CASO PARA TITULACIÓN

Modelos de Aprendizaje de Máquina para la Construcción de Portafolios de Productos de Telefonía Celular

Escrito por: Ana Luisa Masetto Herrera

Contenido

1 Introducción **4**

2 Descripción de la Problemática **5**

2.1 Información sobre la Compañía 5

2.2 Problemática que enfrenta la Empresa 8

2.3 Definición del Problema en Términos de Aprendizaje de Máquina 8

2.4 Supuestos y Restricciones 10

3 Descripción de los Datos **11**

3.1 Descripción de los Datos proporcionados por la Empresa 11

3.1.1 Catálogo de Puntos de Venta 11

3.1.2 Registro de Ventas 11

3.2 Limpieza de Datos: Problemas de Calidad en los Datos y cómo tratarlos 12

3.2.1 Problemas de Calidad: Catálogo de Puntos de Venta 12

3.2.2 Problemas de Calidad: Registro de ventas 12

3.2.3 Limpieza de los Datos 12

3.3 Ingeniería de Características 14

4 Modelos empleados actualmente por la Empresa ABCD **18**

4.1 Estructura de los Modelos Base 18

4.1.1 Estructura de los Datos 18

4.1.2 Descripción de los Modelos Base 18

4.2 Resultados de los Modelos Base 19

5 Validación Cruzada para Series de Tiempo **21**

6 Propuestas de Modelos de Aprendizaje de Máquina **23**

6.1 ¿Por qué aplicar modelos de aprendizaje de máquina? 23

6.2 Modelos 23

6.2.1 Árboles de Decisión 24

6.2.2 Bosques Aleatorios 24

6.3 Construcción de los modelos 25

6.4 Modo de empleo 26

6.5 Resultados de los Modelos propuestos 26

6.5.1 Resultados: Bosques Aleatorios 26
6.5.2 Resultados: Árboles de Decisión 27

7 Resultados finales **30**
7.1 Modelo Base vs. Modelos Propuestos 30

8 Conclusiones y Recomendaciones **32**

9 Anexos **34**
9.1 Anexo 1: Herramientas 34
9.2 Anexo 2: Análisis Exploratorio de los Datos 35
9.3 Anexo 3: Variables empleadas en los modelos de aprendizaje de máquina 39
9.4 Anexo 4: Errores de Entrenamiento de los Modelos propuestos 41

1 Introducción

La industria de las telecomunicaciones ha crecido drásticamente en los últimos años debido a los diversos avances tecnológicos que se han alcanzado; dentro de esta industria se encuentra el sector enfocado a la telefonía celular, sector que también ha presentado un aumento considerable en su demanda de productos.

De acuerdo con información recopilada en conjunto por el Instituto Nacional de Estadística y Geografía (INEGI), la Secretaría de Comunicaciones y Transportes (SCT), y el Instituto Federal de Telecomunicaciones (IFT), el uso de la telefonía celular ha ganado lugar como una de las tecnologías con mayor penetración en la población mexicana, estimando que en el 2018 había un total de 69.6 millones de personas que tenían un teléfono inteligente, indicando un incremento de usuarios del 7.57% en comparación con el 2017 [1].

En la actualidad, se sabe que prácticamente todas las personas poseen un celular, no solamente para facilitar la comunicación entre amigos, familiares, compañeros de trabajo, clientes, etc. sino que también se ha convertido en una herramienta que facilita algunas de las actividades cotidianas de las personas, como: buscar direcciones, pedir comida, solicitar información bancaria, realizar documentos para tareas o trabajos, o simplemente funciona como fuente de entretenimiento gracias a su capacidad para conectarse a redes sociales y almacenar juegos, videos, fotos y música.

Si bien dijo Oswaldo Contreras Saldívar, presidente del Instituto Federal de Telecomunicaciones: "No sólo se usa el dispositivo móvil por lo práctico, sino porque lo queremos usar para todo, todo el tiempo: lo queremos al alcance de la mano" [2]; es muy importante resaltar que, aunque la necesidad de tener un dispositivo móvil es de la mayoría de las personas, no todas buscan las mismas características en estos. Hay personas que se enfocan únicamente en el rango de precios (gamma del producto), en la marca y en la apariencia; pero también hay personas que se guían más por la construcción en sí del modelo, como son la capacidad de memoria, la vida útil del producto, la calidad de la cámara, el software que utiliza, etc.

Es por eso, por lo que las compañías enfocadas a la venta de productos de telefonía celular enfrentan como **reto principal** el **pronosticar el número unidades a vender de cada producto en sus diversos puntos de venta**, ya que de no hacerse propiamente esto podría generar problemas relacionados con la pérdida de clientes o problemas como gastos adicionales en trasporte o almacenamiento.

Esto es porque las compañías de este sector cuentan con puntos de venta en distintas zonas geográficas con características distintivas de sus respectivas poblaciones, por ende, es muy probable que el mercado al que se dirige cada uno de estos puntos no sea el mismo.

Partiendo de esta situación, este documento presenta un **proyecto de Ciencia de Datos**, el cual utiliza información real de una empresa de telecomunicaciones en México. Este proyecto tiene como finalidad **proponer modelos de aprendizaje de máquina como solución a la problemática de construcción de portafolios de telefonía celular para la compañía "ABCD"**.

Cabe mencionar que este proyecto se realizó en las instalaciones de IBM México y los datos utilizados corresponden a uno de sus clientes, por lo tanto, y dado el **contrato de privacidad** que se tiene con los clientes de IBM, el **nombre real de la empresa** en cuestión **permanecerá en anonimato** asignándole un alias.

2 Descripción de la Problemática

En esta sección del documento se determina el contexto general de la compañía y de la problemática a tratar. La idea general es extraer información fundamental que permita estructurar las bases del proyecto.

2.1 Información sobre la Compañía

La **compañía ABCD** es una empresa dedicada a la industria de las telecomunicaciones enfocada al sector de telefonía móvil. Actualmente, la compañía cuenta con **1,909 puntos de venta** en toda la República Mexicana, donde dispone de **455 productos** diferentes provenientes de **12 marcas** distintas.

La siguiente tabla muestra un resumen de las características más generales de la compañía:

Característica de la Compañía	Valores
Puntos de venta:	1,909 tiendas en la República Mexicana
Rango de fecha de los datos:	01 de junio de 2018 al 31 de marzo de 2019 - 10 meses de registro
Productos:	455 productos distintos, cada uno identificado por un código único (SKU)
Marcas:	12 marcas telefónicas que vende la compañía ABCD. - Affix, Alcatel, Apple, Hisense, Hauawei, Lannix, Lenovo, LG, Motorola, Samsung, Sony, y ZTE.
Gammas de producto:	4 gammas en la que fueron clasificadas los productos de acuerdo a su costo promedio, - Premium, Alta, Media y Baja.
Zonas en las que esta divido el territorio:	8 zonas - Centro sur, Centro Occidente, Golfo de México, Norte, Pacífico Sur, Península de Yucatán, Noreste y Noroeste.
Estados :	32 estados de la república mexicana en los que tiene presencia la compañía.
Ciudades :	228 ciudades en las que tiene presencia la compañía.

Figura 1: Características más importantes de la Compañía ABCD.

Como visualización adicional, se presenta el siguiente conjunto de mapas que, además de mostrar la ubicación geográfica de los distintos puntos de venta, provee información relacionada las ventas de cada punto de venta en los 10 meses de registro con los que se cuenta.

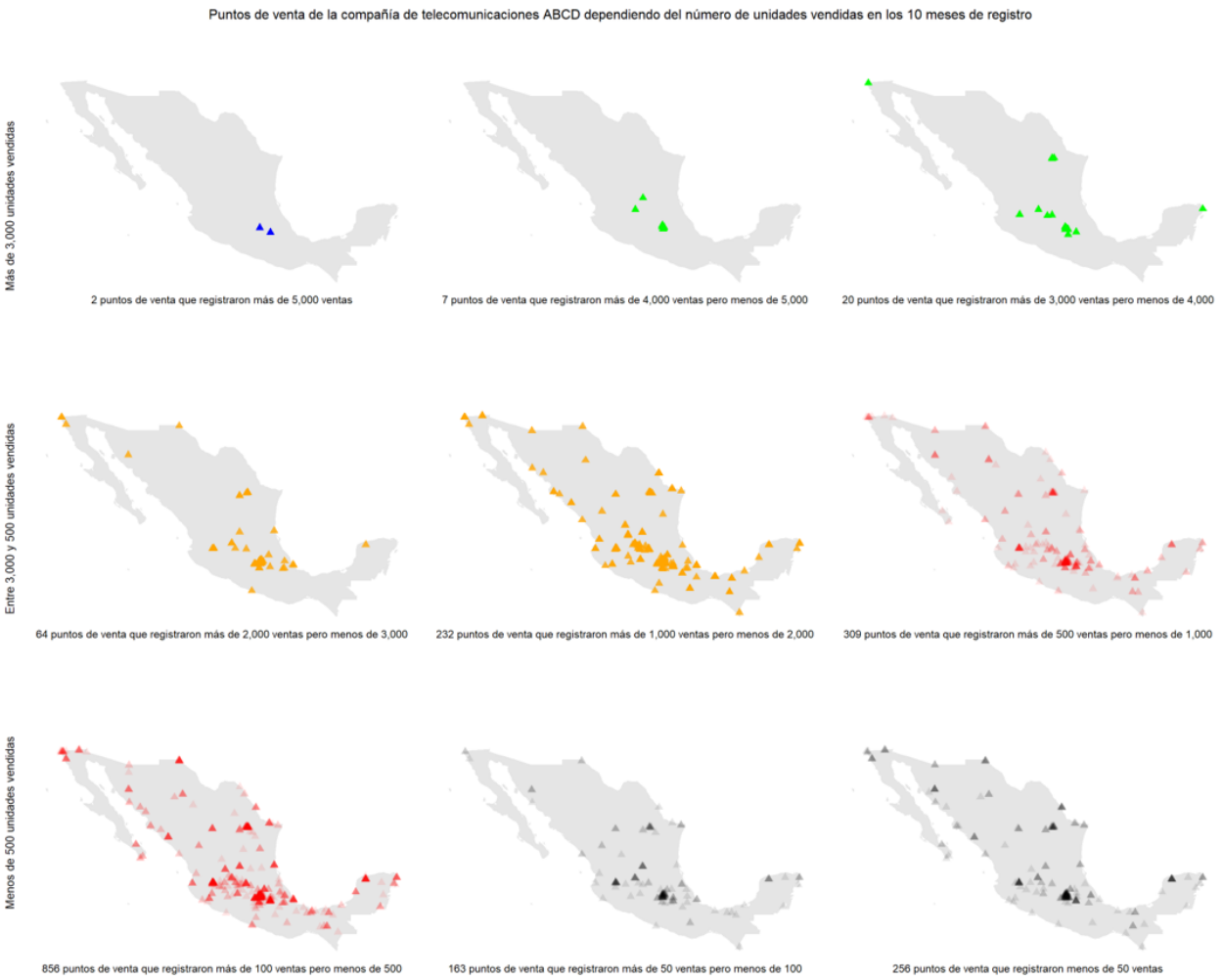


Figura 2: Puntos de venta de la Compañía ABCD clasificados dependiendo del Número de Unidades totales que vendieron en los 10 Meses de Registro.

En la figura anterior se puede observar que son únicamente dos puntos de venta en la zona *centro sur*, en la Ciudad de México y en Puebla, donde se vendieron mas de 5,000 unidades en los 10 meses de registro. También se puede observar que más de la mitad de los puntos de venta (1,275 puntos de venta) vendieron menos de 500 unidades en 10 meses; y que únicamente 29 puntos de venta, la mayoria ubicados en la zona *centro sur*, vendieron más de 4,000 unidades.

Por lo tanto, con este breve análisis, se puede observar que **no** todos los puntos de venta venden el mismo número de unidades, ni los mismos productos como se puede observar en la siguiente gráfica que resume **el número de unidades que se vendieron de cada marca en cada uno de los 32 estados de la República Mexicana**.

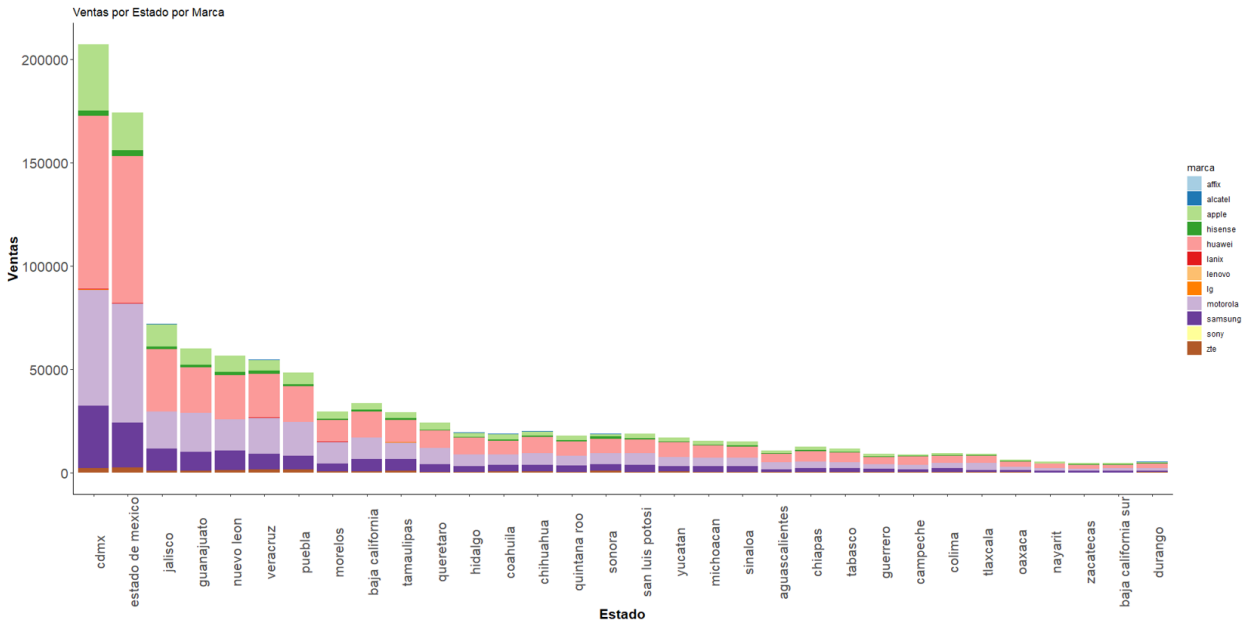


Figura 3: Número de ventas por estado dependiendo de la marca en los 10 meses de registro.

