔ **Problema:** Nos encontramos con un problema de regresión, ya que la variable de respuesta “y” es del tipo numérico y la predicción se realiza en términos de valores continuos.

➔ **Pregunta:** La pregunta que queremos responder es: ¿Cuántas unidades de cada producto de la marca “Zte”, se van a vender en todos los puntos de venta, al siguiente mes de registro “abril”?

➔ **Limpieza de datos:** Los registros deben ser limpiados primero, corrigiendo entradas erróneas para poder tener todas las variables clasificadas y en un formato estándar manipulable. De lo contrario, no será posible tener cálculos certeros y útiles.

➔ **Cálculos:** Una vez que la base de datos esté lista, se pronosticará con ella la cantidad de dispositivos que se venderán en cada punto de venta registrado. Se espera pronosticar la cantidad de dispositivos a vender (variable cuantitativa). Al ser la fecha parte del registro y una variable a pronosticar (cantidad de dispositivos a vender), este modelo de regresión resulta de utilidad para pronosticar las ventas futuras, para esto el algoritmo hará cálculos como, pero no limitado a: promedio,operaciones aritméticas, desviación estándar, etc.

➔ **Interpretación:** El algoritmo regresará los valores que se esperan ser vendidos en cada punto de venta; regresa información útil para la toma de decisiones orientada al objetivo del proyecto.

1.4 Plasmar los objetivos.

Objetivo para la empresa: El objetivo principal que buscamos para la empresa es el de construir un eficiente portafolio de productos mediante la predicción de la demanda de sus dispositivos móviles, esto provocará una disminución de costo de inventarios, transporte y logística, una mejor planeación del ticket de producción y compra de las materias primas, mejor comunicación con los demás departamentos y proveedores y sobre todo una excelente satisfacción al cliente.

Objetivo de la materia: Consideramos como objetivo de la materia el poder ser capaces de aplicar los conocimientos que adquiramos durante el semestre en este proyecto de una forma dinámica, para poder poner en práctica la ciencia de datos y todo lo relacionado con ella en este proyecto, con una aplicación

efectiva cumpliendo con los objetivos de aprendizaje de la materia, ayudando a nuestro crecimiento y desempeño como equipo en la misma ya que la consideramos bastante importante para el resto de nuestra carrera y también para la aplicación en la vida profesional.

Objetivo personal: Uno de los objetivos personales que buscamos es el poder utilizar con mayor profundidad de una manera práctica la ciencia de datos para poder responder las preguntas que se realicen al principio y así con los datos obtenidos, tanto los estructurados como los no estructurados poder identificar patrones y lograr encontrar varias de las cosas más importantes que ayuden a la toma de decisiones. También el poder aprovechar, identificar y utilizar las diversas herramientas, habilidades, tipos de tareas y conocimientos que se tengan que utilizar en el proceso.

1.5 Estructurar el proyecto y hacer un plan preliminar.

1.5.1 Realizar y plasmar un plan, organización de tiempos y distribución de tareas, para poder realizar el proyecto.

Consideramos que es de suma importancia que todos los integrantes se involucren en todas las etapas que se desarrollaran para la creación y análisis de este proyecto, es por ello que hemos realizado un listado de tareas y entregas específicas que nos ayudarán a tener una noción del avance del proyecto y las áreas que se pueden mejorar, sabemos que es importante que todos los integrantes están involucrados, por lo que las revisiones serán hechas por todas antes de cada entrega, sin embargo al ser un proyecto largo nos dividiremos las tareas para que el desarrollo sea más organizado, comprometiéndonos todos a cumplir con las tareas que nos toquen teniendo en cuenta que somos un equipo y que el trabajo de cada miembro impacta en el proyecto, la tabla de organización será un anexo en esta entrega.

Vea Anexo No.1 - “Desarrollo Preliminar del Proyecto”

1.5.2 Elaborar un Gantt con información del punto anterior.

En base a la serie de actividades que desarrollamos en el punto anterior realizamos un diagrama de Gantt para tener de forma más visual un recurso que nos ayude a contemplar nuestras fechas de entrega, el tiempo que debemos de invertir en cada actividad y el avance general que tenemos sobre nuestro proyecto, lo presentamos como en Anexo No.2.

Vea Anexo No.2 - “Diagrama de Gantt del Desarrollo del Proyecto”

Etapa 2

2.1 Describir los datos crudos.

Los datos crudos se encuentran en archivo con terminación “csv” correspondiente a un archivo de valores separados por comas que se puede visualizar con Microsoft Excel pero con mayor facilidad se puede trabajar en “R”.

Los datos tienen una dimensión de 24,089 registros o renglones en 14 variables o columnas.

Diccionario de las variables en los conjuntos de datos:

1. punto_de_venta: Diferentes puntos de venta donde, donde la empresa tiene operaciones.
    a. Tipo de variable: Character. Trata de un texto ya que específicamente es un carácter (no es número con el que se realice alguna operación)
2. fecha: Fecha en la que se hizo la compra del dispositivo móvil con formato DD/MM/AAAA
    a. Tipo de variable: Character. Usado como texto ya que específicamente es un carácter y los “/” no indican ninguna operación sino que integran el carácter.

3. mes: Mes en el que se hizo la compra del dispositivo móvil con formato MM (1 para enero, 2 para febrero, 3 para marzo y así consecutivamente)

    a. Tipo de variable: Factor. Es un elemento contribuye a la producción del resultado, el cual es dividido en diversos niveles, es decir que no se mueven en orden pero sí a través de ciertos grupos. Este elemento se le relaciona un nombre al número ya que con el número es posible la realización del proceso y obtención óptima del resultado y con el nombre la relación del número.

4. anio: Año en el que se hizo la compra del dispositivo móvil con formato AAAA

    a. Tipo de variable: Factor. Elemento que para la contribución en los diversos resultados se maneja en niveles.

5. num_ventas: Cantidad de dispositivos que se vendieron.

    a. Tipo de variable: Integer Variables tomadas como números enteros a pesar de que el proceso o resultado obtenido tenga algún decimal solo es tomado y usado numeros enteros.

6. sku: “Stock-keeping unit” es el código de referencia asignado a los dispositivos móviles para poder identificarlos en los inventarios.

    a. Tipo de variable: Factor. Variable con cantidad finita de valores donde sus elementos contribuyen a la producción del resultado a través del uso de categorías.

7. marca: Se refiere a la marca del dispositivo móvil.

    a. Tipo de variable: Character. Trata de un texto ya que es un carácter, el cual no es número con el que se realice alguna operación o influya a la obtención de los resultados.

