Instituto Tecnológico y de Estudios Superiores de Monterrey

Campus Toluca

Reporte Final

Profesora: M.C. Ana Luisa Masetto Herrera

Integrantes Equipo #4

Víctor Sánchez Arroyo	A01364835
Karla Guadalupe Solórzano Martínez	A01368196
Diego Carlos Garcia Gonzalez	A01368191
Andrés Benavides Domínguez	A01368269
Andrea Romo Hernández	A01367773

Fecha de entrega: viernes 19 de noviembre de 2021

Presentación

https://www.canva.com/design/DAEwARBHrGI/share/preview?token=VyC6vdZmb0ZtjGaGXbnGUQ&role=EDITOR&utm_content=DAEwARBHrGI&utm_campaign=designshare&utm_medium=link&utm_source=sharebutton

Introducción

Durante el semestre, se ha estado analizando un conjunto de datos correspondientes a la marca de celulares Motorola, una marca ampliamente reconocida en la actualidad. Al estar aprendiendo y aplicando las normas de la ciencia de datos, no solamente hemos podido observar estos datos sino acondicionarlos para que lo que una vez era simplemente una enorme tabla con números y variables ahora pueda darnos información que nosotros podemos convertir en conocimiento.

Esto es, a final de cuentas, lo que permite la ciencia de datos. La primera tarea que debía completarse era comprender el contexto de la marca en la actualidad para así obtener un panorama de qué nos dirían los datos eventualmente. Junto con la comprensión de la situación, fue importante definir qué queríamos obtener con el proyecto, establecer objetivos para guiarnos en el proceso, y organizar la manera en que íbamos a llevar a cabo cada etapa.

Utilizando las variables que conforman el conjunto inicial de datos, entre ellas el SKU, punto de venta, fecha de la venta, las dos anteriores divididas a su vez en más variables para mayor comprensión, y luego de una extensa limpieza de datos que se desarrollará más a detalle en el contenido del reporte, se pudo analizar los datos para identificar sus características al igual que algunos patrones no muy obvios. La limpieza de datos dio lugar a un análisis exploratorio de los datos y a la ingeniería de características para finalmente llegar a la etapa de modelado, en la cual se enfocó en el modelo de promedios móviles para identificar patrones en los datos. Los resultados obtenidos con el modelo se desglosan al final del reporte.

La ciencia de datos puede aplicarse en una amplia variedad de ocasiones; es un claro proceso de cómo se pueden convertir datos a información y luego a conocimiento, lo cual no solamente puede ayudar a tener este conocimiento sino también a tomar decisiones, contestar preguntas, prácticamente cualquier propósito que se venga a la mente.

Etapa 1: Comprensión del Negocio

1) Descripción de la situación actual

situación a la que se enfrenta con relación a la marca

Actualmente, el panorama para los teléfonos inteligentes en México, a pesar de la crisis económica que generó la pandemia, no es malo, pues en el país ya existen

más líneas móviles que población total. Según The Competitive Intelligence Unit (CIU) existen 126,014,528 líneas, mientras que el número de personas es 126,014,024, de acuerdo con el Inegi.Según el análisis del cuarto trimestre del 2020 realizado por la consultora CIU , el estudio elaborado por la firma especializada en telecomunicaciones, a lo largo del país las marcas que tienen más presencia entre la población son Samsung, con el 32.2%, Motorola, con el 20.3%, Huawei, con el 14.2%, y Apple, con el 10.4%.

Motorola ha ido a la alza, pues en el periodo anterior, sus cifras eran de 18.7%. El principal mercado para esta marca son los usuarios enfocados en teléfonos de gama media y por primera vez superó el 20% de participación, además de que según los analistas de The CIU, podría ir estableciendo el terreno para competir con Samsung a mediano plazo.

Por qué es importante realizar este tipo de proyectos, y cómo aplicar herramientas y conceptos de Ciencia de Datos a un problema que enfrenta un ingeniero industrial

La ciencia de datos es un tipo de procesos y algoritmos para extraer patrones no obvios y útiles de grandes conjuntos de datos. Su importancia en estos proyectos es que permite tomar decisiones al igual que permite localizar ventanas de oportunidad y así poderse adelantar a lo que sus consumidores necesitan así como sus tendencias de consumo, por lo que este tipo de aportaciones y el análisis que te permite realizar brindan gran ayuda a los problemas a los que se podría enfrentar un ingeniero industrial.

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

Describir la problemática a la que se enfrenta el equipo en terminología de negocio en el contexto de tu carrera profesional.

Una de las empresas que ha tenido un alza en sus ventas ha sido Motorola, siendo que se encuentra en el segundo puesto como la segunda marca con mayor presencia en el país. Es por eso que las compañía en sus diferentes puntos de venta enfrenta un reto que es el de pronosticar el número de ventas de cada producto, ya que de no hacerse apropiadamente, podría generar diversos problemas para la compañía , como lo son costos logísticos excesivos o insatisfacción de clientes al no contar con los productos y servicios necesarios para satisfacer las necesidades del mercado.

Comprender y explicar por qué la tarea asignada es importante para la compañía

Desarrollar un proyecto de ciencia de datos enfocado a resolver un problema de construcción de portafolios de productos y con el fin de predecir la demanda de la marca Motorola para los diferentes puntos de venta de la empresa. Es importante la tarea asignada, ya que se obtiene valiosa información que se usa para tomar decisiones en la empresa. Puede utilizarse también el pronóstico de ventas para asegurarse que cuentan con la mercancía necesaria para cubrir las necesidades demandas por la clientela.

3) Entender y describir la problemática

Definir el tipo de problema que se está enfrentando

Para definir el tipo de problema al que se está enfrentando, se hizo una breve observación de los datos sucios. Al identificarse las variables, pudo proponerse cuáles podrían tener una relación. Para detectar si se tenía una relación, se elaboró un diagrama de dispersión (presente en el anexo). Las variables que se consideraron relevantes para esto fueron el costo y el nivel de gamma. El diagrama de dispersión mostró que estas variables no tienen una relación lineal visible, y este fue el caso para las demás variables que se compararon. Por lo tanto, no podríamos tomar una regresión para este problema, sino que estamos viendo un problema de clasificación. Con estos, no se busca predecir valores continuos sino más bien predecir a qué clases pertenecen ciertos conjuntos de datos.