En la gráfica anterior, a pesar de ser únicamente a nivel estado, se puede apreciar que no en todos los estados tienen la misma presencia todas las marcas; aunque en casi todos los estados *Huawei* es la marca con mayor número de ventas registradas en los 10 meses de registro, se puedeo observar que hay marcas como: *ZTE*, *Affix* y *Lanix* donde sus unidades vendidas ni siquiera se alcanzan a apreciar.

Por ende, se puede observar con estas visualizaciones el problema que enfrenta la empresa, donde determinar el número de unidades a vender en cada punto de venta resulta una tarea muy importante, ya que a simple vista se puede apreciar la diferencia en tanto cantidad, como en diferencia de productos consumidos en los diferentes estados y puntos de venta.

En caso de querer indagar más a detalle en la situación de la empresa, en la sección de anexos (anexo 1) se encuentra la liga con la carpeta correspondiente al análisis exploratorio completo de los datos en el repositorio de github, de la misma manera, en el anezxo 2 se encuentra un análisis exploratorio más completo.

2.2 Problemática que enfrenta la Empresa

Hoy en día, la compañía *ABCD* presenta el *reto* de **determinar al final de cada mes cuál es el número de unidades a vender por producto en cada punto de venta para el mes siguiente**. Se sabe que la compañía **no** utiliza modelos de aprendizaje de máquina para enfrentar dicha situación y es por eso que el propósito de este documento es **plantear propuestas de modelos de aprendizaje de máquina** buscando que su desempeño en tanto al pronóstico de unidades sea **mejor** y **más estable** al actual.

2.3 Definición del Problema en Términos de Aprendizaje de Máquina

Para construir propuestas de modelos que permitan lidiar mejor con la problemática de la empresa, es necesario desglosarlo de tal manera que se pueda entender mejor y así su planteamiento sea más sencillo. En primer lugar, se tiene un problema que requiere de **modelos de aprendizaje supervisado**, es decir, modelos que tienen como tarea estimar una variable respuesta a partir de ciertos datos de entrada. En segundo lugar, se tiene un **problema de regresión** dado que la variable a predecir va a tener valores continuos. Finalmente, se tiene un **problema de series de tiempo** ya que los datos proporcionados por la compañía contienen valores observados correspondientes a sus ventas a lo largo de 10 meses, valores que están secuencialmente ordenados y, por ende, no pueden ser permutados.

Ya que se definió el problema de esta manera, se puede tomar como base el **proceso generador de datos (modelo teórico)** [3]. De este proceso se puede rescatar que Y es la variable que se busca predecir; X es una variable, o conjunto de variables, que se espera que pueda mejorar la predicción de Y; y que Y y X están relacionadas de la siguiente manera:

$$Y = f(X) + \epsilon$$

Donde: f expresa la relación sistemática que hay entre X y Y, y ϵ representa el efecto de variables que no se han medido o procesos aleatorios que determinan la respuesta.

En problemas de Aprendizaje de Máquina, f no se conoce, por ende, hay que estimarla $\hat{f}$ con ayuda de los datos (aprender de los datos). Para llevar a cabo esta tarea, es importante dividir los datos en conjuntos de entrenamiento y de prueba, para después revisar su desempeño considerando métricas que permita comparar el valor real con el de las predicciones [4].

> La idea general de un algoritmo de aprendizaje de máquina es aprender de una muestra de entrenamiento y así generar una $\hat{f}$ que permita hacer predicciones.
>
> $$\hat{Y} = f(\hat{X})$$

Para el caso de esta compañía, es de suma importancia definir claramente cada uno de los conceptos base que se explicaron previamente. En primer lugar, **la variable respuesta (Y)** es:

> $Y_{i,j,k+1}$: Número de unidades del producto con índice i en el punto de venta con índice j que se van a vender en el mes siguiente del mes de registro k.

Donde:

i : 1, 2, 3, ... , 455;

j : 1, 2, 3, ... , 1909;

k : 0, 1, 2, ... , 8.

Con esto presente, lo que busca este proyecto es **estimar** el valor de la variable $Y_{i,j,k+1}$:, por lo tanto, se quiere calcular:

$\hat{Y_{i,j,k+1}}$: Predicción del número de unidades del producto con índice i en el punto de venta con índice j que se van a vender en el mes siguiente del mes de registro k.

Donde: i, j, k toman los mismos valores que se enlistaron previamente.

En segundo lugar, es importante mencionar que los modelos que se proponen para construir la función $f(\hat{X})$ son: **Árboles de Decisión** [5] y **Bosques Aleatorios** [6].

En tercer lugar, hay que definir cuáles serán los **datos de entrada** (X) que permitirán que los modelos aprendan. La compañía proporcionó información relacionada con los registros de venta de los productos y algunas características de estos y de los puntos de venta. La forma en la que se trata esta parte se describe con mayor detalle más adelante en la sección de limpieza e ingeniería de características de los datos.

En cuarto lugar, es importante considerar que los modelos requieren dividir los datos en dos conjuntos; entrenamiento y prueba, sin embargo, el problema requiere de un planteamiento adicional y más complejo para tratar su temporalidad, es decir, requerirá de un proceso llamado: **validación cruzada para series de tiempo**, proceso que se detalla más adelante.

Finalmente, es necesario determinar las métricas que evaluaran el desempeño de los modelos a nivel mensual, en este caso se opta por utilizar 3 **métricas**:

- **Error Absoluto Medio:** Mide el promedio de la magnitud de los errores en un conjunto de predicciones [7].

$$MAE(Y, \hat{Y}) = \frac{1}{n} \sum_{i=1}^{n} |y_{i,j,k} - \hat{y_{i,j,k}}|$$

- **Raíz del Error Cuadrático Medio:** Muy parecido al error absoluto medio, esta métrica permite medir la magnitud del error con la diferencia de que esta métrica tiene la particularidad de dar mayor peso a errores más grandes [8].

$$RMSE(Y, \hat{Y}) = \sqrt{\frac{1}{n} \sum_{t=1}^{n} (y_{i,j,k} - \hat{y_{i,j,k}})^2}$$

- **Error Cuadrático Medio:** Al igual que la raíz del error cuadrático medio, esta métrica mide la magnitud del error de predicción, sin embargo, las unidades de esta métrica terminan siendo cuadráticas, por ende, su interpretación no es tan fácil.

$$MSE(Y, \hat{Y}) = \frac{1}{n}\sum_{t=1}^{n}(y_{i,j,k} - \hat{y_{i,j,k}})^2$$

2.4 Supuestos y Restricciones

Adicional al planteamiento anterior, es importante mencionar los supuestos y restricciones que se consideran a lo largo del proyecto.

Desde un punto de vista de producción, lo ideal sería contar con información más detallada de la empresa y de sus diversos puntos de venta, por ejemplo: modelos de pronóstico que se emplean actualmente, capacidad de espacio de almacenamiento en cada punto de venta, unidades en inventario en cada punto de venta, tiempo que se tarda un orden en ser cumplida (tiempo de reabastecimiento), etc. Sin embargo, dicha información no fue proporcionada, por lo tanto, los siguientes supuestos se construyen:

1. Los modelos que se emplean actualmente corresponden a: **Pedir lo del mes anterior** o **pedir un promedio** de los meses anteriores. Modelos que se construyen y describen más adelante.
2. La **capacidad de almacenamiento** en cada punto de venta con relación a cada producto es de **100** unidades.
3. Cuando se realiza la orden de pedido con las unidades pronosticadas, **la orden se cumple inmediatamente**.

3 Descripción de los Datos

Antes de discutir los modelos base de pronóstico de la empresa y los modelos propuestos como solución alternativa a éstos, es importante describir la información que fue proporcionada, con el fin de dar una visualización del modo incial en el que se tenían los datos y definir la estructura que deben tener para poderse emplear en los modelos. La empresa proporcionó dos bases de datos: la primera corresponde a un catálogo de tiendas y la segunda a un registro de ventas.

3.1 Descripción de los Datos proporcionados por la Empresa

3.1.1 Catálogo de Puntos de Venta

El catálogo fue proporcionado en un archivo con extensión csv y tiene un total de 1,911 renglones y 79 columnas. Las 79 variables poseen diferentes propiedades relacionadas con los puntos de venta, sin embargo, de éstas sólo se tomarán en cuenta las siguientes:

- **Nombre del pdv:** Nombre del punto de venta.
- **Nuevo nombre del pdv:** Algunos de los puntos de venta fueron renombrados.
- **Regiones homologadas:** Región a la que pertenece cada punto de venta (norte, sur, etc.).
- **Estado:** Estado donde se encuentra el punto de venta.
- **Ciudad:** Ciudad donde se encuentra el punto de venta.
- **Latitud:** Ubicación con coordenadas geográficas del punto de venta.
- **Longitud:** Ubicación con coordenadas geográficas del punto de venta.

Las razones por las cuales no se consideran las demás variables son: En primer lugar, no es claro a qué se refieren algunas variables y no se proporcionó su respectiva descripción; en segundo lugar, existen muchos valores faltantes y la información para completarlos no es posible de obtener por cuenta propia y la empresa no accedió a proporcionarla; por último, hay variables con información estimada de la cuál no se sabe con certeza las unidades ni la forma en la que se calcularon.

3.1.2 Registro de Ventas

El segundo documento contiene los registros de ventas de la empresa, registros que tienen lugar a partir del 1 de junio del 2018 al 31 de marzo del 2019. El archivo con extensión csv con dicha información tiene 1,048,575 renglones y 10 columnas. Las 10 variables que se tienen son las siguientes:

- **Punto de Venta:** Nombre del punto de venta donde se realizó la compra.
- **Plan tarifario:** Plan tarifario bajo el cual se vendió la unidad.
- **Sku:** Código único del producto (códigos internos de la compañía, por lo tanto, **no se sabe el nombre comercial de los productos**).
- **Fecha:** Fecha en la que se registró la venta de la unidad.
- **Precio:** Precio de la unidad.
- **Costo:** Costo de la unidad. Estos valores son muy cercanos a los de la variable anterior.
- **Marca:** Marca de la unidad vendida.
- **Ventas:** Cada registro dentro del documento corresponde a una venta.
- **Mth:** Mes en el que se hizo la venta de la unidad.
- **Yr:** Año en la que se hizo la venta de la unidad.

3.2 Limpieza de Datos: Problemas de Calidad en los Datos y cómo tratarlos

Ambos registros tenían diversas fallas de calidad, las cuales se enlistan a continuación:

3.2.1 Problemas de Calidad: Catálogo de Puntos de Venta

- **Máyusculas y minúsculas:** Había celdas con información en mayúsculas y otras en mínusculas.
- **Caracteres especiales:** Algunos de los registros tenían acentos, guiones y dobles espacios.
- **Valores faltantes:** La columna **Nuevo nombre del pdv** poseía muchos valores faltantes, al igual que las columnas de **longitud y latitud**.
- **Tiendas faltantes en el catálogo:** Para detectar este problema, se intentó hacer un join del registro con el catálogo y se detectaron algunas tiendas en el registro que no estaban en el catálogo.
- **Tiendas repetidas en el catálogo:** Habían tiendas registradas más de una vez en el catálogo.
- **Registros erróneos:** La columna de estados tenía más de 32 estados registrados; la columna de las regiones tenía registros que no hacían sentido; Por un lado, la columna de longitud tenía coordenadas registradas positivas y eso no es posible dado que los estados de la república mexicana únicamente abarcan coordenadas de longitud entre -86 y -116 aproximadamente. Por el otro lado, la columna de latitud tenía coordenadas registradas fuera de rango, por ejemplo:Registro cuya latitud era de 20 millones.

3.2.2 Problemas de Calidad: Registro de ventas

- **Mayúsculas y minúsculas:** No se tenía un formato definido en tanto a la forma en la que se debía de hacer los registros. Había celdas con información en mayúsculas y otras en mínusculas.
- **Valores faltantes:** La columna de precio tenía más de 7,000 valores faltantes; la columna de marca tenía 9 valores faltantes; y la columna de costo tenía 7 valores faltantes.
- **Formato no homogéneo en la columna de fecha:** Los registros de la columna de fecha tienen diferentes formatos (01-03-18, 01MAR18, 01-03-2018, entre otros).
- **Caracteres especiales:** Algunos de los registros tenían acentos, guiones y dobles espacios.
- **Errores de registro:** Hay marcas escritas de diversas maneras y tiendas registradas con un nombre erróneo.