8. gamma: Se refiere a la gamma(suma de productos divididos en cierta categoría) a la que pertenece el dispositivo móvil.

    a. Tipo de variable: Character. Se trata de un texto ya que este es un carácter, el cual a pesar de ser resultado de una operación no se usa como número que afecte al resultado.

9. costo_promedio: Precio o importe del dispositivo móvil.

    a. Tipo de variable: Numeric. Esta variable representa un número con el cual se realizan operaciones aritméticas y estas logran definir ciertos resultados.

10. zona: Zona de México donde se encuentra el punto de venta donde se hizo la compra.

    a. Tipo de variable: Factor. Se realiza en diversos niveles (número finito) otorgando de la facilidad de pertenencia a uno o más grupos o categorías.

11. estado: Estado de México donde se encuentra el punto de venta donde se hizo la compra.

    a. Tipo de variable: Factor. Variable hecha de estructura de datos para el manejo categórico de la variable dando la facilidad de pertenencia a uno o unos grupos o categorías.

12. ciudad: Cuidad donde se encuentra el punto de venta donde se hizo la compra.

    a. Tipo de variable: Factor. Variable que pertenece a una o más categorías con una cantidad finita de valores.

13. latitud: Distancia angular que hay desde el punto de venta donde se hizo la venta, hasta el paralelo del ecuador; se mide en grados, minutos y segundos sobre los meridianos.

a. Tipo de variable: Numeric. Variable la cual representa algo(grados, minutos, segundos, etc) numéricamente para lograr realizar operaciones aritméticas y así obtener resultados de operaciones.

14. longitud: Distancia angular que hay desde el punto de venta donde se hizo la venta, hasta el meridiano de Greenwich; se mide en grados, minutos y segundos sobre los meridianos.

a. Tipo de variable: Numeric. Variable la cual representa algo (grados, minutos, segundos, etc) numéricamente para lograr realizar operaciones aritméticas y así obtener resultados de operaciones.

2.2 Detectar problemas de calidad (solamente enunciar los problemas, en la siguiente etapa se mencionará cómo se hizo la limpieza).

En la primer variable de puntos de venta se encuentran problemas de que algunos puntos son escritos en mayúsculas y en otras en minúsculas (hablando del mismo punto), también problemas de que una palabra omite una letra (por ejemplo azul sin a) y puntos de venta escritos con espacio y sin espacio (nuevamente hablando del mismo punto)

En el caso de la variable de mes hay algunos meses que son escritos en letra y otros escritos en número.

Para la variable de anio algunos vienen escritos con los 4 dígitos y otros con los últimos 2.

La variable de marca viene escrita de 6 diferentes formas, en esta caso mayúsculas, minúsculas, con doble t, escrita 2 veces en mayúscula y en minúscula, y escrita con doble Z.

En la variable de zona la zona centro occidente está mal escrita, occidente no tiene i.

Para la variable de estado hay varios que no son estados como por ejemplo Monterrey que es una ciudad en Nuevo León.

Para las últimas variables de longitud y latitud hay errores de valores fuera de rango.

Estos son problemas de calidad ya que son datos no homogéneos (puntos de venta, mes, anio marca), faltas de ortografía (marca, zona, puntos de venta) e incongruencia en los datos (estado, latitud y longitud).

Etapa 3: Preparación de los datos

3.1 Limpiar los datos: todos los problemas de calidad que se encontraron deben de corregirse.

La limpieza de datos corresponde a un proceso que nos permite asegurar la calidad de los datos que se emplearán durante el proyecto, este proceso se encarga de corregir los problemas de calidad que se detectaron en nuestra etapa 2 del CRISP-DM, para que al finalizar contemos con un conjunto de datos listos para su análisis e implementación, es por ello que este es uno de los pasos más importantes y meticulosos de nuestro proyecto.

La metodología que seguimos fue bastante sencilla, primero hicimos un análisis general de los datos sucios, como se nos fueron otorgados, posteriormente realizamos un diccionario de las variables en los conjuntos de datos y en este punto determinamos el tipo de variable que debía de ser cada una, realizamos el código para conocer las variables actuales y detectamos los errores, finalmente mediante la aplicación de los conocimientos adquiridos durante nuestra clase realizamos el código pertinente para realizar los cambios de cada una de las variables que no estaban correctas con la ayuda de RStudio.

Al realizar nuestra limpieza de datos pudimos encontrar algunos problemas de calidad de los datos y variables con errores, durante esta actividad pudimos corregir dichas falta, entre las cuales se encontraron:

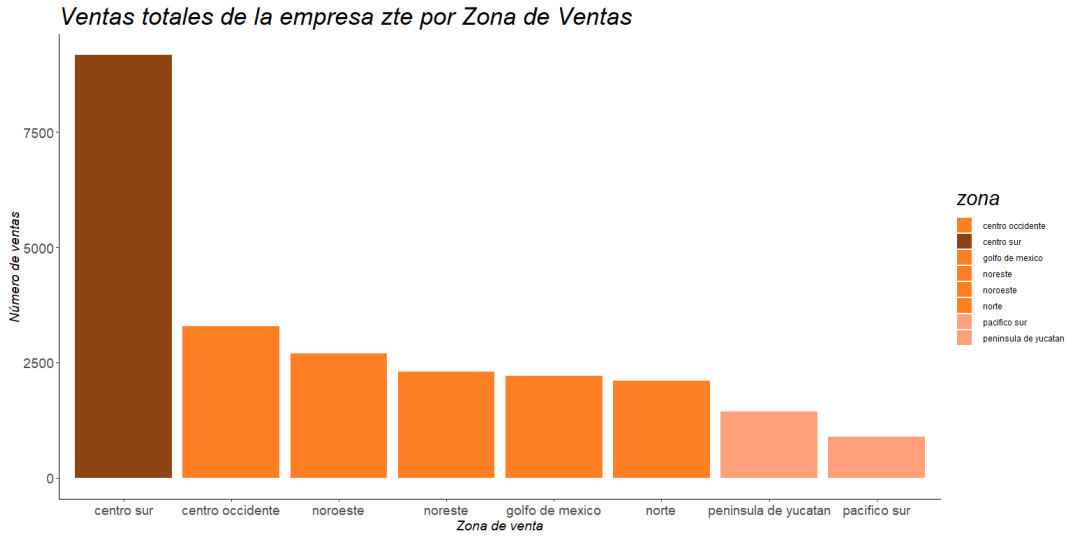
- En la primer variable de puntos de venta se encuentran problemas de que algunos puntos son escritos en mayúsculas y en otras en minúsculas (hablando del mismo punto), también problemas de que una palabra omite una letra (por ejemplo azul sin a) y puntos de venta escritos con espacio y sin espacio (nuevamente hablando del mismo punto). Se identificaron utilizando la función df %>% select() %>% unique y se corrigieron utilizando la función str_replace().
- En el caso de la variable de mes hay algunos meses que son escritos en letra y otros escritos en número. Para la variable de anio algunos vienen escritos con los 4 dígitos y otros con los últimos 2. Se utilizaron las mismas funciones para estos errores.
- La variable de marca viene escrita de 6 diferentes formas, en esta caso mayúsculas, minúsculas, con doble t, escrita 2 veces en mayúscula y en minúscula, y escrita con doble Z. Se utilizaron las mismas funciones para estos errores.
- En la variable de zona la zona centro occidente está mal escrita, occidente no tiene i. Se utilizaron las mismas funciones para este error.
- Para la variable de estado hay varios que no son estados como por ejemplo Monterrey que es una ciudad en Nuevo León. Se utilizaron las mismas funciones para estos errores.
- Para las últimas variables de longitud y latitud hay errores de valores fuera de rango. Se identificaron los valores fuera de rango con la función df %>%select() %>% max, df %>%select() %>% max y se corrigieron los valores con str_replace().