Enunciar la pregunta que van a contestar a lo largo del proyecto.

Como ya se definió que el problema al que nos estamos enfrentando es uno de clasificación, una pregunta que puede contestarse a lo largo del proyecto es la siguiente:

¿Cómo puede acomodarse el nivel de gama de los dispositivos con respecto a las demás variables?

4) Plasmar los objetivos.

Definir los objetivos del proyecto

Con base en la planeación que tenemos actualmente para el proyecto, podemos enunciar los siguientes objetivos:

a) Definir las cuestiones teóricas que se deben comprender antes de comenzar a trabajar con los datos.

b) Verificar la calidad de los datos mediante su limpieza.

c) Preparar los datos para su análisis (EDA)

d) Tener los datos organizados de tal manera que permita seleccionar una técnica de modelado adecuada.

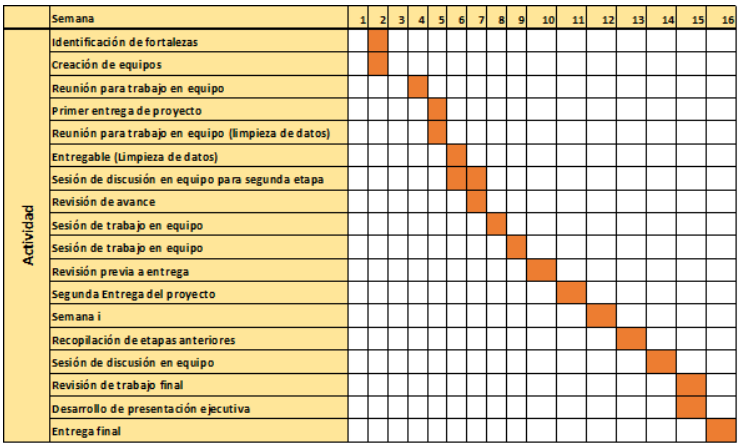
e) Evaluar los resultados que se obtengan y verificar que contesten la pregunta planteada.

f) Planear el despliegue apropiado de lo que encontremos mediante el proyecto.

5) Estructurar el proyecto y hacer un plan preliminar.

Elaborar un Gantt.

El siguiente diagrama de Gantt representa lo mencionado en el punto siguiente:



Realizar y plasmar un plan

Consideraremos la duración del semestre respecto a las fechas establecidas en el curso para hacer la división de las etapas del proyecto, y se buscará enfocarse en cada etapa durante una semana. Existen excepciones a esto, ya que la identificación de fortalezas de los miembros y la creación de los equipos fueron ambas vistas durante la misma semana. Así mismo, tenemos contemplada la sesión de discusión en equipo sobre la segunda etapa para que se lleve a cabo en las semanas 6 y 7 del semestre. Hemos definido las demás actividades a realizar, y estas se llevarán a cabo de manera secuencial hasta la semana 16 del semestre.

Etapa 2 - Comprensión de los datos

Describir los datos crudos.

Los datos se encuentran en un archivo .csv, un formato utilizado para compartir datos entre plataformas (Comma separated values), es considerado como el formato más flexible, los datos se encuentran en 14 columnas:

Punto de venta: El lugar donde se compró el equipo	Sku: Stock Keeping Unit
Fecha: El día de venta del equipo	Marca: La marca del equipo vendido
Mes: El mes de venta del equipo	Gamma: La gama del modelo, baja, media, alta
Anio(año): El año de venta del equipo	Costo Promedio: El costo promedio de la venta
Latitud y Longitud: La ubicación en GPS del punto de venta	Zona: La zona del país donde se encuentra el estado donde se realizó la venta
Estado: El estado donde se realizó la venta	Ciudad: La ciudad donde se realizó la venta

Número de venta: La cantidad de equipos en la venta	

Detectar problemas de calidad

Todas las columnas de datos presentan problemas, principalmente la falta de estandarización; En las fechas los meses se escriben de varias maneras, incluyendo o no el cero antes del número de mes o este escrito con letras en lugar del número de mes, en los años podemos encontrar casos donde estos se escriben como 18 o 2018, en la marca podemos observar como esta se escribe con mayúscula al inicio o sin ella.

En algunos otros casos podemos notar como en el estado quien capturó los datos puso la ciudad y en otros el costo promedio está ausente.

Todos esos problemas deben ser corregidos antes de realizar cualquier análisis ya que pueden causar que el programa que realice la agregación de datos pierda alguno, ya que por la cantidad de datos sería una tarea imposible para un ser humano.

6 El equipo termina el entregable enunciando las actividades en las que tienen que enfocarse para proceder con las siguientes fases del proyecto o áreas de oportunidad que han detectado.

Para llegar a la segunda entrega de este proyecto, será necesario enfocarnos en las actividades de la semana 5 a la semana 11 como está estipulado en nuestro diagrama de Gantt; esto incluye discusión en equipo, revisión de avance, trabajo en equipo y revisión previa a la segunda entrega.

Después de realizar un análisis rápido de los datos se puede observar que los datos contienen múltiples errores como descritos previamente pero estos errores son relativamente sencillos de corregir ya que estos son los mismos, esto tomaría varios días si se hiciese a mano pero con el uso de R se pueden cambiar con los comandos vistos.

Por el momento, se tiene una comprensión sobre lo que nos muestran los datos crudos y sabemos qué es lo que se tiene que limpiar en ellos para que podamos trabajar con estos datos.

Etapa 3: Preparación de los datos

Limpieza de datos

La etapa en que se realizó la limpieza de los datos representa en la gran mayoría de los proyectos el mayor tiempo de dedicación del proyecto, incluso llega a consumir el 80% del tiempo. Este no fue el caso, ya que los datos habían sido previamente limpiados, sin embargo habían recibido un tratamiento para ser ensuciados a propósito. Como equipo de trabajo tuvimos que enfocar bastante dedicación en poder comprender e identificar qué variables eran las que presentaban problemas.