3.2.3 Limpieza de los Datos

Una vez que se detectaron los problemas de calidad en ambas bases de datos, se hizo la limpieza de datos a cada conjunto de datos. A continuación, se plantean las actividades principales que se llevaron a cabo para limpiar los datos.

1. **Homogeneizar todos los registros a minúsculas:** Todos los registros de ambas bases de datos se cambiaron a minúsculas para facilitar su manejo.

2. **Eliminar caracteres especiales:** En el registro de ventas y en el catálogo de puntos de venta, todos los acentos se remueven, al igual que la letra ñ, los dobles espacios y otros tipos de espacios (Non Breaking Spaces).

3. **Imputar valores faltantes:**

- En el catálogo de tiendas, los puntos de venta que cambiaron de nombre fueron sustituidos por su nuevo nombre.
- En el catálogo de tiendas, los valores faltantes en las columnas de longitud y latitud fueron imputados. Tras un proceso sumamente artesanal de buscar cada una de las 39 tiendas con valores faltantes en sus coordenadas geográficas en Google Maps se logró recopilar la información faltante.
- Para completar las tiendas faltantes en el catálogo de tiendas se hizo un join con el registro de ventas y se detectaron las tiendas que no empataban, luego estas se intentaron buscar dentro del catálogo con otros nombres y las que no se encontraban se buscaron manualmente en Google Maps para luego agregarlas al catálogo.
- En el registro de ventas, la columna de precio tenía muchos valores faltantes, sin embargo, se habló con un experto de IBM en la industria de telecomunicaciones y se llegó al acuerdo de únicamente utilizar una variable entre precio y costo dado que los valores de estas dos variables eran prácticamente los mismos. Por ende, esta columna se descarta.
- En el registro de ventas, la columna de costo tenía 7 valores faltantes, los cuales se imputaron al buscar en los demás registros el sku de los productos con costos faltantes, luego se filtraron los resultados por punto de venta y fecha y al final se obtuvieron los valores a imputar.
- En el registro de ventas, la columna de marca tenía 9 valores faltantes, los cuales se imputaron de manera similar al caso anterior, buscando el sku de los productos con valores faltantes en esta columna en los demás registros, y al ser el sku el código único del producto fue muy sencillo encontrar a cuál marca se referían.

4. **Errores de Registro:**

- En el catálogo de tiendas, hay registrados más de 32 entidades federativas, por lo tanto, se homogenizan de tal manera que sólo se tenga información de las 32 entidades de la Repúbica Mexicana. Por ejemplo, en la columna de estado estaba escrito matamoros y este se sustituyó por tamaulipas, o el estado de méxico estaba escrito de diferentes formas y este se homogeneizó.
- En el catálogo de tiendas,los registros positivos dentro de la columna de longitud se cambiaron a valores negativos y se verificaron de tal manera que estás fueran correctas y, efectivamente, se refirieran a algún punto de venta de la compañía ABCD.
- En el catálogo de tiendas, los registros fuera de rango de la columna de latitud se corrigieron a valores dentro de rango.

- Para el catálogo de tiendas se buscó información sobre las zonas en las que se divide el territorio mexicano y se encontró en la página de Comisión Nacional para el Conocimiento y Uso de la Biodiversidad (CONABIO) que la república se puede dividir en 8 regiones, por lo tanto, se actualizan las regiones originales de la base de datos y se sustituyen por la nueva división [9].
- En el catálogo de tiendas habían 26 tiendas que estaban registradas más de una vez, éstas se analizaron para ver si su repetición era justificada o no. Por ejemplo, la tienda llamada *Chapultepec* tenía 4 registros distintos, los cuales tenían distinta ubicación (zona, ciudad y estado), por ende, lo que se hizo fue cambiar los nombres de los puntos de venta de tal manera que estas tiendas se pudieran identificar por separado. Sin embargo, hubo casos en que la tienda se repetía con los mismos valores y, por lo tanto, los registros repetidos se eliminaron.
- En el registro de ventas se les da un formato particular (AAAA-MM-DD) a todas las fechas registradas.
- En el registro de ventas hay tiendas que están registradas con un nombre erróneo, por lo tanto, se detectaron estas tiendas y se les cambió el nombre al nombre correcto dentro del catálogo.
- En el registro de ventas, en un principio se tenían 32 marcas distintas, por lo tanto, se detectó cuáles eran los registros errónes (marcas escritas de diversas formas) y se cambió de tal manera que al final solamente se tuvieron 12 marcas.

5. **Combinar información:**

Finalmente, se hizo una fusión de ambas bases de datos para tener un archivo donde se tenga información relacionada con las ventas y con algunas características geográficas de los puntos de venta.

3.3 Ingeniería de Características

Con los datos ya limpios, se trató de crear más características a partir de las que se tienen, con el fin de mejorar la eficacia predictiva de los algoritmos de aprendizaje de máquina que se proponen más adelante en el reporte. A continuación, se presentan las nuevas variables generadas:

- **Variable** *Gamma***:** Se observó que el costo de los productos cambiaban ligeramente en los diferentes periodos de venta (meses), por lo tanto, se optó por obtener el **costo promedio por producto** y utilizar este valor para construir una nueva variable llamada gamma con posibles valores: premium, alta, media y baja.

Costo Promedio	*Gamma correspondiente*
Costo promedio ≤ 5,000	*Baja*
5,000 ≤ Costo promedio ≤ 10,000	*Media*
10,000 ≤ Costo promedio ≤ 15,000	*Alta*
Costo promedio > 15,000	*Premium*

Figura 4: Construcción de la Variable Gamma.

- **Construcción de índices:** Para facilitar el uso de variables categóricas dentro de los algoritmos, se crearon índices para las variables: punto de venta, sku, marca, gamma, y fecha.

Característica	Posibles valores
Punto de venta	1 a 1,909 - *Asignados dependiendo del orden descendente de los nombres de los puntos de venta (A-Z).*
Marca	1 a 12 - *Asignados dependiendo del orden descendente de los nombres de las marcas (A-Z).*
Gamma	1 a 4 - *Asignados dependiendo de su valor en tanto a costo promedio.* - *4:Premium, 3:Alto, 2:Medio y 1:Alto*
Bloque de fecha	*0 a 9* - *Asignados dependiendo del mes de registro.* - *El mes más antiguo va a tener el valor más pequeño y el valor más alto va a ser el último mes del que se tiene registro.* - *0:junio 2018, 1:julio 2018, 2:agosto 2018, 3:septiembre 2018, ... , 8:abril 2019, y 9:marzo 2019*
Sku	1 a 455 - *Asignados dependiendo del orden descendente de los códigos únicos de los productos (A-Z).*

Figura 5: Índices para las Variables Categóricas.

- **Agrupación:** Se hace una agrupación por punto de venta, sku, marca, gamma, y fecha, para calcular el número total de ventas relacionadas con estas caracteristicas, es decir, construir una variable que permita identificar a final de mes las unidades totales que se vendieron dependiendo de su agrupación.

- **Completar serie de tiempo:** Con la agrupación anterior, se puede observar que la serie de tiempo no está completa, es decir, hay bloques de meses (0-9) e índices de productos que no aparecen en todos los puntos de venta, por lo tanto, esto se debió de completar.

  - Lo primero que se hizo fue obtener las combinaciones existentes entre punto de venta, sku y bloque de fecha, únicamente estas tres variables son consideradas para obtener el número total de combinaciones, dado que cada sku tiene asociado únicamente una gamma y una marca. El número total de combinaciones y de registros de la serie de tiempo completa es: **8,685,950**, longitud final que van a tener los registros.
  - A continuación, se relacionaron los valores obtenidos en la agrupación con la serie de tiempo, donde los valores nulos significan que no hubo ventas registradas con esas características (punto de venta, sku, marca, gamma, y fecha).

A partir de este punto, la serie de tiempo ya está completa, sin embargo, para enriquecer los modelos propuestos, se hacen conteos y promedios por duplas de características, y después se hacen rezagos de estas nuevas catacterísticas, con el fin de recopilar información adicional.

- **Conteos por grupo:** Se sacan conteos y promedios de 4 duplas de características:

  - **Ventas promedio por tienda por mes.**
  - **Ventas promedio por marca por mes.**
  - **Ventas promedio por gamma por mes.**
  - **Ventas promedio por producto por mes.**
  - **Ventas totales por tienda por mes.**
  - **Ventas totales por marca por mes.**
  - **Ventas totales por gamma por mes.**
  - **Ventas totales por producto por mes.**

- **Rezagos a tres tiempos:** Se crearon rezagos a 3 tiempos (1 mes, 2 meses y 3 meses) para contar con información del pasado para hacer las predicciones.

- **Creación de la variable respuesta:** Esta fue la parte más importante dentro de esta sección dado que es la construcción de la variable que se busca predecir. Al ser la variable respuestas: **Número de unidades de cada producto que se van a vender en cada punto de venta al siguiente mes de registro**, se tuvo que hacer un rezago más, donde los valores relacionados con las unidades vendidas a final de mes se recorren un periodo atrás en el tiempo, para que a finales de mes se pueda predecir lo que se va a vender el mes siguiente.

Figura 6: Proceso para creación de variable respuesta.

Finalmente se tiene el archivo limpio y con ingeniería de características con: 39 Variables y 8,685,950 registros; la visualización de sus primeros 10 registros corresponde a la imagen anterior. Cabe mencionar que para la implementación de los modelos, todas las observaciones correspondientes al último mes de registro (9:marzo) son eliminadas, por lo tanto, los modelos consideran **7,817,355 registros** y las variables a utilizar dependen si es modelo base o modelo propuesto de aprendizaje de máquina.

4 Modelos empleados actualmente por la Empresa ABCD

Previamente se mencionó en la sección de **Supuestos y Restricciones** que no se sabe con certeza cuál es el modelo que emplea la compañía en la actualidad, por lo tanto, se **asume** que el modelo podría corresponder a: Pedir lo del mes anterior o pedir un promedio de los meses anteriores (promedios móviles), ya que son modelos empleados con frecuencia gracias a que su implementación no es complicada.

4.1 Estructura de los Modelos Base

4.1.1 Estructura de los Datos

Al ser estos algunos de los modelos más sencillos de utilizar, no es necesario considerar las variables de conteos, promedios y rezagos, es por eso que para la construcción de estos modelos las variables a emplear son:

- **tiendaID:** Variable que indica los índices de los diferentes puntos de venta.
- **skuID:** Variable que indica los índices de los diferentes productos.
- **fechaID:** Variable que indica los índices de los periodos de los que se tiene registro.
- **ventas_por_mes:** Variable que indica el número total de unidades vendidas en cada tienda, de cada producto en los diferentes periodos de registro.
- **variable_y_ventas_del_siguiente_mes:** Variable respuesta. Variable que indica el número de unidades de cada producto que se deben de vender en cada punto de venta al siguiente mes de registro.

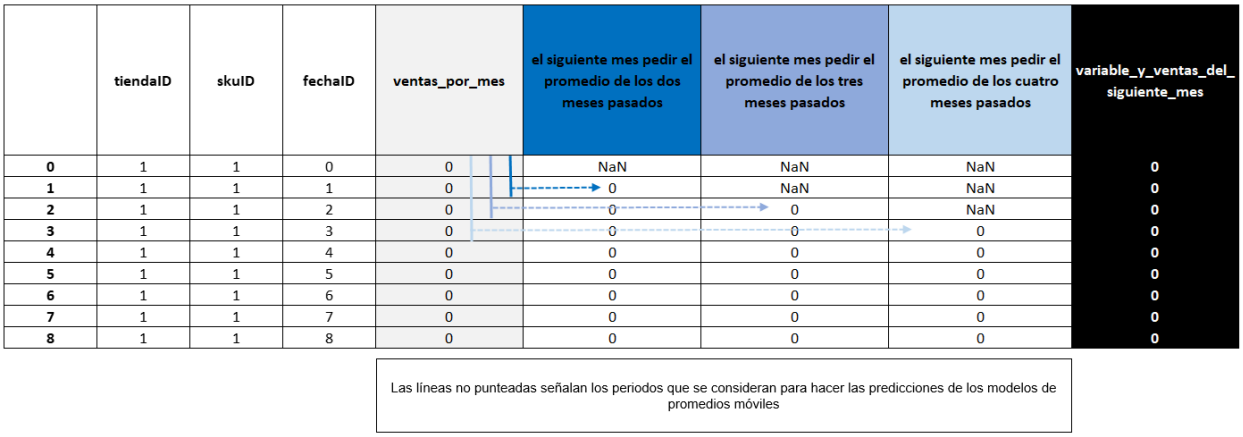
4.1.2 Descripción de los Modelos Base

Modelo Base 1: El primer modelo construye la predicción al tomar el número de unidades vendidas en cada punto de venta en el último periodo de registro y utilizar este mismo valor como su pronóstico para el siguiente mes.

Figura 7: Modelo Base 1: Pedir lo del mes anterior.

Modelo(s) Base 2: Los modelos que llevan a cabo promedios móviles crean su predicción al tomar el promedio de algunos meses anteriores al mes de pronóstico y utilizar este valor como pronóstico. En este proyecto se consideran 3 promedios distintos:

- **Promedio de los dos meses anteriores.**
- **Promedio de los tres meses anteriores.**
- **Promedio de los cuatro meses anteriores.**



	tiendaID	skuID	fechaID	ventas_por_mes	el siguiente mes pedir el promedio de los dos meses pasados	el siguiente mes pedir el promedio de los tres meses pasados	el siguiente mes pedir el promedio de los cuatro meses pasados	variable_y_ventas_del_siguiente_mes
0	1	1	0	0	NaN	NaN	NaN	**0**
1	1	1	1	0	0	NaN	NaN	**0**
2	1	1	2	0	0	0	NaN	**0**
3	1	1	3	0	0	0	0	**0**
4	1	1	4	0	0	0	0	**0**
5	1	1	5	0	0	0	0	**0**
6	1	1	6	0	0	0	0	**0**
7	1	1	7	0	0	0	0	**0**
8	1	1	8	0	0	0	0	**0**

Figura 8: Modelo(s) Base 2: Pedir un promedio de los meses anteriores.