Estos son problemas de calidad ya que son datos no homogéneos (puntos de venta, mes, anio marca), faltas de ortografía (marca, zona, puntos de venta) e incongruencia en los datos (estado, latitud y longitud). Fueron corregidos durante nuestra actividad 6 del semestre, la cual anexamos a continuación.

Vea Anexo 3. Actividad (grupal) 6. Limpieza de datos

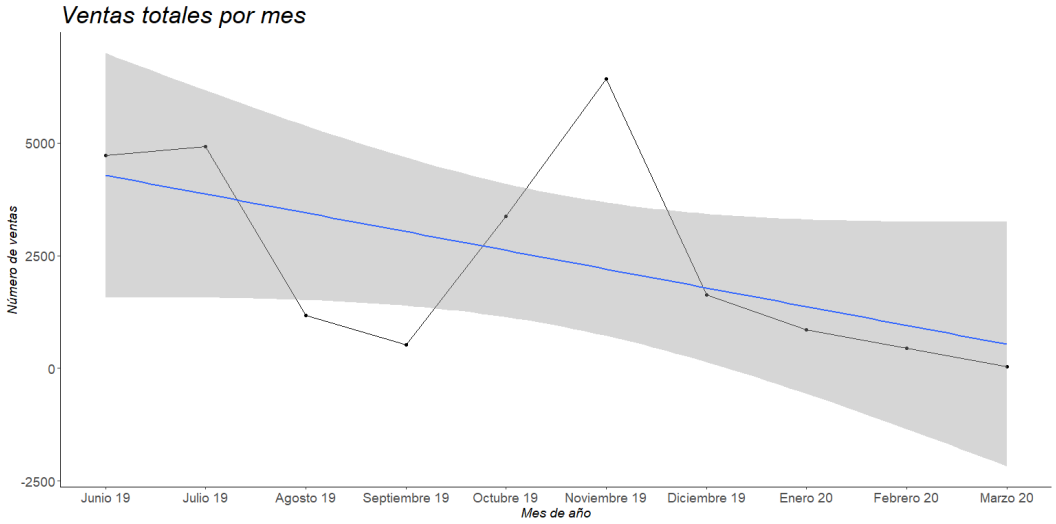
3.2 Realizar un Análisis Exploratorio que permita entender mejor la situación actual.

Para realizar el análisis exploratorio de los datos, el objetivo consiste en comprender el comportamiento regular de los datos a través de su análisis y encontrar si hay relaciones entre ellos. Consideramos que la mejor manera es visualmente a través de las gráficas, pero con una clara interpretación o contexto para comprender mejor nuestros datos, al igual que una conclusión en cada una de ellas con los aprendizajes más valiosos que nos llevamos de ellas, a continuación se presentan.



Gráfica 1. ***Ventas totales de la empresa zte por Zona de Ventas***

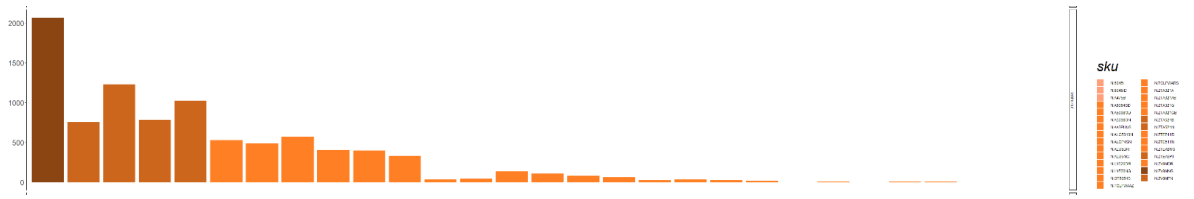
La gráfica nos muestra que claramente la zona que más genera ventas es la del Centro Sur, y que la zona con menos ventas es las Pacifico Sur, también podemos observar que las zonas del Norte, Golfo de México y el Noreste tienen una ligera variación pero se encuentran en el mismo rango, también nos podemos percatar que hay una gran diferencia en el total de ventas de las zonas, se observa que hay un incremento que es el más notorio en el Centro Sur, esto nos ayuda a poner atención y realizar un plan de acción para que las ventas aumenten en las demás zonas.



Gráfica 2. ***Ventas totales por mes de la empresa zte, de junio 2019 a marzo 2020***

Se puede deducir que hay una tendencia negativa a lo largo del tiempo marcada por la recta azul, se puede observar que conforme han pasado los meses el número de ventas ha caído por el punto inicial en junio de 2019 y el punto final en marzo de 2020, podemos observar un crecimiento muy precipitado desde el mes de Septiembre hasta el mes de Noviembre del 2019, se recomienda analizar que medidas se tomaron para conocer lo que beneficia a las ventas de la empresa, por otro lado luego de Noviembre las ventas han caído y han permanecido en caída desde ese mes, se recomienda hacer un plan de acción para mejorar o incentivar las ventas.

Ventas totales de la empresa zte por SKU y zona de ventas



Extracto de la Gráfica 3. Ventas totales de la empresa zte por SKU y Zona de Ventas.
Para consultar la gráfica completa, de favor recurrir a los anexos del presente documento

El SKU recoge información para identificar cada producto en función de su color, precio, marca, talla, tamaño, fabricante por lo que detectamos que esta variable en especial es de suma importancia en el desarrollo del proyecto. La combinación y el orden de las letras y números depende de la prioridad que necesite cada vendedor, siempre determinada por las necesidades del comprador y el objetivo de la empresa, con esta grafica podemos observar que es lo que vende más la empresa, en este caso el bestseller es el N.ZV8MPN en 7 de las 8 zonas de venta marcado en el color más oscuro y seguido por el SKU NZTA521N, y el que ha registrado menos ventas ha sido el N.A475B, marcado en el color más claro pero casi imperceptible por sus pocas ventas, esta gráfica es de gran ayuda ya que nos indica que es lo que más utilidades nos genera y también lo que representa pérdidas de una forma más visual y nos ordena los datos de manera descendente para identificar fácilmente nuestros focos rojos.