Como se menciona en el párrafo anterior, el poder comprender las variables que estamos trabajando nos permite saber en donde existen posibles errores de captura y cómo es que se podrían arreglar esos errores. Después de comprender las variables, logramos encontrar errores en la variable mes, unificando este tipo de entrada con una variable de tipo numérica. Por otra parte encontramos errores en la variable año, en ese caso no estaba uniforme en términos de 4 dígitos. La variable marca también presentaba errores de captura, estos debieron corregirse para que todos fueran “motorola”. Otro error en las variables se encontró en la zona, un ejemplo de errores en la captura fue “GOLFO DE MEX” el cual fue cambiado por “golfo de méxico”. La variable “estado” fue otra que presentó fallas en las capturas. Las demás variables que tienen incongruencias en las capturas fueron; latitud, longitud, ciudad y punto de venta (de hecho esta presenta bastantes errores en las capturas).

El procedimiento para poder realizar esta limpieza de datos fue el siguiente; primero se leyeron los datos que fueron entregados por la profesora, esto con la ayuda del siguiente comando *datos1 <- read.csv(“equipo_4_motorola_datos_sucios.csv")*

El siguiente es un ejemplo del comando que nos permitió realizar la limpieza de los datos en la variable mes:

datos1$mes <- str_replace(datos1$mes, "JUL", "7")

Para las siguientes variables solo la variable (que en esta caso es mes) y se selecciona en primer el dato que está mal, seguido del nuevo valor que se le asigna.

Actualidad

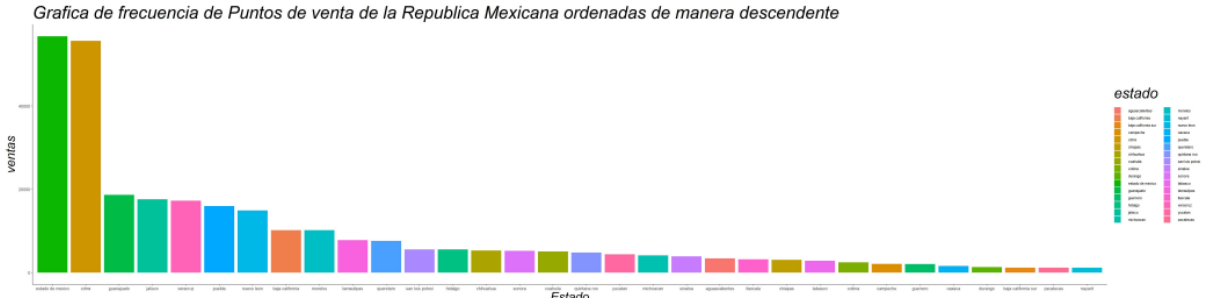
En este momento del proyecto se encuentran los datos totalmente limpios y de esta forma se han podido realizar los siguientes avances del proyecto, los cuales han sido; análisis exploratorio de los datos, ingeniería de estadísticas para la etapa de modelado y etapa de modelado (promedios móviles).

Análisis Exploratorio de los datos

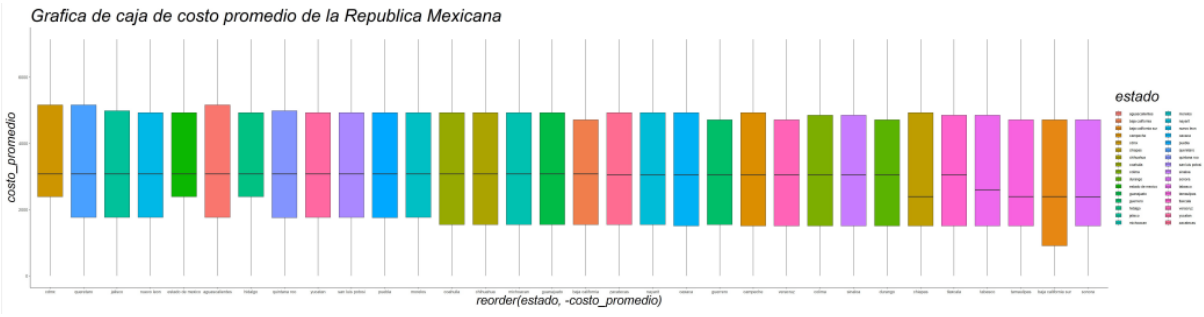
Descripción de las variables

De todas las variables que tenemos, consideramos que las variables punto_de_venta, num_ventas, sku, mes, gamma y costo_promedio son as que representan la mayor importancia para nuestro estudio. Algunos detalles importantes de cada una pueden apreciarse a continuación: punto_de_venta: los datos están constituidos por aproximadamente 1,914 puntos de venta distintos distribuidos alrededor de la República Mexicana, es decir, actualmente hay 1,914 ubicaciones distintas en las cuales se realizan ventas de los productos Motorola dentro de México. Estos puntos de venta pueden organizarse a su vez con ayuda de las variables zona, estado, ciudad, latitud y longitud. num_ventas: ya que se está