En total se construyeron 4 modelos base; de esos modelos, el que tenga el mejor desempeño va a ser considerado como el modelo base que la empresa utiliza y el modelo cuyo desempeño se espera mejorar con los modelos de aprendizaje de máquina propuestos en las siguientes secciones.

4.2 Resultados de los Modelos Base

Ya construidos los modelos base, se procede con la obtención de los desempeños a nivel **mensual** de cada uno, dentro de las tres métricas propuesta. A continuación, se muestran los resultados:

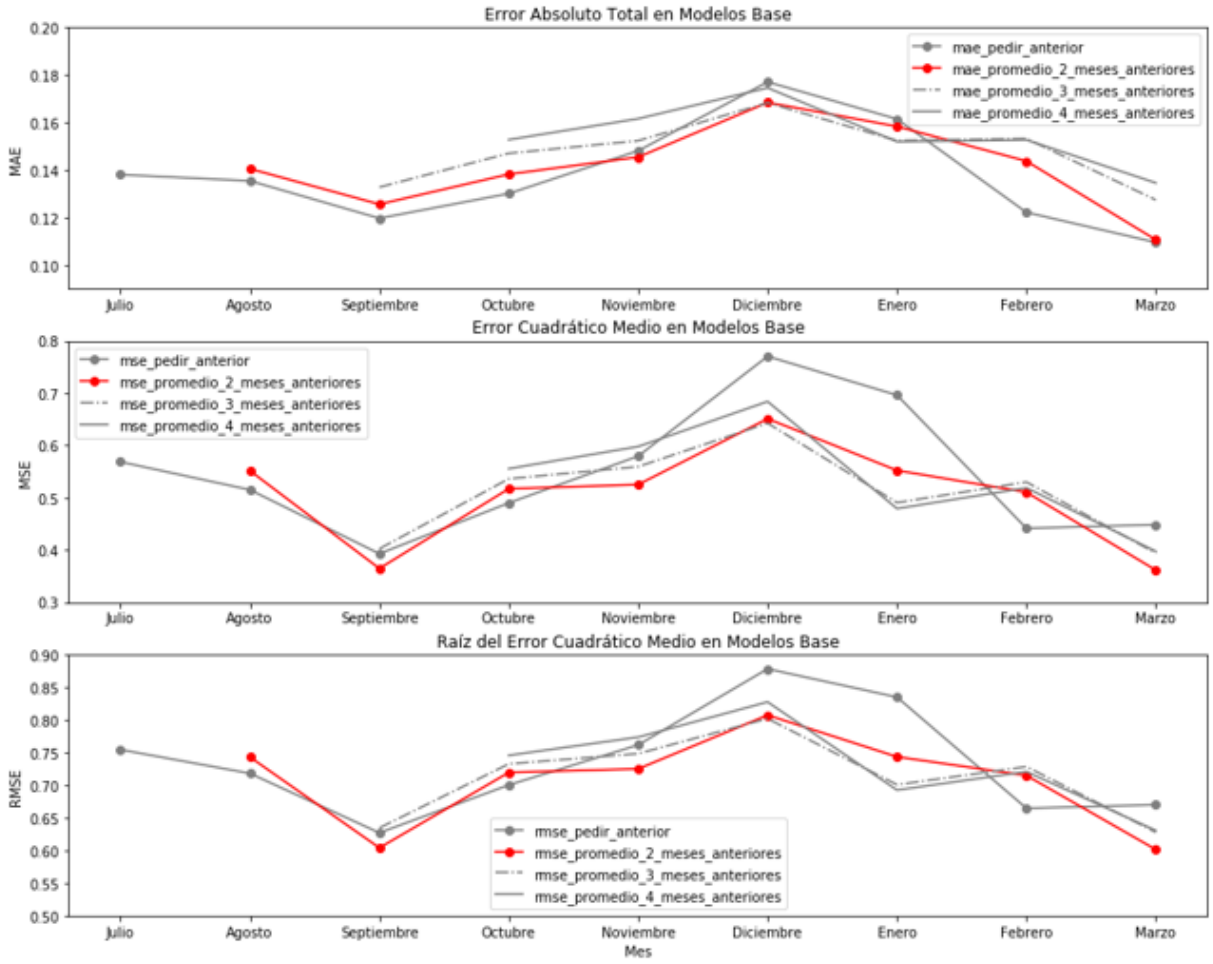


Figura 9: Desempeño mensual de los modelos base propuestos.

En la figura se puede observar cómo los modelos tienen diferentes comportamientos dependiendo del mes de predicción; a primera instancia pareciera que el modelo base que pide lo del mes anterior tiene mejor desempeño, sin embargo, su error se dispara en los meses de diciembre y enero; de la misma manera, se puede observar como los modelos que piden el promedio de 3 y 4 meses anteriores tienen el peor desempeño al estar constantemente con valores de error más altos que los otros dos modelos, por lo tanto, el modelo con el mejor desempeño es el que considera el **promedio de los dos meses anteriores**, al ser el modelo más estable. **Este es el comportamiento que se espera mejorar con los modelos de aprendizaje de máquina propuestos más adelante**.

5 Validación Cruzada para Series de Tiempo

Ya que se determinó el modelo base; y antes de pasar a la sección de modelos propuestos, es importante mencionar una de las consideraciones más importantes que se tomó en cuenta para realizar este proyecto, dicha consideración es utilizar *Validación Cruzada para Series de Tiempo*.

Cuando los registros tienen una secuencia periódica que respetar, se debe de aplicar una técnica llamada *Validación Cruzada para Series de Tiempo* con el fin de evitar el sobreajuste de los modelos y comprobar su efectividad en observaciones no vistas con anterioridad. La idea general de esta técnica consiste en particionar los datos en varios conjuntos de entrenamiento y prueba, respetando su secuencia temporal [10].

En el caso de este proyecto, se cuenta con 8 particiones, cada una con su respectivos conjuntos de entrenamiento y prueba. A continuación, se muestra el diagrama que ilustra esta técnica con los datos de la compañía ABCD.

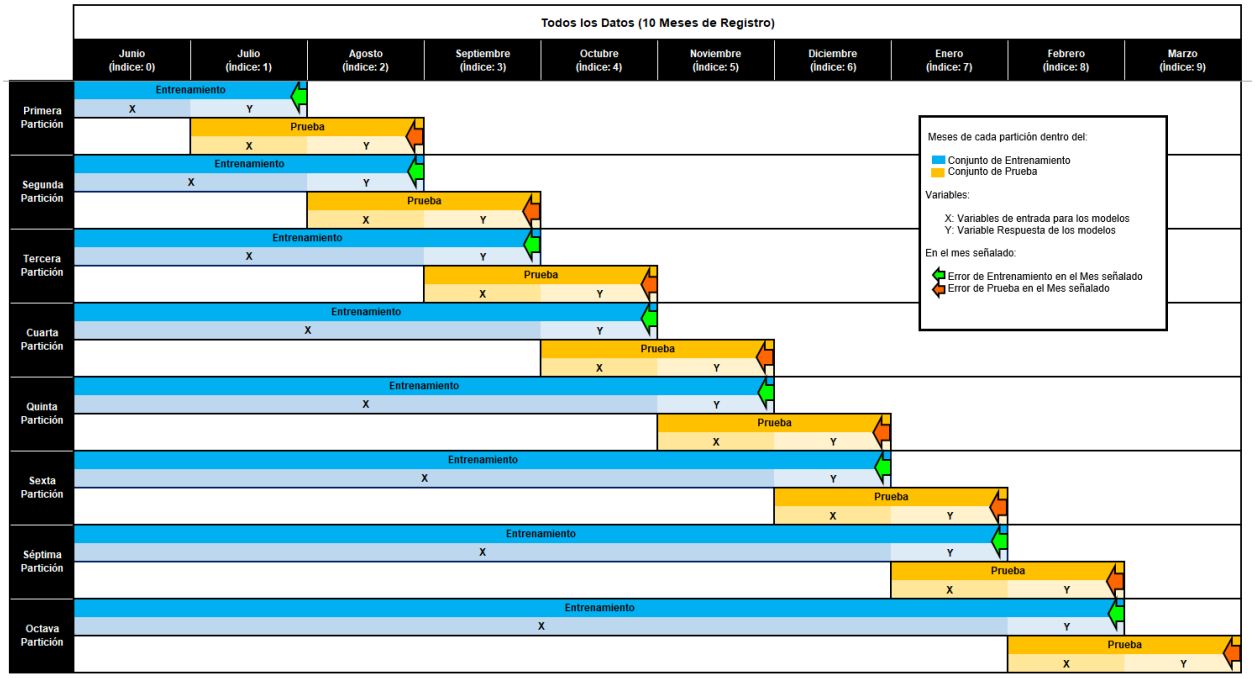


Figura 10: Validación Cruzada para Series de Tiempo.

Por un lado, se puede observar en la figura anterior que con cada mes que pasa, el conjunto de entrenamiento va creciendo, comportamiento que es apropiado, ya que de esta manera, y conforme pasen los meses, los algoritmos propuestos van a tener una mayor ingesta de datos buscando aprender cada vez más de éstos. Por el otro lado, se puede observar que el conjunto de prueba en todas las particiones corresponde a un mes únicamente y éste se va recorriendo conforme pasa el tiempo. Finalmente, se puede observar que en cada partición se van a calcular dos errores, el error de entrenamiento y el error de prueba, sin embargo, el más importante a tomar en cuenta va a ser el de prueba.

A continuación se presenta una gráfica que permite visualizar el número de observaciones que cada partición posee en sus conjuntos de entrenamiento y prueba.

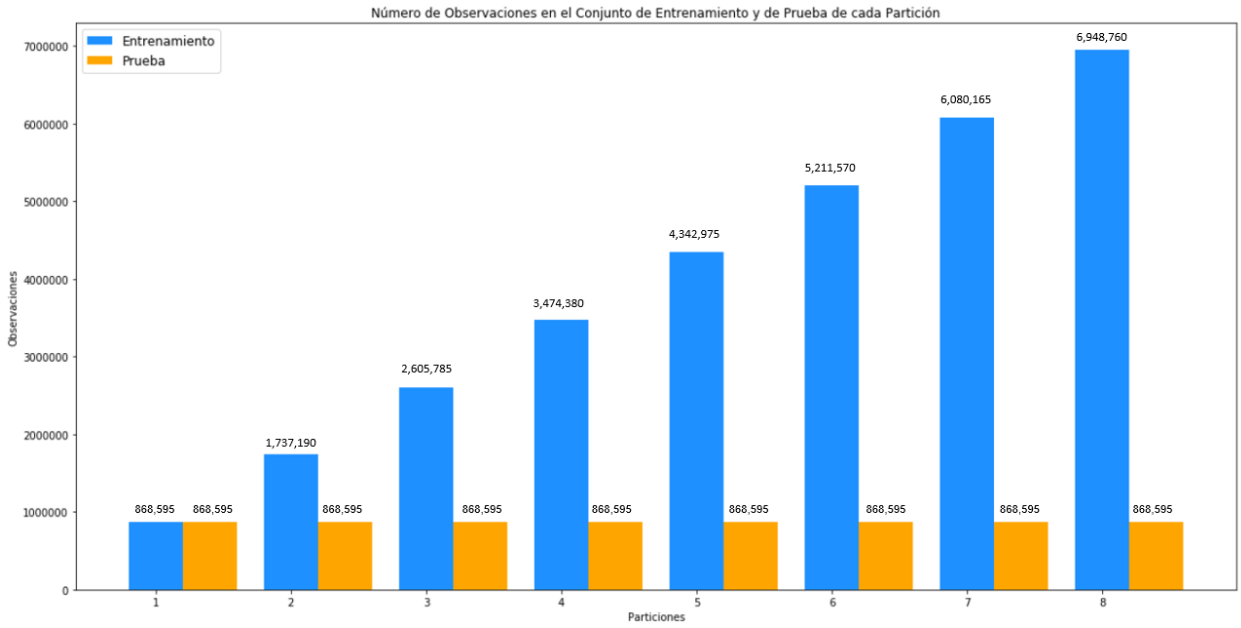


Figura 11: Validación Cruzada para Series de Tiempo.

Las figuras anteriores muestran la estructura de validación cruzada que se aplica al problema en general, pero en caso de que su representación no sea tan clara, a continuación se extrae la última partición y se explica con más detalle:

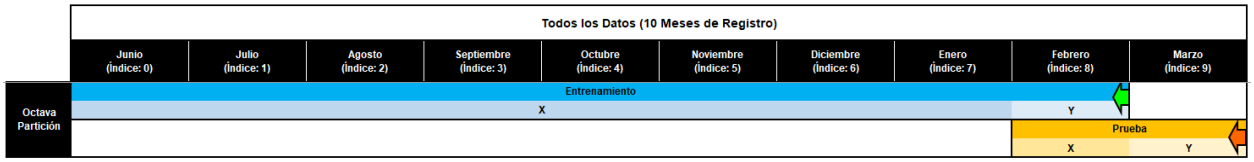


Figura 12: Validación Cruzada para Series de Tiempo de la última Partición de los Datos.