3.3 Seleccionar y construir (ingeniería de características) variables para la etapa de modelado

En la etapa 2 cuando se realizó la limpieza de los datos, contábamos con un Dataset “Datos limpios” de **24,089 registros y 14 variables o columnas**, con la ingeniería de características debemos ver cuales son las variables de mayor importancia para nuestro proyecto y crear las variables adicionales necesarias para poder proseguir con la siguiente etapa la cual consiste en el modelado de los datos.

El primer paso que se tomó, fue cambiar la asignación de 3 variables para facilitar el manejo de los datos, que la variable fecha tuviera una asignación de “Date”, y que las variables mes y año tuvieran una asignación de “Numeric” para facilitar el uso de ellos específicamente en este ejercicio. Posteriormente, procedimos a crear índices para las 3 variables más importantes de los datos, ya que de esta manera podemos ordenarlas y manejarlas de una mejor manera, las cuales son Punto de venta, mes y SKU. En el caso de la variable que nos indica el mes cuando se realizó la compra, nosotros no contamos con más de un año de registros de ventas, contamos solamente con 10 meses, pero estos 10 meses están divididos entre mediados del 2019 y mediados del 2020, por lo que ordenarlos e identificarlos con números resulta mucho más sencillo, ya que evita confusiones con cual mes va antes que otro si no tomamos en cuenta el año.

Las variables que no nos proporcionan información relevante al proyecto son: Número de ventas, esta variable sirve más como un contador ya que todas las ventas son de 1 unidad, por lo que sabemos que cada registro corresponde a la venta de un equipo móvil y no necesitamos revisar el dato que viene en esta variable, debido a que ya lo conocemos. La variable marca tampoco nos proporciona información útil ya que todos los equipos pertenecen a la misma marca “zte”. En el caso de la variable gamma, todos los dispositivos vendidos corresponden a la gamma baja y aprovechamos para revisar la variable Costo promedio que tampoco nos da información extra, ya que este costo promedio va en relación con la gamma baja. Las variables zona, estado y ciudad si nos dan información importante sobre las ventas de los celulares, pero como en este proyecto buscamos pronosticar por punto de venta, están demás las variables ya que estas corresponden al Punto de venta. Finalmente para el caso de las variables de latitud y longitud, estas van de acuerdo al Punto de venta también, por lo que sólo son informativas.

Las columnas de índices las incluimos al conjunto de datos para poder manejar de mejor forma el conjunto con la función “Left join”, que consiste en unir data frames entre ellos, une todas las filas del primer data frame (conjunto de datos limpios de la marca "zte) con los valores correspondientes del segundo (índices que creamos de mes, punto de venta y sku). Esto nos ayudó para posteriormente poder filtrar por punto de venta, sku y mes y de tal forma poder obtener las ventas totales de cada filtro, al realizar esto se puede observar que la serie de tiempo no está completa, ya que para las combinaciones que no tienen puntos de venta no están incluidas en esta, así que para completar esta serie utilizamos la función de merge(), esta función se usa para crear datasets con combinaciones, así que ahora para esos datos faltantes se coloca automáticamente un NA, que posteriormente cambiamos por un 0 para poder manejar la variable como un valor cuantitativo.

Posteriormente los datos del noveno mes, de Marzo de 2020 son eliminados, ya que no son datos que se usarán para el pronóstico, ya que se utilizan datos de meses anteriores, el único utilizado es el de ventas pero ya ven incluidas en el mes anterior ya que está la columna de ventas del siguiente mes, así que no son necesarios todos los datos de ese mes.

Ya que tenemos que contestar la pregunta de ¿Cuántas unidades de cada producto de mi marca, se van a vender en todos los puntos de venta al siguiente mes de registro?, generamos las siguientes variables de conteos para poder responder a esta pregunta al momento de realizar los pronósticos, estas variables son: ventas_totales, ventas_totales_tienda_y_mes, ventas_promedio_tienda_y_mes, ventas_totales_tienda_y_sku y ventas_promedio_tienda_y_sku, estas variables se generaron por medio de filtros.

Para poder generar los pronósticos agregamos las columnas de 1, 2 y 3 meses pasados para las variables de ventas_totales, ventas_totales_tienda_y_mes, ventas_promedio_tienda_y_mes, ventas_totales_tienda_y_sku y ventas_promedio_tienda_y_sku, esto utilizando la función de lag, esta función genera un desfase así que el primer valor se copia 1, 2 o 3 veces dependiendo de la cantidad de meses pasados para poder generar este desfase.

Al finalizar nuestro proceso de Ingeniería de características, concluimos un Dataset al cual denominamos “Datos completos”, el cual cuenta con **380,277 registros y 24 variables** o columnas, esto aumentó inmensamente nuestra cantidad de datos en comparación con los Datos limpios de la etapa 2; en más de 15 veces la cantidad de registros por todas las ventas faltantes por mes, punto de venta y SKU, y en casi el doble la cantidad de variables, por la eliminación de variables que nos proporcionaban información relevante y la creación de nuevas variables que nos ayudarán a realizar un buen pronóstico como lo son: los índices, las ventas totales, las ventas totales del siguiente mes, por tienda, SKU y mes.

Etapa 4: Modelado

4.1 Construir (codificar) el modelo de promedios móviles, cambiando 2 veces el periodo móvil.

Los promedios móviles son métodos utilizados para realizar pronósticos de aquellas series de tiempo que no tienen tendencia, únicamente se tiene una variación natural (una variación esperada), en este caso es muy útil ya que en la mayoría de los meses no se tiene una tendencia de ventas y entre ellos hay una variación normal, otro modelo que se pudo utilizar era el de suavización simple, pero y este es un poco más complicado porque es necesario contar con ponderaciones ya que las observaciones más recientes son las que tendrían mayor peso, así que para resolver la pregunta principal del proyecto de ¿Cuántas unidades de cada producto de mi marca, se van a vender en todos los puntos de venta, al siguiente mes de registro? es suficiente con el método de promedios móviles

Un promedio móvil simple se obtiene encontrando el promedio de un conjunto específico de datos y utilizándolo después para pronosticar el siguiente valor.