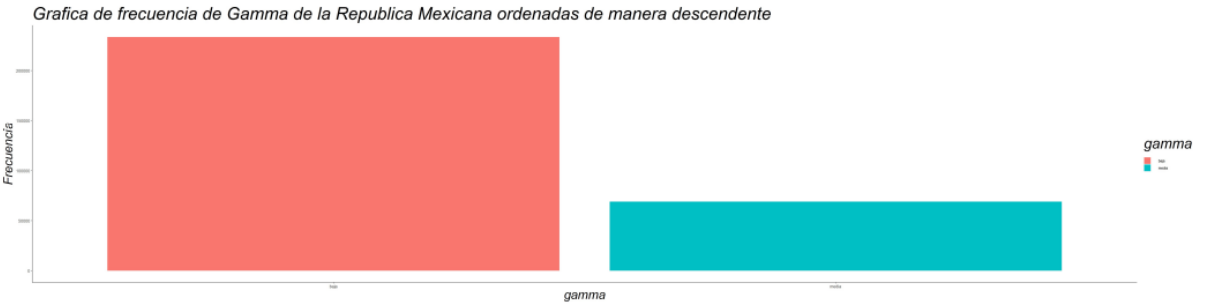
analizando cada fila de forma individual, el valor de esta variable es de 1 para cada registro. A pesar de esto, la variable nos podrá ayudar en nuestros cálculos. Teniendo datos sobre la cantidad de ventas realizadas, es posible obtener información sobre los distintos productos considerados en el estudio y cómo contribuye cada uno a las ganancias de la empresa. Por ejemplo, podemos analizar qué productos venden más, qué productos venden menos, y con base en esto se puede hacer una toma de decisiones de manera informada. sku: contamos con 34 modelos distintos para este estudio. mes: se cuenta con información para los doce meses del año, todos identificados mediante su valor numérico. gamma: se consideran los niveles alto, medio y bajo de gamma. costo_promedio: el costo promedio difiere de acuerdo con el modelo y el nivel de gamma del mismol.



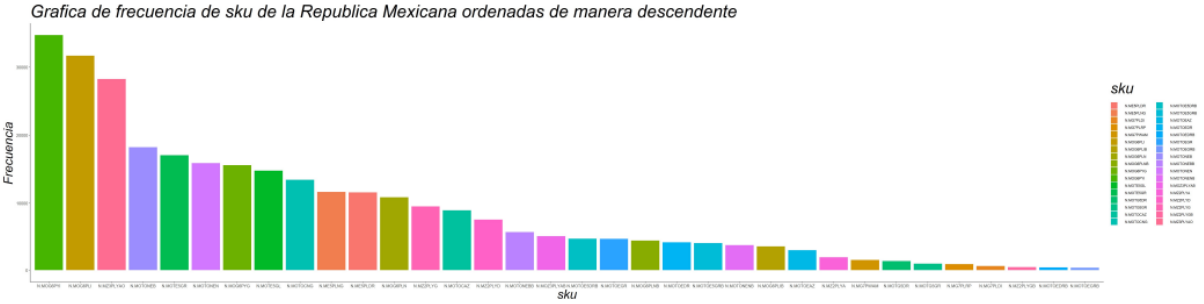
En esta gráfica de frecuencia podemos ver los puntos de venta que existen a través de la república mexicana en cada uno de los estados de forma de manera descendente es decir de los puntos de ventas de mayor a menor y los resultados nos muestran que el Estado de México cuenta con la mayor cantidad de puntos de venta con una cantidad mayor de 40000. Le sigue la CDMX con una cantidad mayor a los 40000 puntos de venta y el tercer lugar en puntos de venta es el estado de Guanajuato con un aproximado de 20000. Los estados con la menor cantidad de puntos de venta de la compañía Motorola son el estado de Nayarit con un aproximado de 2000 y le sigue el estado de Zacatecas con un cantidad similar a 2000. El tercer lugar con la menor cantidad de puntos de ventas es el estado Baja California sur con un cantidad cercana a 2000. Es importante contar con esta información porque nos proporciona una idea sobre los lugares de los principales mercados para la empresa, ya que si encuentra una mayor cantidad de puntos de venta esto quiere decir que existen en esa zona los lugares con los principales mercados y esto con los futuros compradores de los productos.



En esta gráfica podemos observar el promedio que existe en el costo en cada estado y en la mayoría de los estados tiene un costo de 3000 y existen 5 estados que están por debajo de ese número y esos estados son Chiapas,tabasco,tamaulipas,baja california sur y sonora. Es importante esta información porque nos pueden ayudar cuales son los estados que tienen los costos más altos y los mas bajos, asi como si nos mostrara datos que fueran atípicos y por lo tanto, nos ayuda a detectar el promedio que existe de los costos en los estados, el valor maximo y minimos que tienen los costos en cada estado.