En esta partición se puede observar:

- **Número de registros:** Esta partición cuenta con 6,948,760 registros en el conjunto de entrenamiento, y 868,565 registros en el conjunto prueba.
- **X Entrenamiento:** Las variables de entrada (33) para entrenar el modelo que abarcan los registros de junio a enero.
- **Y Entrenamiento:** La variable de salida con la que se va a evaluar el conjunto de entrenamiento, corresponde a las ventas del mes de febrero (se calcula error de entrenamiento en el mes de febrero).
- **X Prueba:** Las variables de entrada (33) del conjunto de prueba que abarcan los registros del mes de febrero.
- **Y Prueba:** La variable de salida con la que se va a evaluar el conjunto de prueba corresponde a las ventas del mes de marzo (se calcula error de prueba en el mes de marzo).

6 Propuestas de Modelos de Aprendizaje de Máquina

En este punto, el problema ya fue definido, los datos se limpiaron, se aplicó ingeniería de características, se tiene un desempeño base a superar y, ahora sí, se puede proceder a la implementación de propuestas de modelos de aprendizaje de máquina para abarcar la problemática .

6.1 ¿Por qué aplicar modelos de aprendizaje de máquina?

En primer lugar, *aprendizaje de máquina* se refiere a utilizar métodos computacionales que puedan aprender de los datos con el fin de producir **reglas** para mejorar el desempeño de algunas tareas [12].

Las principales razones por las que se proponen modelos computacionales son:

1. Al aplicar un modelo computacional este podría obtener una respuesta barata, rápida y automatizada con la precisión necesaria.
2. Para mejorar el desempeño actual.
3. Para identificar variables o patrones importantes y así entender mejor las problemáticas.

En este caso, se quiere intentar mejorar el desempeño actual de la compañía ABCD, al utilizar algún modelo de aprendizaje de máquina que permita mejorar la predicción de unidades a vender en cada punto de venta al siguiente mes de registro.

6.2 Modelos

Dado el planteamiento que se hizo en la sección de *Descripción de la Problemática*, en particular, la parte que describe el problema en términos de aprendizaje de máquina, se propone aplicar dos tipos de modelos que permitan lidiar con este problema de regresión, los modelos propuestos son: **Árboles de Decisión** y **Bosques Aleatorios**.

A continuación, se presenta una breve descripción de los algoritmos propuestos, seguido por la forma en la que se contruyeron en código y su modo de empleo, para después mostrar y analizar sus respectivos resultados.

6.2.1 Árboles de Decisión

De forma general, un árbol de decisión es un modelo de aprendizaje supervisado que puede tratar problemas tanto de regresión como de clasificación.

A continuación, se presenta una sencilla visualización de un árbol de decisión:

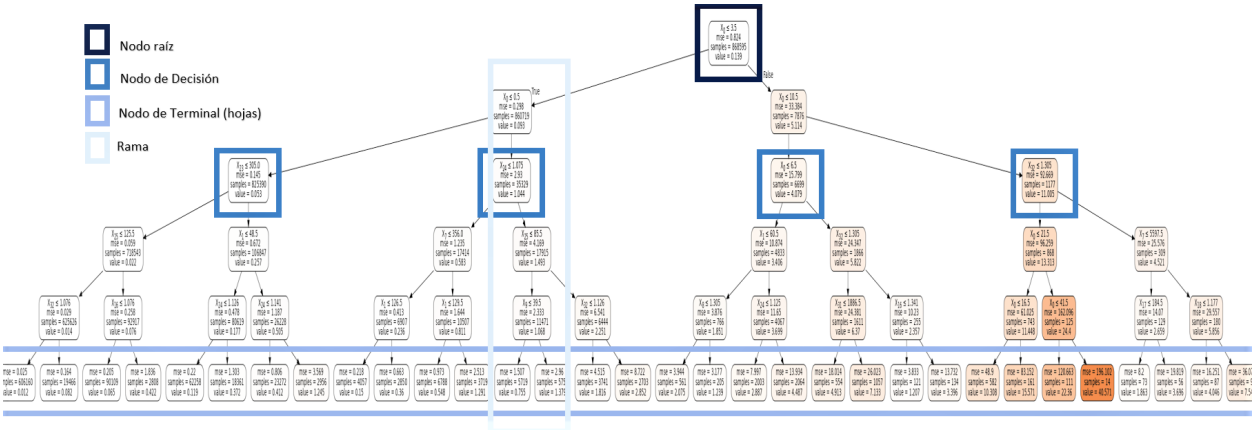


Figura 13: Visualización general de un Árbol de Decisión con máxima profundidad de 5.

La idea básica de estos modelos es buscar puntos de corte en las variables de entrada X (crear nodos de decisión en el árbol) para hacer predicciones (llegar a un nodo final que indique el valor que va a tomar la variable $\hat{y}$) [13].

Para este proyecto, se propone utilizar este modelo cambiando únicamente uno de sus parámetros: **Profundidad del árbol**. Esta característica dentro de un árbol de decisión corresponde a la longitud máxima que puede alcanzar una rama dentro del árbol; los valores que se le van a asignar en este proyecto son: **1**, **5** y **sin profundidad establecida**, es decir, el modelo en sí determinará hasta cuando el árbol deja de crecer (tiende al sobreajuste).

6.2.2 Bosques Aleatorios

Al igual que los árboles de decisión, los bosques aleatorios son modelos de aprendizaje supervisado que puede abarcar tareas tanto de predicción como de clasificación.

A continuación se presenta una visualización simplificada de un bosque aleatorio:

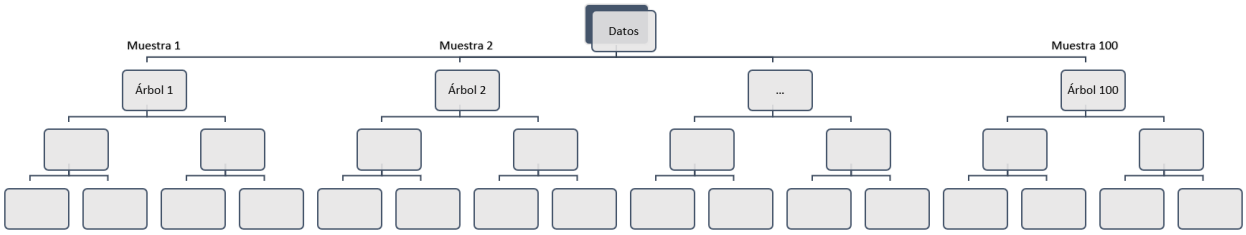


Figura 14: Visualización general de un bosque aleatorio con 100 árboles.

De forma general, los bosques aleatorios crean varios árboles de decisión y los combina buscando tener una predicción más precisa y estable, esta construcción se genera al perturbar la muestra de entrenamiento de distintas maneras y así construir árboles distintos [14].

Para este proyecto se propone utilizar este modelo cambiando únicamente uno de sus parámetros: **Número de árboles**. Los valores que se van a tomar en este caso son: **100 árboles** y **500 árboles**.

En total, este proyecto va a construir 5 propuestas de modelos de aprendizaje de máquina: 3 árboles de decisión y 2 bosques aleatorios.

6.3 Construcción de los modelos

Ya que se tiene en mente el concepto general de los modelos que se van a abordar, se procede con su implementación en código, implementación que puede consultarse en la liga referente a la carpeta de *Modelado* en el repositorio de github (La liga puede consultarse en la sección de anexos - anexo 1). Tanto los árboles de decisión como los bosques aleatorios contruidos para este proyecto, siguen una estructura de implementación en código similar (Jupyter Notebook).

1. Exportar librerías necesarias para correr los diferentes algoritmos.
2. Leer los datos.
3. Hacer los últimos arreglos para tener los datos en el formato correcto.

- Eliminar los registros correspondientes al último mes de registro (9:marzo).
- Eliminar las columnas de *gamma* y *marca* ya que el ID del producto (skuID) lo incluye implícitamente.
- Indexar las columnas de: *tiendaID*, *skuID*, y *fechaID*.

Finalmente, se tiene un archivo con **7,817,355 registros** y **34 columnas** (33 variables de entrada y 1 variable respuesta). En caso de querer ver el nombre de las 34 variables que se utilizaron en los modelos, estas se pueden consultar en el anexo 3 al final de este documento.

4. Realizar las 8 particiones de los datos con sus respectivos conjuntos de de entrenamiento y prueba para poder aplicar validación cruzada.
5. Entrenar los 8 modelos distintos con los diferentes conjuntos de entrenamiento.
6. Evaluar el desempeño de los modelos con el conjunto de entrenamiento. Si se desean ver estos resultados, se pueden consultar en la sección de anexos (anexo 4).
7. Hacer predicciones con el conjunto de prueba para de esta manera evaluar (con las tres métricas) las predicciones contra los valores reales (error de prueba).

6.4 Modo de empleo

El modo de empleo de los modelos es sumamente sencillo, lo único que se debe de hacer es cargar los datos limpios y completos en formato csv y después correr los códigos que se encuentran en un Jupyter Notebook dentro del repositorio del proyecto en github (la liga se encuentra en la sección de anexos, anexo 1).

Los modelos en sí reciben un *dataframe* con 33 columnasy éstos tienen como valores de salida tantos valores como registros tenga el dataframe, es decir, por cada arreglo de 33 valores correspondientes a las variables que requiere el modelo, éste va a explulsar un valor $\hat{y}$, donde $\hat{y}$ corresponde a la variable de salida $\hat{Y_{i,j,k+1}}$, es decir, la predicción del número de unidades del producto con índice i en el punto de venta con índice j que se van a vender en el mes siguiente del mes de registro k.

Por ejemplo, todos los conjuntos de prueba cuentan con 868,565 registros, por lo tanto, los modelos ya entrenados van a recibir estos registros como un dataframe y el modelo va a expulsar 868,565 valores correspondientes a cada una de las predicciones de unidades a vender de cada producto, en cada tienda, al siguiente mes de registro (1,909 puntos de venta x 455 productos distintos x 1 mes = 868,595 valores de salida).

6.5 Resultados de los Modelos propuestos

6.5.1 Resultados: Bosques Aleatorios

Los primeros resultados que se muestran son los relacionados con los modelos de **bosques aleatorios**. En la siguiente figura se muestran las gráficas correspondientes al **Error de Prueba** de cada partición de validación cruzada en ambos bosques aleatorios (100 y 500 árboles).

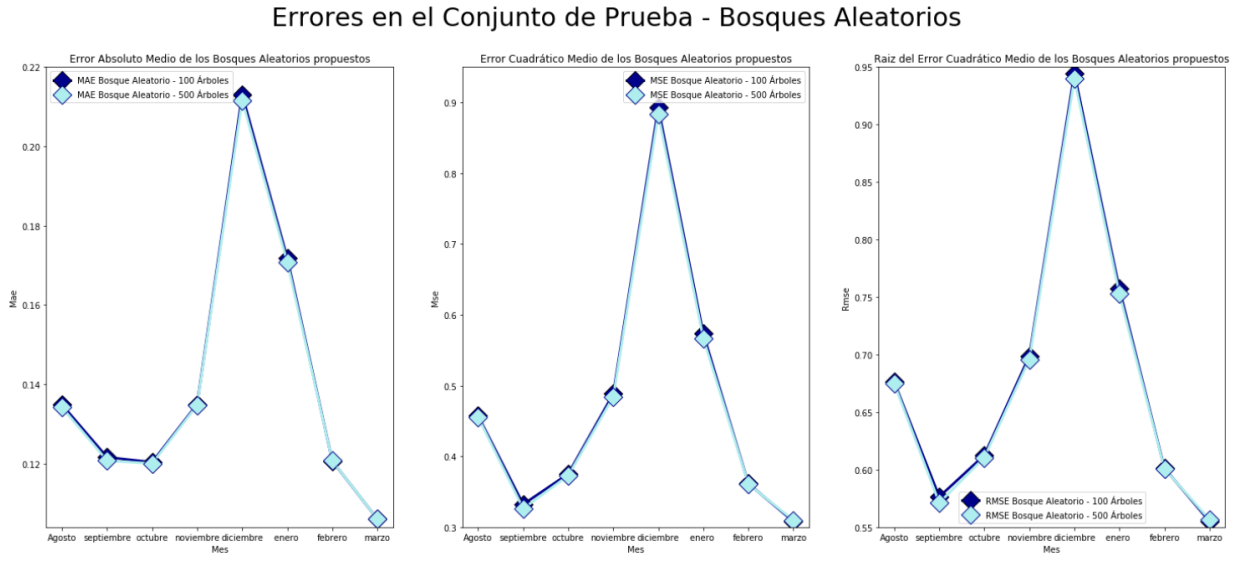


Figura 15: Gráfica de los Errores de prueba (por mes) de los dos Bosques Aleatorios propuestos.

En la primera gráfica se puede observar el *Error Absoluto Medio* que tuvieron los bosques aleatorios en los meses de agosto 2018 a marzo 2019. De esta gráfica puede recalcarse que el comportamimento de ambos árboles es muy similar, de hecho, la diferencia que hay en sus valores de error son mínimas.

En la siguiente gráfica, se observa el comportamiento que tienen los dos modelos con respecto el *Error Cuadrático Medio*, al igual que en el caso anterior, se puede ver como el comportamiento de ambos en tanto a esta métrica es prácticamente el mismo. Finalmente, en la última gráfica se puede observar la *Raíz del Error Cuadrático Medio*, donde una vez más se puede apreciar que el comportamiento de los modelos no difiere notoriamente.