Para realizar el modelo en python se utilizamos los datos previamente obtenidos de la etapa 3.3, para realizar los cálculos de los promedio móviles simples únicamente necesitamos los datos de 'pdv_id', 'mes_id', 'sku_id', 'ventas_totales', 'y_ventas_siguiente_mes', así que realizamos un filtro para solo trabajar con los datos de estas columnas.

Construimos los modelos de promedios móviles simples para 1, 2 y 3 meses, para un mes únicamente colocamos el dato del mes anterior, para los promedios de 2 y 3 meses utilizamos la siguiente línea de código:

“datos_ma['m2_promedio_de_dos_meses_anteriores'] =
datos_ma.groupby(['pdv_id','sku_id']).rolling(2)['ventas_totales'].mean().reset_index(drop=True)”

para 3 meses el único cambio fue en la función de rolling, en la cual cambiamos 2 por 3.

Para poder comparar cual de los 3 promedios móviles es el mejor es necesario compararlos con indicadores de desempeño, en este caso utilizaremos el MAE (Mean Absolute Error), para esto lo primero que hicimos fue dividir el conjunto de datos por meses, para esto utilizamos la siguiente línea de código (utilizaremos de ejemplo la del mes de julio):

“error_julio= datos_ma[datos_ma.mes_id == 0]

error_julio.head(5)”

Si bien el mes de junio es el mes con id igual a 0, lo colocamos como el error de julio ya que en esas líneas son en las que vienen los pronósticos, realizamos esto para cada mes.

Posteriormente utilizamos la siguiente línea de código para ya calcular el MAE:

“error_m1_julio = mean_absolute_error(error_julio['y_ventas_siguiente_mes'],
error_julio['m1_pedir_lo_del_mes_pasado'])

error_m2_julio = None

error_m3_julio = None”

Recordemos que la fórmula de MAE es:

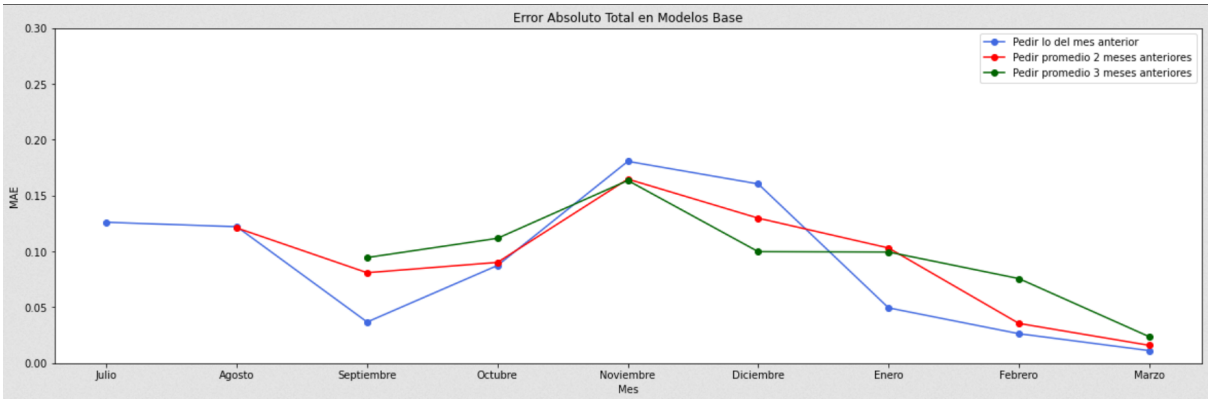
$$MAE = \frac{\sum \left| Y_t - \hat{Y}_t \right|}{n} = \frac{\sum |e_t|}{n}$$

Imágen 1. Fórmula del Error Absoluto Medio

Así que con la función utilizada ya realizamos la fórmula siendo y_ventas_siguiente_mes= Yt y m1_pedir_lo_del_mes_pasado/m2_promedio_de_dos_meses_anteriores/m3_promedio_de_tres_mes es_anteriore= Y^t, para los primeros 2 meses existe una diferencia ya que para julio no contamos con 2 ni 3 mese previos así que no se realizó ese promedio movil y tampoco se calcula ese error, para agosto no contamos con 3 meses previos y de igual maner no se realizó ese promedio movil y tampoco se calculó ese error.

Con los errores de cada mes creamos un data frame en el que aparecieran las columnas de 'Mes', 'mae_pedir_anterior','mae_promedio_2_meses_anteriores' y 'mae_promedio_3_meses_anteriores'.

Finalmente realizamos una gráfica de líneas con los errores de cada mes con respecto a cada promedio móvil y obtuvimos los siguientes resultados:



Gráfica 4. ***Error absoluto Total en Modelos Base***

Por lo que podemos observar en la gráfica, concluimos que el mejor modelo podría ser el promedio móvil de un mes, es decir la opción de pedir lo del mes anterior, ya que aunque en el mes de Noviembre y Diciembre tiene los errores más altos, en el resto de los meses tiene los errores más pequeños así que creemos que ese sería el mejor modelo a utilizar, aunque creemos que el modelo de 2 meses anteriores tambien podria ser util ya que tambien tiene errores pequeños y se ve más estable.

4.2 Construir (codificar) el modelo de aprendizaje de máquina.

En esta etapa elaboramos un modelo de aprendizaje de máquina en el que usamos métodos computacionales (programación en python & r) para aprender de los datos (ventas totales), generar reglas (pronóstico de ventas por tienda) y así mejorar el desempeño (contra métodos tradicionales - promedios móviles). Dentro de los usos del aprendizaje de máquina, usamos el de estimar, ya que estamos prediciendo la demanda de celulares para el siguiente mes, de una manera más eficiente, rápida y eliminando el posible error humano.