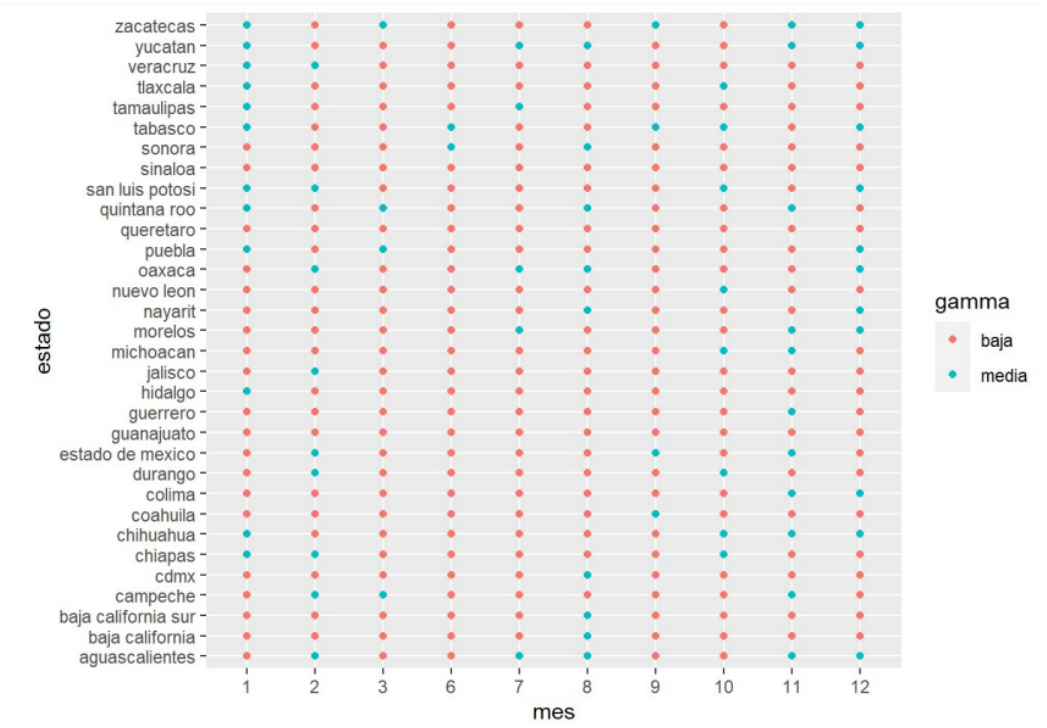


En esta gráfica de frecuencia, podemos observar los niveles de gamma baja y media. Siendo una gráfica relativamente sencilla, notamos que el nivel bajo de la gama constituye la mayor parte de los artículos vendidos en la República Mexicana. A su vez, el nivel medio no representa ni la mitad de la frecuencia del nivel bajo. Esto nos indica que la mayor parte de los artículos que forman parte de este estudio tienen una baja saturación de colores en cuanto a su pantalla y relativamente pocos tienen una saturación media. Esta información puede servir para evaluar la preferencia que tiene el público respecto a la gama de sus teléfonos.



Los distintos rubros de la variable sku se han graficado de forma descendente de tal forma que nos permite apreciar fácilmente qué modelos son más vendidos y qué

modelos son los que menos se venden. En cuanto a los tres modelos que representan las mayores frecuencias en nuestros datos se encuentran NLMOG5PYI, NLMOGP6PLI y el modelo NLMZ3PLYAO. Por lo contrario, los tres modelos que representan la menor frecuencia en nuestros datos son NLMZ2PLYGB, N.MOTOEDRB y N.MOTOEGRB. Esta gráfica puede proporcionarnos información muy valiosa ya que es importante conocer cuáles son los modelos mayormente vendidos en la compañía para poder darles el tratamiento adecuado y de esta forma poder maximizar ventas. Con conocimiento como este, es posible tomar decisiones informadas respecto a las ventas que se realicen.



En esta gráfica podemos observar en qué meses hay más incidencia de gama media o baja. En enero en 12 estados predominó la gama media, siendo el mes con mayor incidencia de esta gama. Por su contraparte en el mes de Junio solo en 2 estados predomina la gama media. También podemos observar que Zacatecas, Tabasco, Aguascalientes y Yucatán son los estados con mayor incidencia de gama media a comparación de Sinaloa, Guanajuato y Querétaro son los estados en los que en ningún mes predominó la gama media. Esta información nos ayuda para detectar en cuales estados predomina la preferencia por dispositivos de cierta gamma para la planeación de ventas, distribución de equipos y en cual seria mejor opción para incorporar equipos de gama alta.

Ingeniería de Características

```r
datos$punto_de_venta <-
as.character(datos$punto_de_venta) #Caracter o factor
datos$fecha <- as.Date(datos$fecha)
datos$mes <- as.numeric(datos$mes) #pueden ser factor /
para usos prácticos de este ejercicio conviene que este
en número
datos$anio <- as.numeric(datos$anio) #pueden ser factor /
para usos prácticos de este ejercicio conviene que este
en número
datos$sku <- as.character(datos$sku) #Caracter o factor
datos$marca <- as.character(datos$marca) #Caracter o
factor
datos$sku <- as.character(datos$sku) #Caracter o factor
```

Para realizar el análisis de características se utilizaron las librerías tidyverse y zoo de la siguiente manera.

En este chunk optimizamos el tipo de los siguientes datos para poder ayudar con su manipulación más adelante, cambiando algunos a tipos que sean más congruentes con la información que proporcionan.

Los siguientes chunks asignan identificadores a los puntos de ventas, fechas y productos individuales, después convirtiéndolos en caracteres para su próxima manipulación.

```r
pdv_id <- datos%>%select(punto_de_venta)%>%unique()%>%arr
ange()
head(pdv_id)
```

Este siguiente chunk es una unión de varios parámetros a la variable datos, que es en donde se está realizando la mayoría de la manipulación.

```r
datos <- left_join(datos, pdv_id, by="punto_de_venta")
head(datos)
```

```r
datos <- left_join(datos, sku_id, by="sku")
datos <- left_join(datos, mes_id, by=c("mes","anio"))
head(datos)
```

- Etapa 4: Modelado

Etapa 4: Modelado

Promedios Móviles

Es un modelo en el cual se utilizan promedios móviles como indicadores de tendencias utilizado para el análisis de datos históricos. De esta forma, puede crearse una serie de promedios de diferentes subconjuntos de datos para suavizarlos durante un periodo de tiempo.

La línea de promedio móvil en sí es una representación del promedio de un activo cualquiera durante un periodo de tiempo en específico.

Con la información que nos puede brindar un promedio móvil, es posible suavizar maniobras y movimientos en los precios ya que su interpretación es fácil de hacer y de esta forma proporciona la dirección de la tendencia de los datos de forma rápida y visual. Además, puede ayudar a la toma de decisiones en cuanto a la dirección en la cual comerciar.

Por ejemplo, si con la gráfica de promedios móviles notamos que el precio se encuentra por encima de la línea de promedio móvil, significa que el precio es más alto que los valores promedio que se tomaron anteriormente, es decir, indica un aumento de precio. Por el contrario, si el precio se encuentra por debajo de la línea de promedio móvil, esto indica una reducción en el precio. Estas indicaciones, si bien son indicadas por el cálculo del promedio móvil, pueden llegar a ser subjetivas ya que se trata de un pronóstico.

Utilizados por sí solos, los promedios móviles no pueden dar una predicción tan precisa como para notar señales en los datos a tiempo. Por lo tanto, una buena práctica es tener tanto promedios móviles a corto plazo como promedios móviles a largo plazo para que las señales se produzcan más adecuadamente. También es posible que el promedio móvil se combine con algún indicador de otro tipo con el mismo beneficio.