A continuación, se presentan las tablas con los valores númericos con los que se contruyeron las gráficas pasadas, donde se resaltan con letras más oscuras los valores de error más pequeños dentro de cada métrica

Año	Mes	MAE		MSE		RMSE	
		RF 100 Árboles	RF 500 Árboles	RF 100 Árboles	RF 500 Árboles	RF 100 Árboles	RF 500 Árboles
2018	Agosto	0.1348170	**0.1341560**	0.4570430	**0.4552280**	0.6760500	**0.6747059**
	Septiembre	0.1216010	**0.1207540**	0.3321710	**0.3260670**	0.5763430	**0.5710228**
	Octubre	0.1203590	**0.1200210**	0.3751590	**0.3726620**	0.6125020	**0.6104605**
	Noviembre	0.1349920	**0.1347430**	0.4881390	**0.4840520**	0.6986690	**0.6957385**
	Diciembre	0.2130520	**0.2115690**	0.8923170	**0.8830010**	0.9446250	**0.9396813**
2019	Enero	0.1718510	**0.1708110**	0.5732460	**0.5667470**	0.7571300	**0.7528260**
	Febrero	**0.1207040**	0.1207490	0.3614420	**0.3607920**	0.6012000	**0.6006596**
	Marzo	**0.1060450**	**0.1060450**	**0.3080350**	0.3098310	**0.5550090**	0.5566246

Figura 16: Errores de prueba (por mes) de los dos Bosques Aleatorios propuestos.

Con respecto al *Error Absoluto Medio (MAE)*, se observa que el bosque aleatorio con 500 árboles tiene mejor desempeño en los meses de agosto 2018 a diciembre 2019, mientras que el bosque aleatorio de 100 árboles tiene un error menor en el mes de febrero; y ambos tienen el mismo error en el último mes de registro (marzo 2019).

En relacion con el *Error Cuadrático Medio (MSE)* y la *Raíz del Error Cuadrático medio (RMSE)*, se puede observar que el bosque aleatorio con 500 árboles tiene mejor desempeño en los meses de agosto del 2018 a febrero del 2019, dejando el último mes de registro como el único mes en el que el bosque con 100 árboles tuvo un error menor.

En conjunto, las gráficas y las tablas anteriores permiten tener una apreciación más detallada de los resultados que tuvieron estos modelos. Finalmente, se puede decir que, de los dos modelos de bosques aleatorios propuestos, el que tuvo mejor desempeño en general fue el que esta construido con 500 árboles.

6.5.2 Resultados: Árboles de Decisión

Los siguientes resultados correspoden a los modelos de **Árboles de Decisión**. La siguiente figura muestran las gráficas correspondientes al **Error de Prueba** de cada partición de validación cruzada en los tres árboles de decisión propuestos (1 de profundidad, 5 de profundidad y sin profundidad establecida).

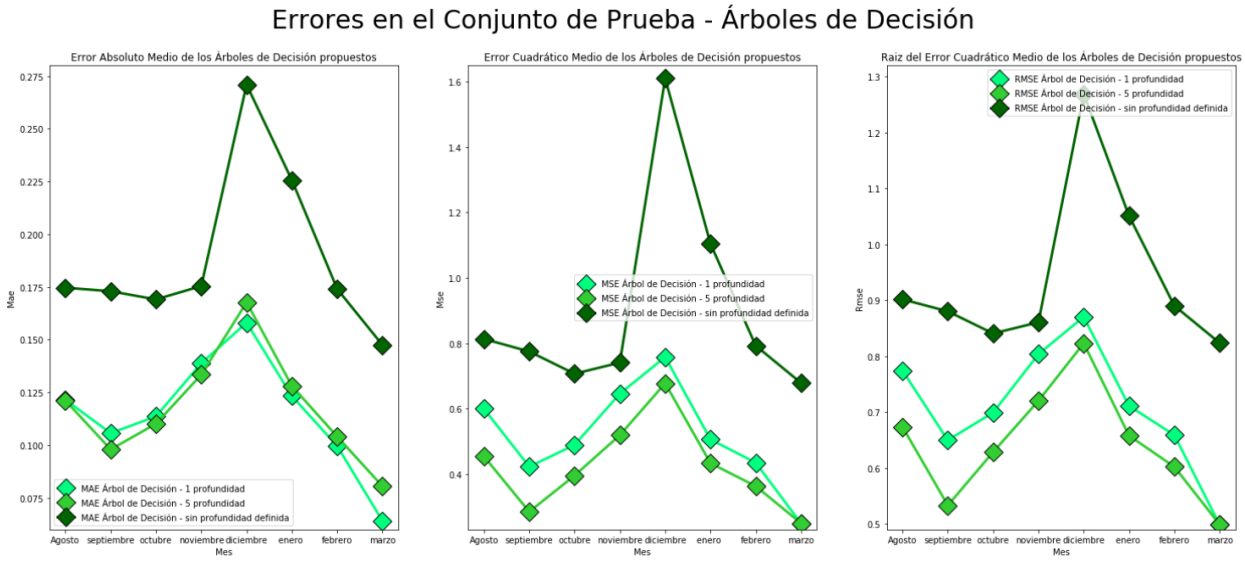


Figura 17: Gráfica de Errores de prueba (por mes) de los tres Árboles de Decisión propuestos.

A primera instancia, se puede observar que los resultados de los árboles varian mucho más entre ellos que los bosques aleatorios, sin embargo, esa comparación se hace en la siguiente sección de este documento y se enuncia con más detalle.

Mientras que en la primera gráfica se puede observar el *Error Absoluto Medio (MAE)* que tuvieron los tres árboles de decisión en los meses de agosto 2018 a marzo 2019, aquí se puede observar que los mejores dos modelos son los que tienen profundidad de 1 y de 5.

En la siguiente gráfica, se observa el comportamiento que tienen los tres modelos con respecto el *Error Cuadrático Medio (MSE)*, donde se observa que el modelo sin profundidad establecida es el que tiene el peor desempeño, mientras que el modelo con profundidad de 5 tiene el mejor, al ser el que presenta los valores de error más pequeños en todos los meses considerados.

Finalmente, en la última gráfica se puede observar la *Raíz del Error Cuadrático Medio (RMSE)*, donde al igual que el caso anterior, el mejor modelo es el que tiene profundidad de 5.

Como apoyo a la visualización anterior, a continuación, se presentan las tablas con los valores númericos con los que se contruyeron las gráficas. Al igual que en las tablas correspondientes a los resultados de los bosques aleatorios, las tablas de esta sección resaltan con letras más oscuras los valores de error más pequeños dentro de cada métrica.

Año	Mes	MAE			MSE			RMSE		
		DT 1 Profundidad	DT 5 Profundidad	DT sin profundidad	DT 1 Profundidad	DT 5 Profundidad	DT sin profundidad	DT 1 Profundidad	DT 5 Profundidad	DT sin profundidad
2018	Agosto	0.121488	**0.121142**	0.174628	0.599140	**0.453430**	0.812912	0.774041	**0.673372**	0.901616
	Septiembre	0.105638	**0.097937**	0.172937	0.422051	**0.284220**	0.774633	0.649655	**0.533123**	0.880132
	Octubre	0.113611	**0.110120**	0.169059	0.488702	**0.394911**	0.706527	0.699072	**0.628419**	0.840552
	Noviembre	0.138805	**0.133387**	0.175417	0.645559	**0.518567**	0.740580	0.803467	**0.720116**	0.860570
	Diciembre	**0.158025**	0.167674	0.270893	0.757501	**0.677320**	1.610366	0.870345	**0.822995**	1.269002
2019	Enero	**0.123257**	0.128010	0.225594	0.505446	**0.432744**	1.104388	0.710947	**0.657833**	1.050899
	Febrero	**0.099547**	0.104010	0.174006	0.434161	**0.362465**	0.792269	0.658909	**0.602051**	0.890095
	Marzo	**0.064064**	0.080479	0.147435	**0.248325**	0.248875	0.678285	**0.498322**	0.498874	0.823581

Figura 18: Errores de prueba (por mes) de los tres Árboles de Decisión propuestos.

Con respecto al *Error Absoluto Medio (MAE)*, se observa con más claridad que los modelos con mejor desempeño son los árboles con *1 y 5 de profundidad*. El comportamiento más interesante

de estos modelos dentro de esta métrica es que, de los 8 meses evaluados, los primeros cuatro tienen mejores resultados con el árbol que tiene profundidad de 5, mientras que los últimos 4 meses tienen mejor desempeño con el árbol de 1 de profundidad. Por lo tanto, hay que analizar el comportamiento de los modelos en las demás métricas para poder determinar el modelo con el mejor desempeño en general.

En relación con el *Error Cuadrático Medio (MSE)* y la *Raíz del Error Cuadrático Medio (RMSE)*, se puede observar que el árbol con 5 de profundidad tiene el mejor desempeño de los meses de agosto de 2018 a febrero de 2019, mientras que el árbol con 1 de profundidad tiene un error menor únicamente en el último mes de registro (marzo 2019).

En conjunto, las gráficas y las tablas anteriores permiten tener una apreciación más detallada en tanto a los resultados que tuvieron estos modelos. Finalmente, se puede decir que de los tres modelos de árboles de decisión propuestos, el que tuvo mejor desempeño fue el que esta construido con una profundidad máxima de 5.

7 Resultados finales

En esta sección se presentan los resultados finales del proyecto, es decir, se contrasta el modelo base contra los modelos propuestos, con el fin de analizar el desempeño de estos y ver si se logró contruir un modelo de aprendizaje de máquina que presente un error de predicción menor al modelo en uso actualmente por la compañía ABCD.

7.1 Modelo Base vs. Modelos Propuestos

La siguiente figura corresponde a la visualización de los errores de prueba del modelo base y de los 5 modelos de aprendizaje de máquina propuestos.

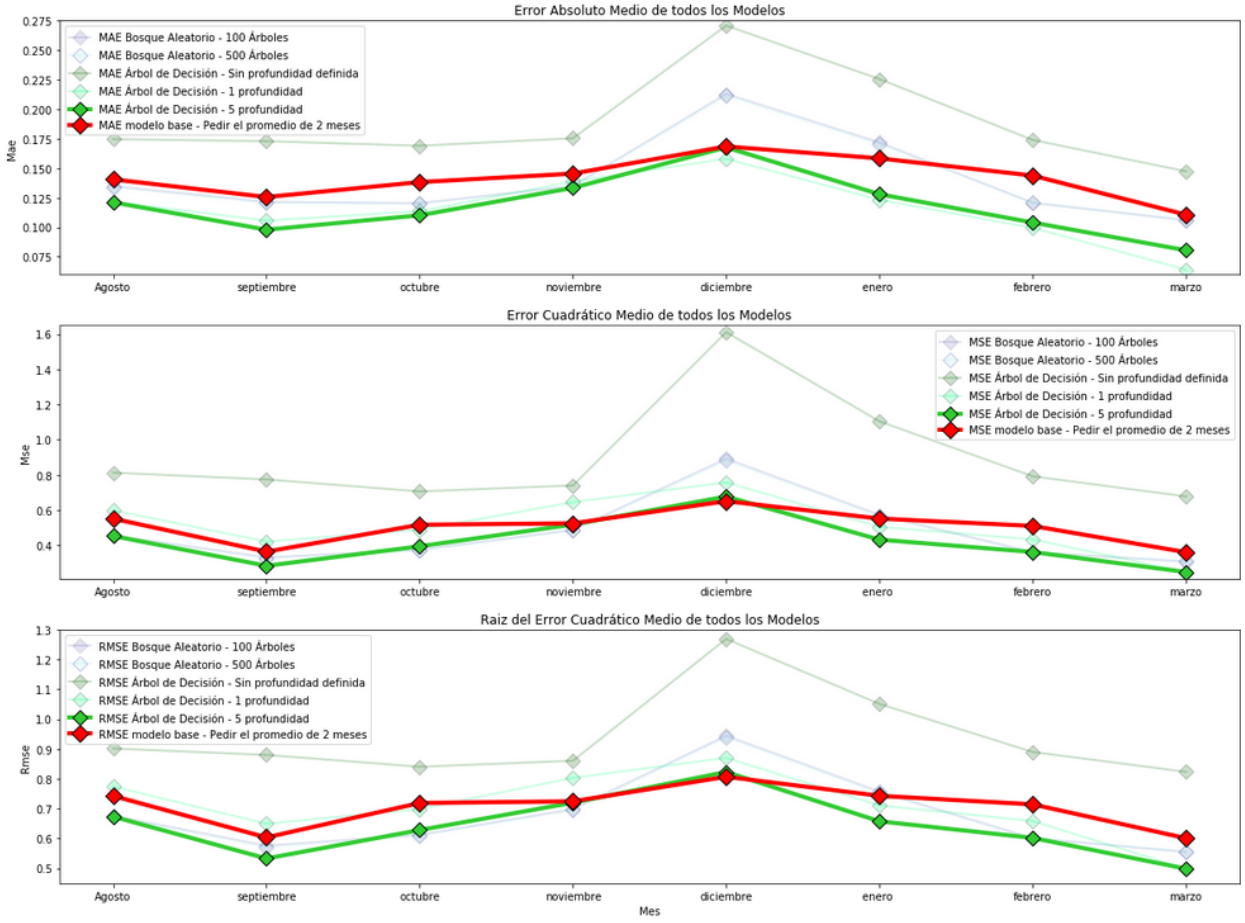


Figura 19: Gráficas de errores de prueba de los modelos propuestos vs. el modelo base.