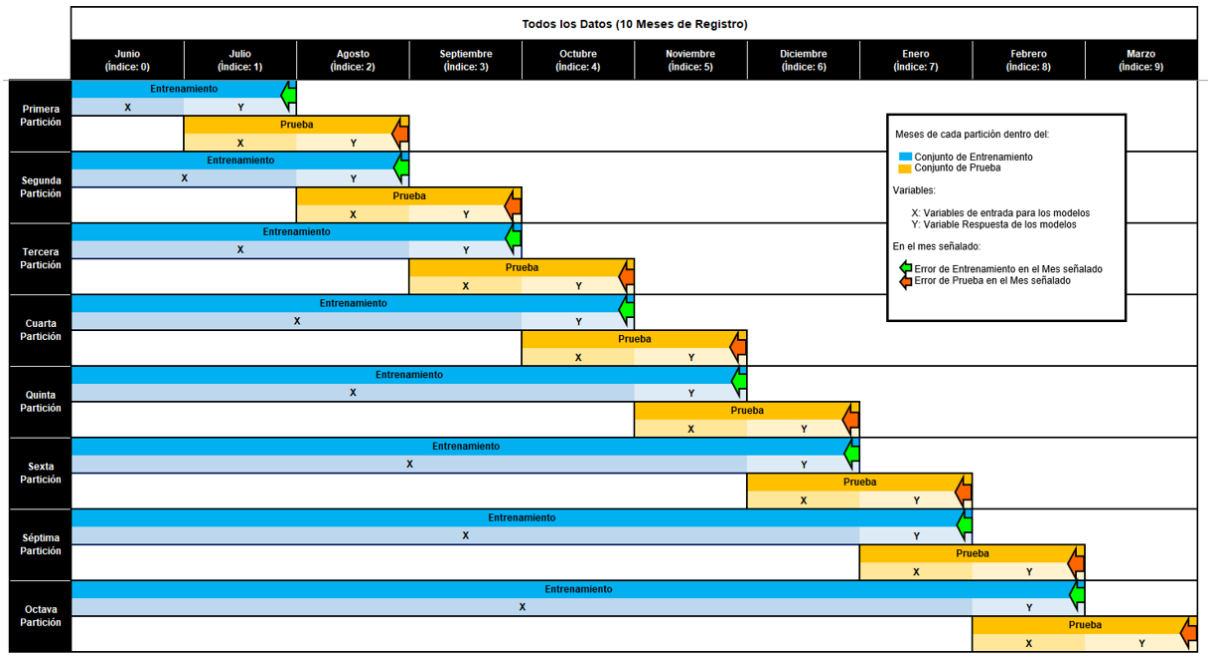
Nos encontramos con un problema de aprendizaje supervisado ya que queremos conocer la variable de salida “y”(ventas siguiente mes por tienda) con base en varias variables de entrada “x” (ventas totales) y al ser la variable “y” del tipo numérico se traduce también en un problema de regresión. Dividiremos los datos en 8 conjuntos de entrenamiento que nos ayudarán a aprender de los datos, no memorizar, y 8 conjuntos de prueba para probar nuestro modelo a datos no antes vistos. También usaremos una validación cruzada de tiempo para evitar el sobreajuste, tomando en cuenta una secuencia periódica respecto al tiempo.

El modelo de máquina que seleccionamos fue el de árbol de decisión, para predecir el valor de una variable en un problema de aprendizaje supervisado de regresión. Este modelo consiste en una serie de toma de decisiones en forma de árbol, en donde la raíz o primer nodo produce la primera división en función de la variable más importante, las ramas o nodos intermedios representan las soluciones y las hojas o nodos finales son las predicciones que buscábamos. Estos árboles pueden tener diferentes profundidades que nosotros determinamos y probamos que van en función del número máximo de nodos por rama.

Este modelo fue considerado ya que nos permite seleccionar la variables más importantes, no es necesario que los datos cumplan con los supuestos de linealidad, normalidad de los residuos u homogeneidad de las varianzas y porque son más simples de crear e interpretar en comparación con otros modelos. Sin embargo, hay que tener en cuenta ciertas restricciones de este modelo como

evitar el sobreajuste que tiende a realizarse, realizar una muy buena preparación de los datos ya que este modelo se ve muy influenciado por los outliers y tener en cuenta que no se deben de crear árboles muy complejos ya que puede que los nuevos datos no se adapten correctamente.

Para realizar el proceso de modelado de aprendizaje de máquina utilizamos phyton igual que en la etapa anterior, empezamos leyendo los datos previamente preparados, estos datos los obtenemos del archivo de datos_completos.csv que preparamos en la etapa 3.3, una breve visualización de sus dimensiones como renglones, columnas y verificación de las variables. Posteriormente procedimos a realizar la preparación de los datos con una validación cruzada de tiempo con 8 particiones para los conjuntos de entrenamiento y prueba de acuerdo a la cantidad de de datos que tenemos y por cada una de estas particiones se llevará a cabo un modelo.



Imágen 2. Modelo de validación cruzada de tiempo para los 10 meses de registro con 8 particiones (Masetto, 2021)

Para realizar la primera participación, designamos los datos del primer mes con información, junio, como “x1” de entrenamiento para estimar el siguiente mes, julio, como “y1” de entrenamiento. De igual manera designamos los datos de julio como “x1” de prueba, para estimar el siguiente mes agosto como “y1” de prueba. Seguido de esto creamos índices con la función set_index, tanto para el conjunto de entrenamiento como el de prueba para generar el mismo número de particiones que de puntos de venta. Finalmente dividimos estos conjuntos en dos secciones, las variables de entrada “x” y las variables de salida “y” utilizando los comandos de igualdades y desigualdades. Repetimos este procedimiento para las siguientes 7 particiones, recorriendo en una posición los meses y tomando en cuenta que la “x” de entrenamiento contiene la información de los meses anteriores también; como se muestra en la imagen 2.

En la siguiente parte realizamos el modelado con el conjunto de entrenamiento apoyándonos de la paquetería “numpy” & “sklearn” con la función “tree” que nos ayudó con la creación de los árboles de decisión y finalmente de “sklearn.metrics” la función de mean_absolute_error (MAE) para calcular los errores de una manera más sencilla, ya que estos errores nos sirven como guía de resultados de desempeño y así poder comparar entre modelos para optar por el mejor de ellos. En la primera partición definimos que queríamos que el árbol tuviera una profundidad de 1, le dimos las variables para que entrenara el modelo y le pedimos que nos diera una predicción en base a estas variables, como los resultados vienen en decimales y los celulares son unidades enteras, también utilizamos

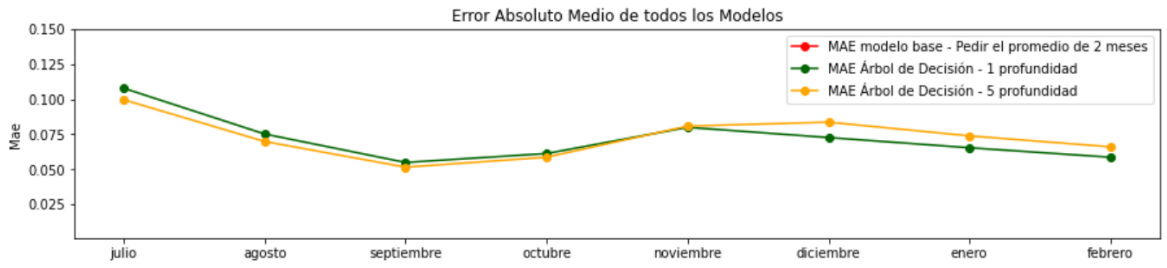
una función para que redondeara estas predicciones y guardamos los valores de estas predicciones en una nueva columna. Finalmente calculamos el error absoluto medio entre la predicción que nos otorgó el árbol y el valor real de las ventas. Repetimos este proceso para cada una de las ocho particiones.