Dentro del modelo de promedios móviles, existen distintas formas de calcularlo:

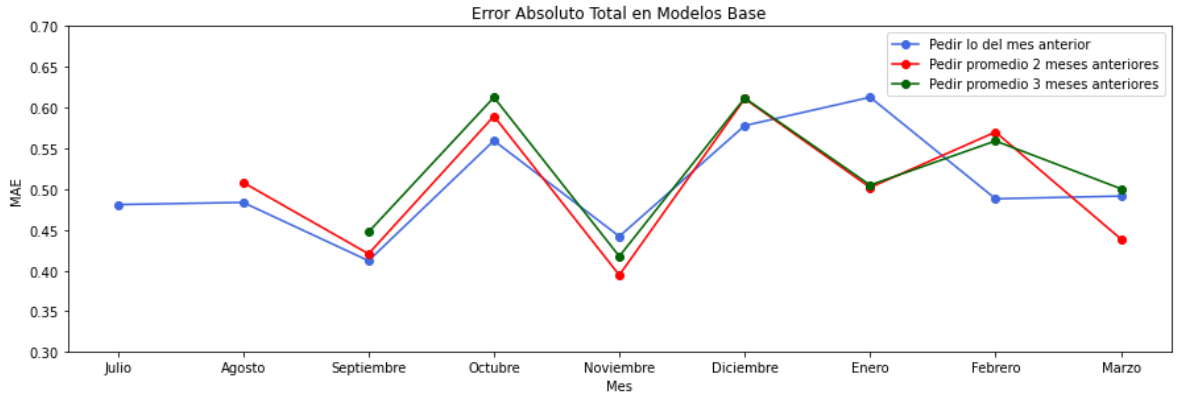
- Promedio móvil simple: es un promedio aritmético que suaviza la curva de precios. En este promedio móvil, se le da importancia igual al primer periodo y al último, solamente basta con elegir el número de periodos que se va a considerar para el promedio.
- Promedio móvil exponencial: también es un promedio aritmético que suaviza la curva de precios, sin embargo, esta variante le pone más importancia a los últimos datos del periodo en cuestión.
- Promedio móvil ponderado: en esta variante también se le da mayor importancia a los últimos datos en el periodo. Aquí se le asignan

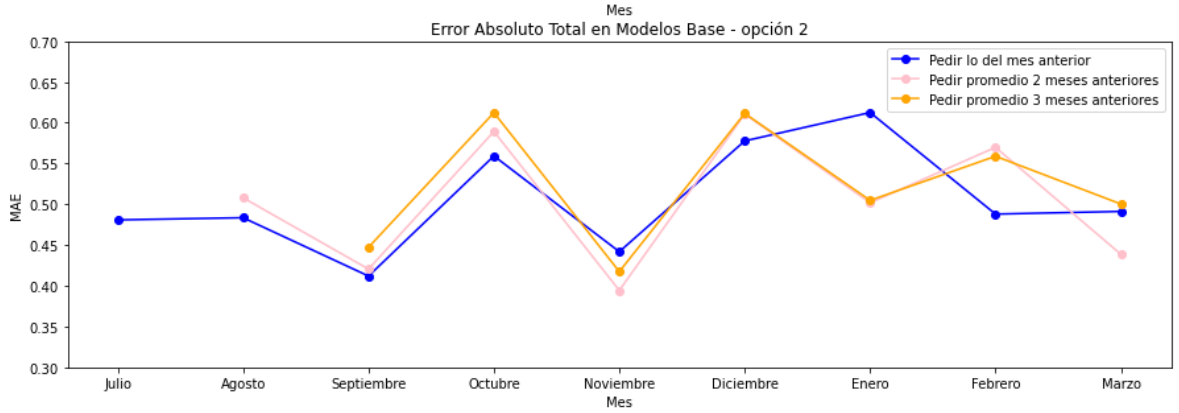
ponderaciones o pesos. Para calcularlo, se considera el valor promedio, el valor actual, el número de periodos a considerar y el factor de ponderación.
- Promedio móvil de Hull: hace uso de promedios móviles ponderados al igual que la raíz cuadrada del periodo en lugar del periodo en sí; de esta manera, puede reducir el retraso.
- Promedio móvil KAMA: parte de un promedio convencional hasta que se consigue una media que elimine el ruido del mercado para descubrir si hay tendencia en un periodo o no. Esto se realiza utilizando el apoyo de una razón de eficiencia (ER por sus siglas en inglés) que elimina el factor de peso.

Además, tenemos algunas métricas que nos apoyan al momento de ver el alcance del modelo:

- Error absoluto medio (MAE): mide la magnitud promedio de los errores en un conjunto de pronósticos. En este cálculo, la dirección de los errores no se toma en cuenta.
- Root mean square error (RMSE): Es la raíz cuadrada del promedio de diferencias entre pronóstico y valor real elevadas al cuadrado, es decir, es la raíz cuadrada del promedio de los errores elevados al cuadrado.
- Mean square error (MSE): Calcula qué tan cercana es la línea de regresión respecto a los datos tomando los errores y elevándolos al cuadrado para eliminar signos negativos. Se utiliza el promedio de todos los errores elevados al cuadrado.

Conociendo las características de los promedios móviles, nos fue posible aplicar este modelo a los datos que ya tenemos.





En qué consiste el modelo de aprendizaje de máquina que seleccionaron

Árboles de decisiones

Un árbol de elección es un modelo predictivo que divide el espacio de los predictores agrupando visualizaciones con valores semejantes para la variable contestación o dependiente. Para dividir el espacio muestral en sub-regiones es necesario utilizar una secuencia de normas o elecciones, para que cada sub-región contenga la más grande proporción viable de personas de una de las poblaciones. Si una subregión tiene datos de diferentes clases, se subdivide en zonas más pequeñas hasta fragmentar el espacio en sub-regiones menores que unen datos de la misma clase. Los árboles de elección están compuestos por nodos y su lectura se hace de arriba hacia abajo.

Dentro de un árbol de elección distinguimos diversos tipos de nodos:

Primer nodo o nodo raíz: en él se genera la primera separación en funcionalidad de la variable de mayor relevancia.

Nodos internos o intermedios: tras la primera separación pudimos encontrar dichos nodos, que vuelven a dividir el grupo de datos en funcionalidad de las cambiantes.

Nodos terminales u hojas: se hallan en la parte inferior del esquema y su funcionalidad es indicar la categorización definitiva.