De forma general, se puede observar que el *árbol de decisión sin profundidad establecida* es el que presenta el mayor error de todos los modelos en las tres métricas. Seguido, se encuentran *ambos bosques aleatorios* que, aunque lograron obtener un menor error al modelo base en algunos meses (los primeros 4 meses y los últimos dos), el error que obtuvieron en los meses de diciembre y enero es mayor en comparación con el modelo base. Finalmente, se tienen los otros *dos árboles de decisión*, cuyo desempeño varia dependiendo del mes y la métrica; el análisis que se detalló en la sección anterior concluyó que el mejor modelo fue el *árbol con 5 de profundidad*.

Por lo tanto, de los 5 modelos propuestos en este proyecto, el que tuvo mejor desempeño fue el **árbol de decisión con profundidad de 5**, modelo que se selecciona para contrastar con el modelo base. Es por eso que en las tres gráficas anteriores se resaltan con una mayor densidad de color: El modelo base (línea roja) y el modelo de aprendizaje de máquina que tuvo el mejor desempeño (línea verde), mientras que los demás modelos se grafican con un degradado de color.

La primera gráfica muestra el desempeño que tuvieron los modelos con relación al *Error Absoluto Medio (MAE)*, donde se puede observar que en casi todos los meses, el *árbol de decisión con un valor de profundidad de 5* tiene un error menor que el modelo base (pedir el promedio de los dos meses anteriores de registro); el único mes donde no se alcanza a apreciar en su totalidad cuál es el mejor modelo es en el mes de diciembre.

Las siguientes dos gráficas correspoden al *Error Cuadrático Medio (MSE)* y a la *Raíz del Error Cuadrático Medio (RMSE)*, en estas gráficas se puede observar que el árbol de decisión con profundidad de 5 tiene el menor error en casi todos los meses; los meses donde no se puede apreciar en su totalidad el desempeño de ambos modelos es en los meses de noviembre y diciembre.

Como apoyo a la visualización anterior, a continuación, se presentan las tablas con los valores númericos de los dos modelos a comparar, resaltando los valores con el menor error dentro de cada métrica.

Año	Mes	MAE		MSE		RMSE	
		Modelo Base	DT 5 Profundidad	Modelo Base	DT 5 Profundidad	Modelo Base	DT 5 Profundidad
2018	Agosto	0.140563	**0.121142**	0.551698	**0.453430**	0.742764	**0.673372**
	Septiembre	0.125616	**0.097937**	0.364511	**0.284220**	0.603748	**0.533123**
	Octubre	0.138270	**0.110120**	0.517241	**0.394911**	0.719195	**0.628419**
	Noviembre	0.145481	**0.133387**	0.525178	**0.518567**	0.724692	**0.720116**
	Diciembre	0.168484	**0.167674**	**0.651252**	0.677320	**0.807002**	0.822995
2019	Enero	0.158390	**0.128010**	0.552302	**0.432744**	0.743170	**0.657833**
	Febrero	0.143794	**0.104010**	0.510564	**0.362465**	0.714538	**0.602051**
	Marzo	0.110653	**0.080479**	0.361328	**0.248875**	0.601106	**0.498874**

Figura 20: Errores de prueba (por mes) del modelo base vs. el mejor modelo propuesto.

Con esta información se puede determinar que, con respecto al *Error Absoluto Medio*, el árbol de decisión con 5 de profundidad tiene el menor error en todos los meses considerados; mientras que en las *otras dos métricas*, el modelo base es superado en casi todos los meses, menos en el mes de diciembre.

> **Este análisis indica que el objetivo de proponer un modelo que mejore el desempeño de la compañía a la hora de hacer prediciones fue cumplido, ya que el árbol de decisión con 5 de profundidad genera un error menor en casi todos los meses de análisis.**

8 Conclusiones y Recomendaciones

Este proyecto desarrolló un proyecto de ciencia de datos aterrizado a una problemática real, utilizando datos de una empresa enfocada a la venta y distribución de productos de telefonía celular.

El objetivo del proyecto fue proponer modelos de aprendizaje de máquina que intentaran mejorar la predicción mensual de unidades a vender en los diferentes puntos de venta de la compañia; se asumió que el modelo empleado actualmente por la empresa correspondía a pedir el promedio de los dos meses anteriores de registro y este era el desempeño que se buscaba mejorar. Dicho objetivo se cumplió con un árbol de decisión con una profundidad de 5, que resultó en obtener un error de predicción menor en todos los meses bajo la métrica del *Error Absoluto Medio (MAE)*, y obtuvo un error menor en 7 de 8 meses en las métricas del *Error Cuadrático Medio (MSE)* y la *Raíz del Error Cuadrático Medio (RMSE)*.

En relación con el **modelo propuesto** en este proyecto, las recomendaciones finales que se les daría a la empresa ABCD son:

1. Cambiar el modelo empleado actualmente, por el modelo propuesto en este documento, un árbol de decisión con profundidad igual a 5. Ya que con este modelo se logró reducir su error de predicción en los 8 meses de registro proporcionados; y se esperaría que las predicciones de los siguientes meses fueran mejores a si se siguiera con el modelo actual.

2. Tener especial cuidado en los meses de noviembre, diciembre y enero. Estos meses, normalmente, tienen mayor número de ventas en comparación con los demás meses del año, ya que son meses con festividades donde se impulsa la compra de bienes y servicios en general. Al no contar con más registros de ventas, es decir, información de varios años, es más complicado encontrar algún patrón o comportamiento particular en estos meses; es por eso que la siguiente recomendación es recopilar más información que permitan que el modelo aprenda más de una ingesta de datos más robusta.

3. Enriquecer el modelo con información adicional. Adquirir información adicional relacionada con los registros de ventas, ya que la información empleada en este proyecto ni siquiera corresponde a un año de registro continuo.

4. Mantener un registro de ventas limpio. En la parte de anexos (anexo 1) se encuentra la liga con el repositorio del proyecto en github donde se puede consultar el código con todos los pasos que se llevaron a cabo para realizar la limpieza de datos. Mantener un registro de datos limpio y homogéneo permitirá seguir empleando el modelo en la forma en la que se presentó en este documento y, de la misma manera, facilitará su modo de consulta.

En relación con el **trabajo adicional** que podría hacerse para seguir buscando otras soluciones, se hacen las siguientes recomendaciones a la compañía:

1. Al ya lograrse un mejor desempeño con un árbol de decisión, se podría considerar probar más modelos, ya que este proyecto solo considera dos modelos distintos dentro de la gran variedad de modelos que existen que pueden aplicarse en esta misma situación.

2. Buscar información adicional que podría ayudar a enriquecer el modelo propuesto, es decir, considerar más variables con información demográfica de los diferentes puntos de venta, estados o divisiones geográficas.

3. Añadir información relacionada con la empresa. En la primera sección del documento se enlistaron los *supuestos y restricciones* que se debieron de plantear dado que no se contaba con más información relacionada con: tiempo de entrega, capacidad de almacenamiento, etc. El contar con este tipo de información, aunque eleva la dificultad del problema, podría ayudar a obtener mejores resultados.

9 Anexos

9.1 Anexo 1: Herramientas

Las herramientas que se utilizaron para llevar a cabo este proyecto fueron las siguiente:

Watson Studio: Al hacerse el proyecto en las instalaciones de IBM y con información de uno de sus clientes, fue requisito utilizar su plataforma en la nube; plataforma que procura facilitar el manejo de grandes volúmenes de datos, y además, cuenta con herramientas integradas como: Rstudio y Jupyter Notebook.

RStudio: Se utilizó Rstudio sobre Watson Studio para crear y ejecutar los códigos correspondientes a las partes de: limpieza, transformación, ingeniería de características y análisis exploratorio de los datos.

Jupyter Notebook: Se utilizó Jupyter Notebook sobre Watson Studio para crear y ejecutar los códigos correspondientes a las partes de modelado y análisis de resultados.

Github: En la cuenta personal de github de la alumna que presenta el caso se encuentran todos los códigos y resultados del proyecto para su consulta. A continuación, se presentan las ligas a las que se puede acceder para ver todos los códigos que se llevaron a cabo para realizar el proyecto:

- **Repositorio en Github del proyecto:**

  https://github.com/AnaLuisaMasettoHerrera/Caso_de_Titulacion_Maestria

- **Carpeta donde se encuentra la *Limpieza de Datos*:**

  https://github.com/AnaLuisaMasettoHerrera/Caso_de_Titulacion_Maestria/tree/master/CASO/1_LIMPIEZA_TRANSFORMACIÓN_E_INGENIERIA_DE_CARACTERISTICAS

- **Carpeta donde se encuentra el *Análisis Exploratorio de los Datos*:**

  https://github.com/AnaLuisaMasettoHerrera/Caso_de_Titulacion_Maestria/tree/master/CASO/2_ANALISIS_EXPLORATORIO_DE_LOS_DATOS

- **Carpeta donde se encuentran los códigos para la parte de *Modelado*:**

  https://github.com/AnaLuisaMasettoHerrera/Caso_de_Titulacion_Maestria/tree/master/CASO/3_MODELADO

9.2 Anexo 2: Análisis Exploratorio de los Datos

Tras llevar a cabo la limpieza de los datos, se realizó un *Análisis Exploratorio* con el fin de tener una idea más clara de la situación general de la compañía. A estas alturas, se contaba con un csv con 932,963 registros y 13 columnas, aún no se completaba la serie de tiempo, ni se hacían los conteos, rezagos y promedios que se llevaron a cabo en el la parte de *Ingeniería de Características*.

A continuación, se presentan los resultados más relevantes obtenidos en el *Análisis Exploratorio de los Datos*; en caso de querer consultar el código junto con los demás resultados, la liga con la carpeta de github se encuentra en el anexo anterior.

Como análisis base, la siguiente gráfica muestra el comportamiento de las ventas de la compañía a nivel república, donde se puede observar que las ventas tienen tendencia negativa, con incrementos drásticos en ventas en los meses de *noviembre* y *diciembre*, situación que es de esperarse, al igual que la caida en ventas que se presenta después de esos meses.

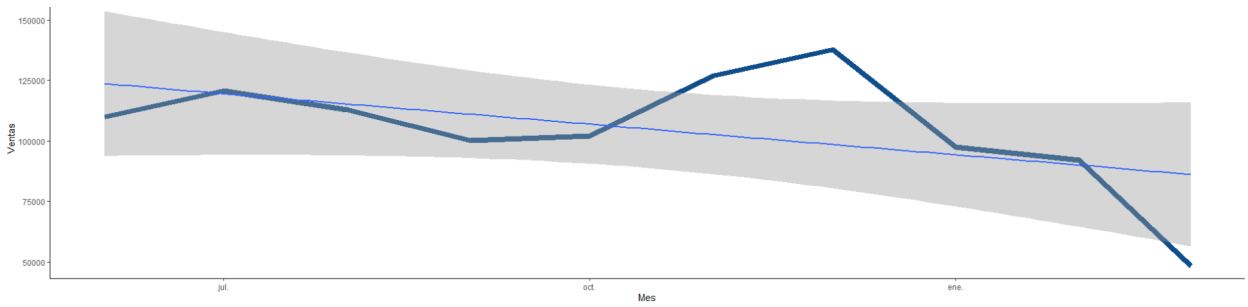


Figura 21: Ventas de los 10 Meses de Registro a Nivel República.

Partiendo de este punto, se procede a hacer un análisis de las ventas más detallado, comenzando por el comportamiento de las ventas a nivel zona geográfica.

La gráfica y el mapa muestran cúales son las zonas que presentaron un mayor número de ventas en 10 meses de registro; se puede observar que la zona correspondiente al *centro sur* es la zona con más ventas, seguido por la zona *centro occidente* y *noroeste*; finalmente, se puede observar que las zonas con menor número de ventas fueron: la *península de Yucatán* y la zona del *pacífico sur*.

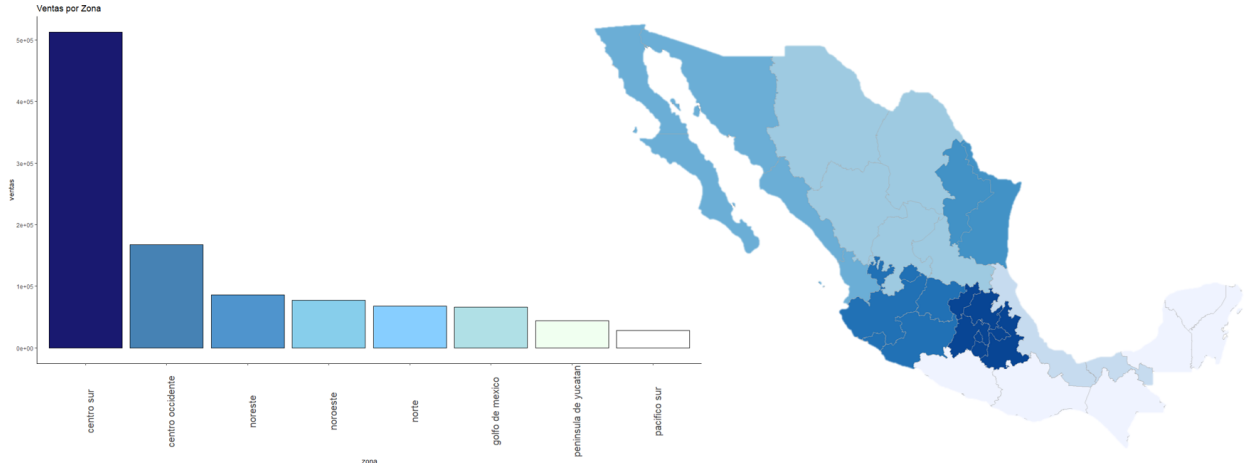


Figura 22: Ventas por Zona geográfica.