Seguido del paso anterior, procedimos a probar el modelo con el conjunto de prueba, con la primera participación le pedimos que predijera y1 a partir de la variable de entrada x1 del conjunto de prueba, redondeamos estos valores a unidades enteras y los agregamos en una nueva columna, también realizamos la suma de todas las predicciones, para observar cuántas unidades en total se venderían en el mes y calculamos el error absoluto medio de la misma manera que con el conjunto de entrenamiento. Repetimos este proceso para las 8 particiones.

Optamos por probar el árbol de decisión con una diferente profundidad para poder comprar entre estos modelos y encontrar la mejor solución, por lo que repetimos todo el proceso desde de la etapa de modelado, cambiando solamente la profundidad del árbol de 1 a 5. Para poder comparar estos modelos, incluyendo el de los promedio móviles, creamos un Excel donde registramos manualmente los errores absolutos medios por mes (nuestro resultado de desempeño), obtenidos por el modelo de promedios móviles, por el conjunto de entrenamiento y prueba de árbol de decisión con profundidad de 1 y por el árbol de decisión con profundidad de 5. Con ayuda de Python leímos este archivo de excel para que todo estuviera correcto, filtramos los errores por conjunto de entrenamiento y por conjunto de prueba y realizamos una gráfica para analizar cada conjunto, las cuales se discutirán en la siguiente etapa de evaluación.

Etapa 5: Evaluación

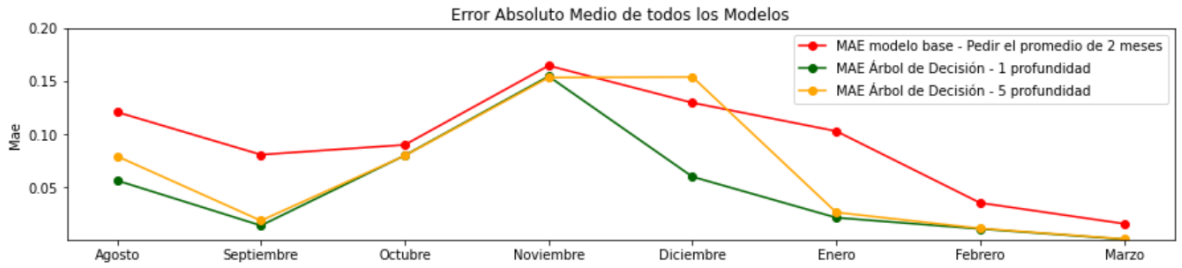
Para la evaluación de los modelos de entrenamiento únicamente se pueden observar los errores de los árboles de decisión ya que para el modelo de promedios móviles simples no fue necesario un conjunto de entrenamiento para ser realizado.



Gráfica 5. ***Error absoluto medio todos los modelos de entrenamiento***

Con la primera gráfica correspondiente al error absoluto medio del conjunto de entrenamiento, nos es muy difícil decidir entre el árbol de decisión de 1 y 5 profundidad, porque en los primeros 4 meses el de profundidad 5 tiene el menor error mientras que el de profundidad 1 tiene el menor error en los siguientes 4 meses.

Por estas razones basaremos nuestra elección en la segunda gráfica que corresponde al error absoluto medio del conjunto de prueba, así mismo es la gráfica que nos muestra el comportamiento que necesitamos, ya que es el comportamiento del modelo cuando es sometido a nuevos datos.



Gráfica 6. Error absoluto medio todos los modelos de prueba

En esta nos es mucho más claro visualizar que el modelo con el menor error absoluto es la de árbol de decisión de profundidad 1, mientras que la que tiene el mayor error absoluto es la de Moving average / modelos tradicionales.

Así que el mejor modelo para predecir la demanda del siguiente mes sería un modelo árbol de decisión con profundidad 1 ya que ofrece las mejores predicciones, esto evaluando con el indicador de desempeño MAE, por lo que es el que le recomendamos a la empresa.

Así que con los datos de la predicción de prueba podemos responder a la pregunta inicial de **¿Cuántas unidades de cada producto de mi marca, se van a vender en todos los puntos de venta, al siguiente mes de registro?**

Para eso utilizamos la siguiente línea de código:

```
#Exportar df excel

prueba_8.to_excel('prueba_8.xlsx')
```

Descargamos los datos de prueba 8 y la respuesta a nuestra pregunta de cuántas unidades se van a vender en el mes de Abril (mes 9) viene en la columna de ventas_por_mes_pred

Así que en el mes de Abril podemos predecir que se van a vender las siguientes unidades:

Punto de Venta	mes	sku	ventas_por_mes_
bca aristeum ermita	Abril	N.LX520DR	1
fgt san esteban naucalpan	Abril	N.ZTEA6PT	1
tda cdmx ecatepec las americas	Abril	N.ZTEA6PT	1
tda cdmx guadalupe fortaleza	Abril	N.ZTEA6PT	1
tda cdmx tintoreto	Abril	N.ZV8MNG	1
		N.ZV8MPN	1
tda cordoba arco	Abril	N.OT5054S	1
tda guamuchil rosales	Abril	N.OT5054S	1
tda los mochis plaza sendero	Abril	N.ZTA521N	1
tda monterrey plaza cumbres	Abril	N.ZV8MPN	1
tda nuevo laredo	Abril	N.A5098ON	1
tda san cristobal de las casas ii	Abril	N.ALC5010N	1
tda tepic los fresnos	Abril	N.ZTA521N	1
		N.ZTA521B	1
tda valle de bravo	Abril	N.ZTA321G	1
	Abril	N.ZTA521B	1
tda veracruz	Abril	N.A5098ON	1
walmart miramontes	Abril	N.LX520DR	1
		N.LX520NG	1
		Total	19

Tabla 1. Predicción de ventas en el mes de Abril

En los puntos de venta y modelos que no se muestran no hay ventas pronosticadas, en el Anexo 5 se podrá ver la información de todos los puntos de venta y todos los modelos.

Conclusiones y recomendaciones a la empresa

La noticias para la empresa “Zte” no son sencillas de entregar, debido a la baja demanda que se tiene pronosticada para el mes de abril con 19 unidades en total de todos sus modelos en todos sus puntos de venta, sin embargo revisando nuevamente los registros proporcionados, este pronóstico no se encuentra alejado de la realidad, como ejemplo podemos presentar el mes antecesor al pronóstico donde en marzo sólo se tuvieron 38 ventas en total y la gráfica 2 presentada en la etapa 3.2, donde claramente se ve una tendencia negativa en los últimos 4 meses, por lo que es correcto que el modelo haya demostrado tal baja debido a esta tendencia.