Limitantes de un árbol de decisiones :

Tienden al sobreajuste de los datos, por lo que el modelo al predecir nuevos casos no estima con el mismo índice de acierto.

Se ven influenciadas por las observaciones anormales, creando árboles con ramas muy profundas que no predicen bien para nuevos casos. Se deben eliminar dichos outliers.

No suelen ser muy eficientes con modelos de regresión.

Crear árboles demasiado complejos puede conllevar que no se adapten bien a los nuevos datos. La complejidad resta capacidad de interpretación.

Se pueden crear árboles sesgados si una de las clases es más numerosa que otra.

Se pierde información cuando se utilizan para categorizar una variable numérica continua.

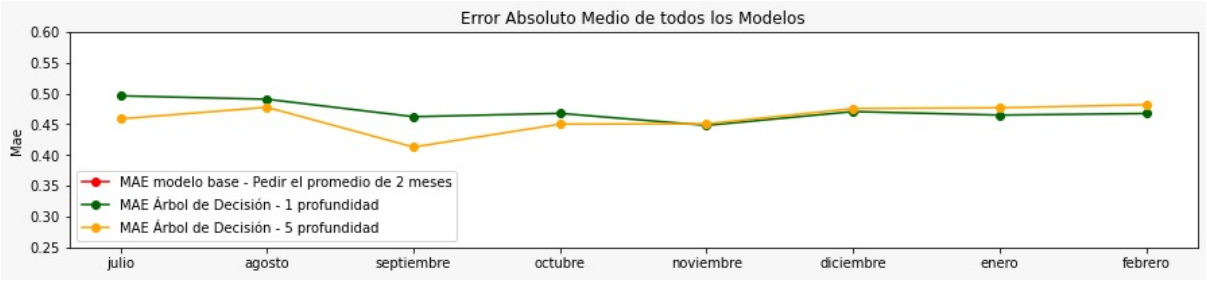
Consideraciones

Un árbol de decisión, comienza con un único nodo y luego se ramifica en resultados posibles. Cada uno de esos resultados crea nodos adicionales, que se ramifican en otras posibilidades lo que va generando que el árbol se expanda. Esto le da una forma similar a la de un árbol en expansión desde un nodo hasta los necesarios..

Existen 3 tipos diferentes de nodos: nodos de probabilidad, nodos de decisión y nodos terminales. Se comienza con la decisión principal, agrega nodos de decisión y probabilidad, se continúa con la expansión hasta que cada línea alcance un extremo.

Dado que depende de cuantos meses se le otorguen de información dependerá la calidad de la información que se tendrá ya que si se tiene un periodo más amplio de análisis estos resultados serán más acertados.


Error Absoluto Medio de todos los Modelos
Mae
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
julio
agosto
septiembre
octubre
noviembre
diciembre
enero
febrero
MAE modelo base - Pedir el promedio de 2 meses
MAE Árbol de Decisión - 1 profundidad
MAE Árbol de Decisión - 5 profundidad

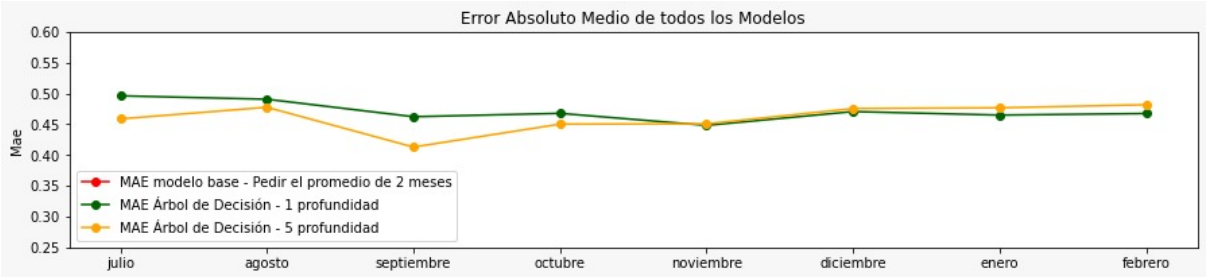

Error Absoluto Medio de todos los Modelos
Mae
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
julio
agosto
septiembre
octubre
noviembre
diciembre
enero
febrero
MAE modelo base - Pedir el promedio de 2 meses
MAE Árbol de Decisión - 1 profundidad
MAE Árbol de Decisión - 5 profundidad

Etapa 5: Evaluación

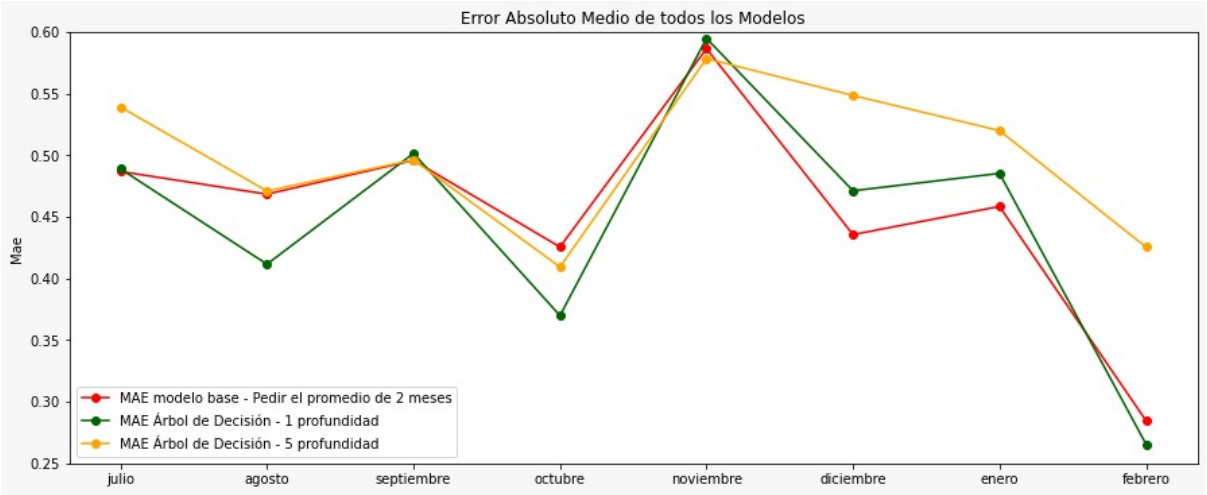
Desde las primeras consideraciones, a la infancia de este modelo ha existido una pregunta, una pregunta que varios miembros de este equipo hemos tenido; Que tan exacto, o que tan útil puede ser un modelo de esta naturaleza, claro que todos quisiéramos que fuese perfecto y exacto aumentando ganancias al 3000% venderlo a Movistar y retirarnos a los 22 pero desafortunadamente la perfección se persigue, no se alcanza, aun así, en esa persecución elaboramos un modelo que, en conjunto con otras métricas puede llevar a una optimización de la distribución de unidades, inventario.