ANEXOS

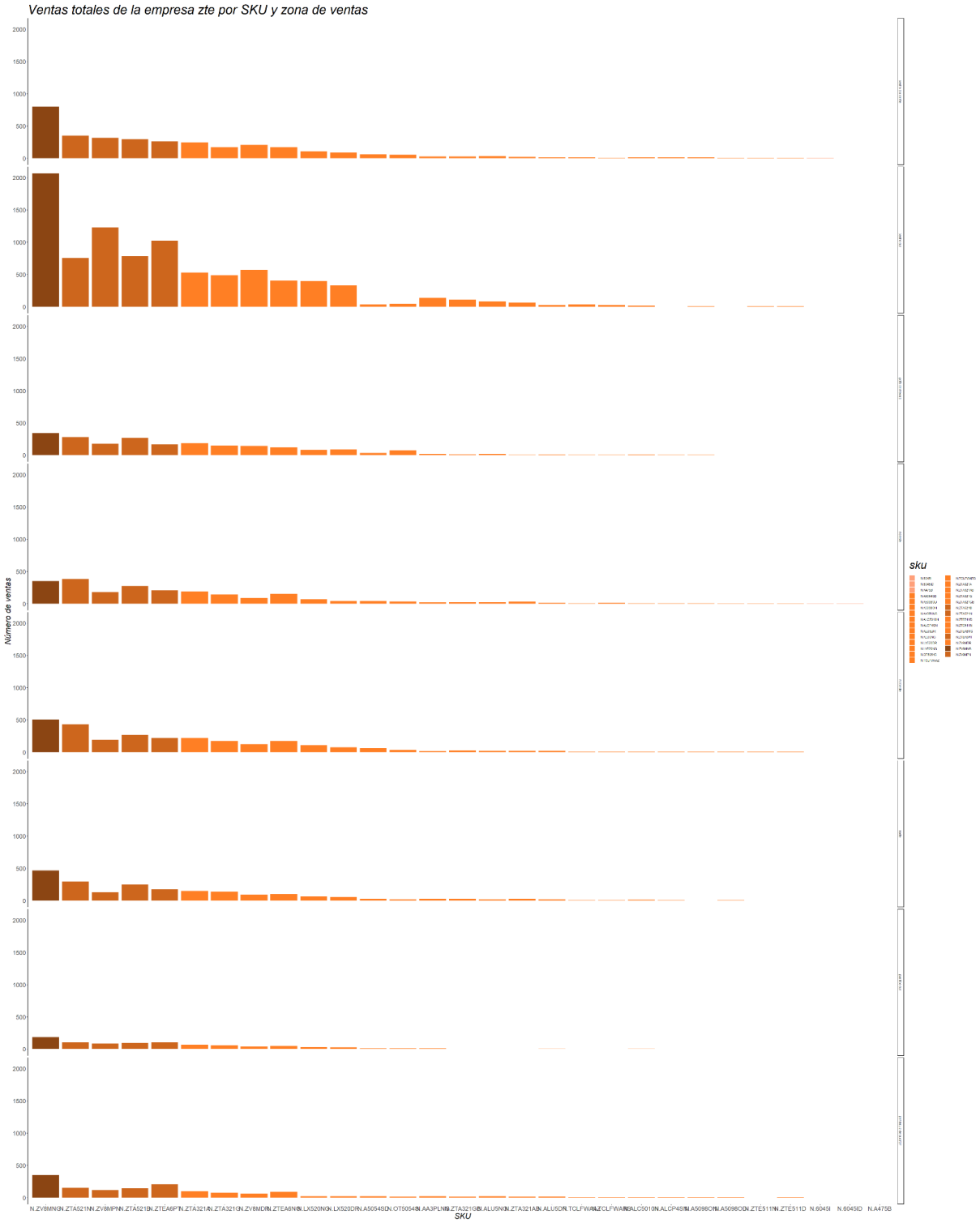
Anexo No.1 - “Desarrollo Preliminar del Proyecto” - Documento PDF. - Relacionado con 1.5.1

Anexo No.2 - “Diagrama de Gantt del Desarrollo del Proyecto” - Documento Excel - Relacion 1.5.2

Anexo No.3 - Limpieza de datos

file:///D:/MIS%20DOCUMENTOS/Laboratorio%20de%20dise%C3%B1o%20y%20optimizaci%C3%B3n%20de%20operaciones/1er%20parcial/Actividad_6_Equipo02_-Limpieza_de_Datos.html

Anexo No. 4 - Gráfica 3. Ventas totales de la empresa “zte” por SKU y zona de venta



Anexo No.5 - "Predicción de ventas mes de Abril" - Documento Excel

Referencias:

[1] Alamilla, R. (2021). Mercado de Smartphones 1T-2021: Reconfiguración del Ecosistema en Curso. Septiembre 4, 2021, de CIU Sitio web: https://www.theciu.com/publicaciones-2/2021/6/7/mercado-de-smartphones-1t-2021-reconfiguracin-del-ecosistema-en-curso

[2] Importancia del pronóstico de la demanda en la cadena de suministro - Demand Solutions. (2020, 17 diciembre). Demand Solutions. https://es.demandsolutions.com/resource-center/abstracts/importancia-del-pron%C3%B3stico-de-la-demanda-en-la-cadena-de-suministro/

[3] Martínez, C. (2017). ZTE cuenta con 9% del mercado en México. Septiembre 4, 2021, de El Universal Sitio web: https://www.eluniversal.com.mx/articulo/cartera/negocios/2017/08/11/zte-cuenta-con-9-del-mercado-en-mexico

[4] Masetto, A. L. (2021). Tema 3_Conceptos_Basicos_Aprendizaje_de_Maquina_Parte_1 [Diapositivas]. Tecnológico de Monterrey. https://experiencia21.tec.mx/courses/189139/files/63779041?module_item_id=10039026

[5] Merayo, P. (2020, mayo). Qué son los árboles de decisión y para qué sirven. Máxima Formación. https://www.maximaformacion.es/blog-dat/que-son-los-arboles-de-decision-y-para-que-sirven/

[6] N.d. (2014). La estrategia de ZTE para conquistar el mercado de smartphones. Septiembre 4, 2021, de Forbes Sitio web: https://www.forbes.com.mx/la-estrategia-de-zte-para-conquistar-el-mercado-de-smartphones/

[7] N.d. Móviles ZTE: características y modelos ¿Son realmente buenos? . Septiembre 4, 2021, de euronics Sitio web: https://www.euronics.es/blog/moviles-zte-caracteristicas-y-modelos-son-realmente-buenos/