Como hemos previamente aludido este modelo utilizó promedios móviles para obtener indicadores de tendencias, utilizando las ventas anteriores para así "predecir" las ventas de los meses siguientes, pero esto pone un problema, mismo que se propaga en el modelo, esté siendo que dependiendo de que tantos "meses" (o unidades de tiempo) se tomen en consideración es que tan acertado va a ser con la predicción, cosa que es complicada por el rápido cambio de modelos y vida útil de los celulares modernos. Esta es la mayor limitante de nuestro modelo y la razón por la cual concluimos que debe ser apoyado por otras métricas en el proceso de toma de decisiones empresariales.

En la siguiente gráfica podemos apreciar el Error Absoluto Medio de los modelos considerados, de profundidad 1 y profundidad 5, podemos observar cómo son similares, aunque no exactamente iguales, estas pequeñas diferencias serán cruciales al momento de determinar cuál de los modelos a utilizar. No obstante la falta de el MAE del modelo base, se puede observar como la línea del modelo de profundidad 5 varía más que la de profundidad 1, sin el modelo base no se puede determinar si esta variación es correcta, pero nos permite conocer la diferencia entre ambos "métodos de aprendizaje".



En la siguiente gráfica podemos observar el resultado de los errores junto con el modelo base, de esta manera confirmando su utilidad. Podemos observar que aunque las 3 líneas siguen un mismo patrón de movimiento, la línea amarilla, correspondiendo al modelo de profundidad 5, es mucho más acertada que la de profundidad 1, esto demuestra lo que hemos mencionado anteriormente, de que estos modelos son muy dependientes de que tanto “tiempo de aprendizaje” les demos, mientras más, serán más acertados, pero no siempre tenemos la oportunidad de darnos ese lujo.



Después de poder revisar ambos entornos (entrenamiento y prueba) nos podemos percatar de los usos y las limitaciones de este modelo, las cuales, afortunadamente no son nada nuevo.

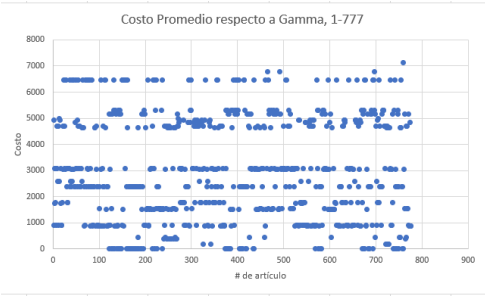
Una vez más llegamos a la conclusión de que el modelo debe ser suplementado de otras métricas y no debe ser tomado por sí solo, ya que sufre de la limitación del tiempo de entrenamiento y la estacionalidad, el tiempo de entrenamiento es algo de lo que hemos hablado extensivamente a lo largo del reporte, mientras mas tiempo de entrenamiento tenga mas acertado sera, pero existen mil y un razones por las que no se nos puede dar el lujo del tiempo, sea que por las políticas de la empresa no se tengan registros extensos, no se permita el uso de tantos datos o querer el modelo antes de que exista ese “historial”, el tiempo simplemente no es un recurso que siempre esté disponible.

Otro factor a considerar es la estacionalidad, y es la razón principal porque la que considerar otra métrica sería vital para el uso del modelo, este modelo tiene dificultad para considerar temporadas de venta, como el buen fin, esto podría llevar a una falta de unidades en las mejores épocas de venta al año.

Para finalizar, no obstante a sus limitaciones este modelo es un método de predicción útil que permitiría una redistribución de unidades a la sedes de venta, eliminando inventario inutil y un aumento de utilidades.

Anexos

Anexo 1. Diagrama de Dispersión Costo vs. Gama



$$M = \frac{\sum_{t=1}^{n} W_t * V_t}{\sum_{t=1}^{n} W_t}$$

Formulación para calcular promedios móviles ponderados

EMA (T) = EMA (T – 1) + K* ((Precio(T) – EMA(T-a))
T = Valor actual
T-1 = Valor previo
K = 2 / (n-1) (n=periodo elegido para EMA)

Formulación para calcular promedios móviles exponenciales

ER = (cambio de precio para el periodo) / (suma de cambio absoluto de precio para cada vela)

Formulación para calcular la razón de eficiencia

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n} |y_j - \hat{y}_j|$$

Formulación para calcular Mean Absolute Error

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

Formulación para calcular Root Mean Squared Error

$$\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2$$

Formulación para calcular Mean Squared Error

- Bibliografía

1. Expansión. (2021, 11 marzo). Motorola se hace peligroso para Samsung en el mercado nacional. https://expansion.mx/tecnologia/2021/03/11/motorola-se-hace-peligroso-para-samsung-en-el-mercado-nacional
2. Noguez, R. (2021, 11 agosto). LG y Huawei empiezan a decir adiós a México; Samsung, Motorola, Apple y Xiaomi ganan terreno. Forbes México. https://www.forbes.com.mx/lg-huawei-adios-mexico-samsung-motorola-apple-xiaomi-ganan-terreno/

¡IMPORTANTE! El reporte NO debe de contener NADA de código, a menos que sea sumamente importante.