Como apoyo adicional a la visualización anterior, enseguida se presenta la tabla con el número de ventas y porcentaje de éstas de cada zona.

Zona	Ventas	Porcentaje
Centro sur	512,223 unidades	48.84944 %
Centro occidente	167,655 unidades	15.98884 %
Noreste	85,779 unidades	8.180531 %
Noroeste	77,573 unidades	7.397945 %
Norte	67,410 unidades	6.428725 %
Golfo de México	66,143 unidades	6.307894 %
Península de Yucatán	43,813 unidades	4.178337 %
Pacifico sur	27,979 unidades	2.668288 %

Figura 23: Número y Porcentaje de Ventas por Zona.

Adicionalmente, se hace el mismo análisis a nivel estado, donde se puede observar que *la ciudad de México*, *el estado de México*, y *Jalisco*, son los estados con mayor número de ventas registradas; mientras que *Durango*, *Baja California Sur* y *Zacatecas* son los estados con menor número de ventas.

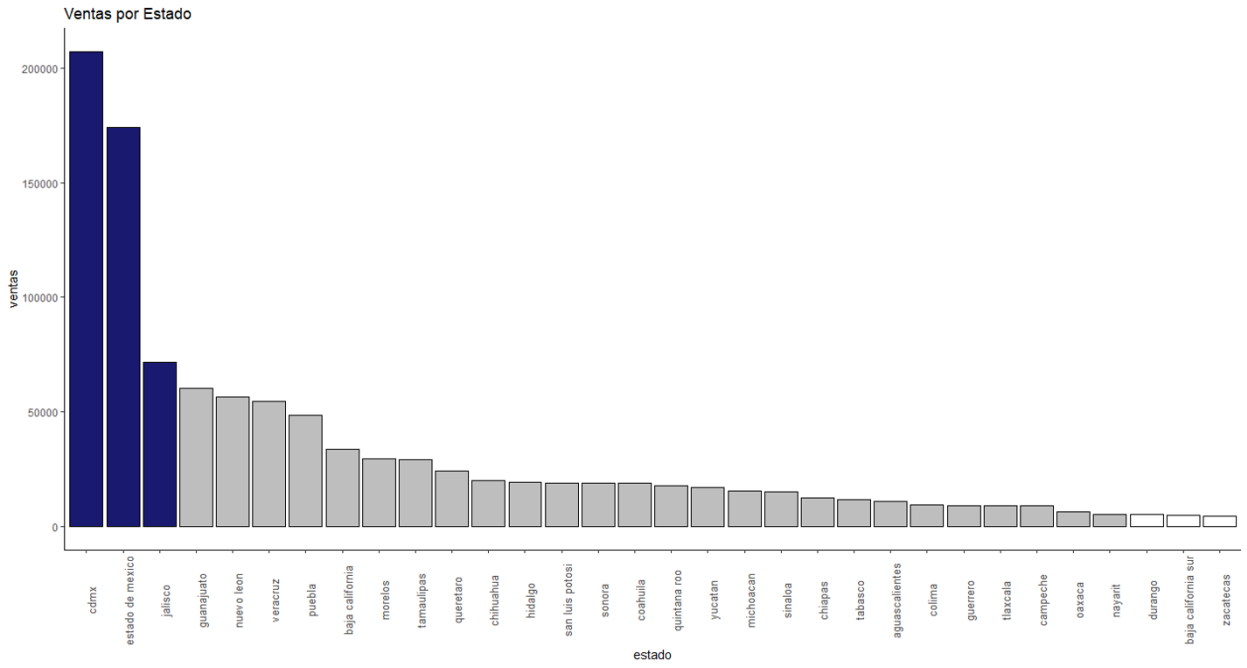


Figura 24: Número de ventas por Estado.

Como segunda parte del análisis, se procede a analizar la situación de la compañía desde un punto de vista de las marcas que esta vende. En primer lugar, se determina el número de productos distintos que cada marca maneja como se muestra en la siguiente imagen.

Marca	Sku's
Apple	129
Huawei	97
Samsung	92
Motorola	60
Hisense	28
Zte	14
Alcatel	13
Sony	10
LG	7
Lanix	2

Figura 25: Productos que maneja cada marca.

En segundo lugar, se determina cuáles son las marcas con más y menos unidades vendidas a nivel república; en la siguiente gráfica se observa que la marca que tuvo más ventas fue *Huawei*, seguida por *Motorola* y *Samsung*; mientras que las marcas con menos ventas fueron *Affix* y *Lenovo*.

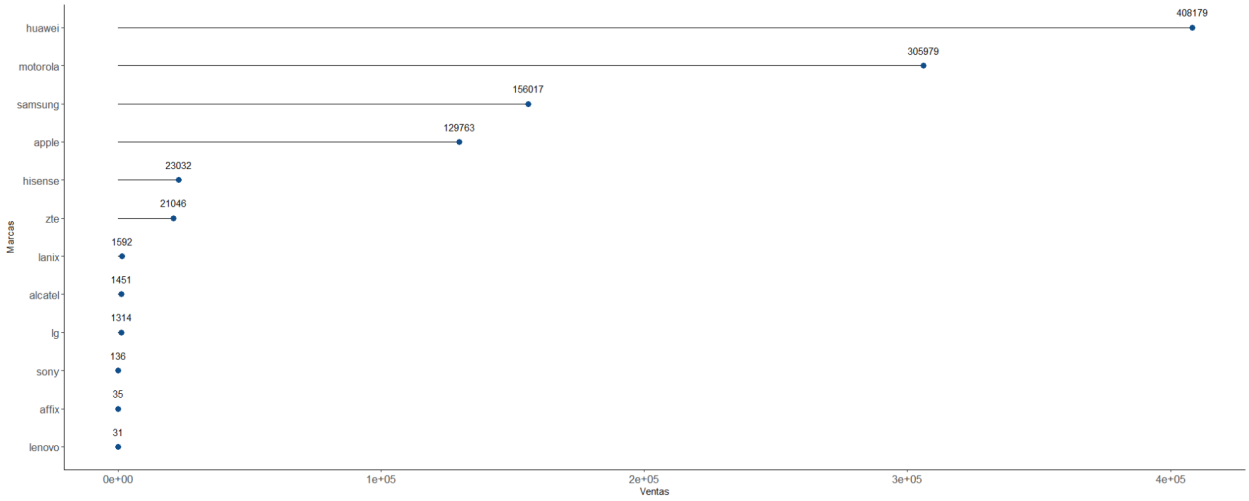


Figura 26: Ventas por Marca.

Finalmente, se hizo un análisis en conjunto, en tanto al número de ventas por marca, por mes en cada uno de los estados del territorio mexicano.

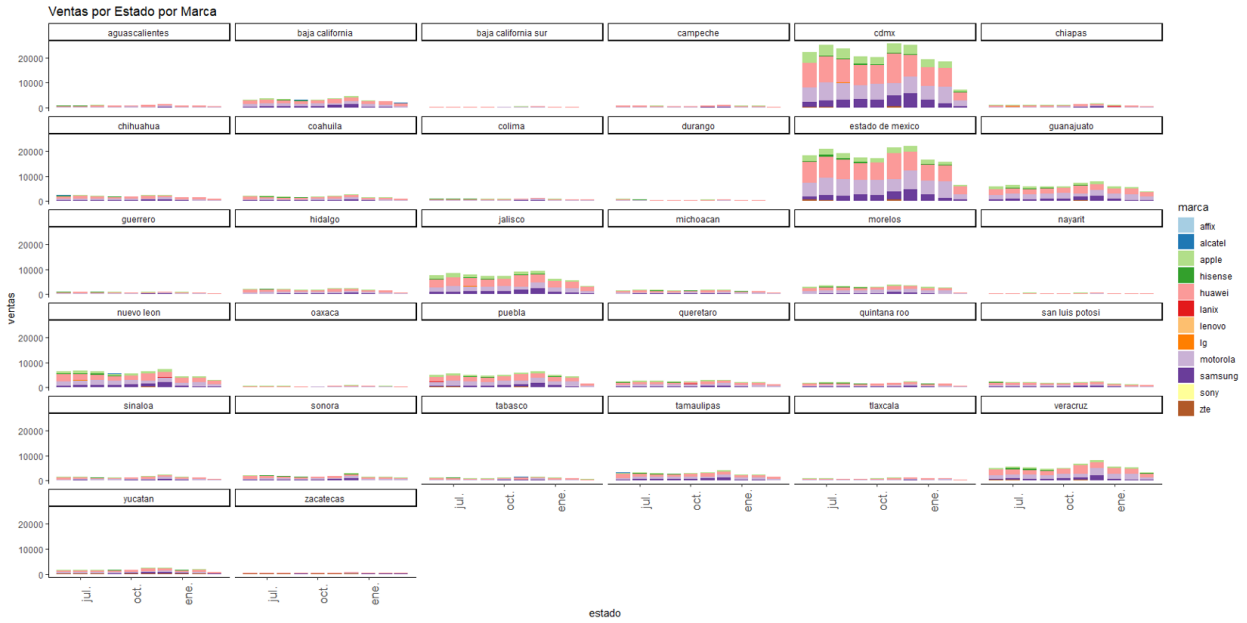


Figura 27: Errores de prueba (por mes) de los tres Árboles de Decisión propuestos.

En la figura anterior, se puede observar que el comportamiento de los meses varia con respecto al mes (eje x), el estado y las marcas; es decir, es una visualización que permite identificar fácilmente cuáles son los estados con mayor y menor número de ventas, cúales son los meses con mayor número de ventas y cómo éstas fluctúan, finalmente, se puede observar que la presencia de las marcas no es la misma.

Con este breve análisis en mente, se puede observar claramente el problema al que se enfrenta la empresa, ya que el número de unidades a vender cada mes es diferente en cada estado (a nivel más desagregado, en cada punto de venta), y las marcas que se venden no son las mismas.

9.3 Anexo 3: Variables empleadas en los modelos de aprendizaje de máquina

Las variables que se emplearon en los algortimos de aprendizaje de máquina son las siguientes:

variable_y_ventas_del_siguiente_mes: Variable respuesta. Variable que indica el número de unidades de cada producto que se deben de vender en cada punto de venta al siguiente mes de registro.

Columnas indexadas:

- TiendaID: Variable con los índices que identifican a cada punto de venta.
- skuID: Variable con los índices que identifican a cada producto.
- PeriodoID: Variable con los índices que identifican a cada periodo del que se tiene registro.

Ventas por mes, por producto, por punto de venta:

1. Ventas por mes: Número de unidades vendidas correspondientes a los valores dentro de las columnas indexadas.

Variables relacionadas con conteos y promedios mensuales:

2. Ventas totales por mes en cada tienda: Unidades totales que vendío cada tienda por mes.
3. Ventas promedio por mes en cada tienda: Promedio de unidades que vendieron en cada tienda por mes.
4. Ventas totales por mes de cada producto: Unidades de cada producto que se vendieron por mes.
5. Ventas promedio por mes de cada producto: Promedio de unidades de cada producto que se vendieron por mes.
6. Ventas totales por mes de cada marca: Unidades de cada marca que se vendieron por mes.
7. Ventas promedio por mes de cada marca: Promedio de unidades de cada marca que se vendieron por mes.
8. Ventas totales por mes de cada gamma: Unidades de cada gamma que se vendieron por mes.
9. Ventas promedio por mes de cada gamma: Promedio de unidades de cada gamma que se vendieron por mes.

Variables con rezagos en el tiempo: A cada una de las 8 variables construidas en el punto anterior se les aplico una fórmula para crear nuevas variables en diferentes rezagos.

10. Ventas totales del **mes anterior** en cada tienda
11. Ventas promedio del mes anterior en cada tienda
12. Ventas totales del mes anterior de cada producto
13. Ventas promedio del mes anterior de cada producto
14. Ventas totales del mes anterior de cada marca
15. Ventas promedio del mes anterior de cada marca

16. Ventas totales del mes anterior de cada gamma
17. Ventas promedio del mes anterior de cada gamma
18. Ventas totales de hace **dos meses** en cada tienda
19. Ventas promedio de hace dos meses en cada tienda
20. Ventas totales de hace dos meses de cada producto
21. Ventas promedio de hace dos meses de cada producto
22. Ventas totales de hace dos meses de cada marca
23. Ventas promedio de hace dos meses de cada marca
24. Ventas totales de hace dos meses de cada gamma
25. Ventas promedio de hace dos meses de cada gamma
26. Ventas totales de hace **tres meses** en cada tienda
27. Ventas promedio de hace tres meses en cada tienda
28. Ventas totales de hace tres meses de cada producto
29. Ventas promedio de hace tres meses de cada producto
30. Ventas totales de hace tres meses de cada marca
31. Ventas promedio de hace tres meses de cada marca
32. Ventas totales de hace tres meses de cada gamma
33. Ventas promedio de hace tres meses de cada gamma

9.4 Anexo 4: Errores de Entrenamiento de los Modelos propuestos

Como se mencionó dentro del documento, una parte muy importante dentro de la construcción de modelos es cuando éstos se entrenan, y al momento de entrenarse es posible calcular el error de entrenamiento.

En la gráfica siguiente, se puede observar el error de entrenamiento que tienen los modelos de **bosques aleatorios**, sin embargo, estos resultados no son del todo confiables, y es por eso que se tiene un error de prueba, ya que al probar el modelo con los mismos datos, el desempeño va a ser mucho mejor.

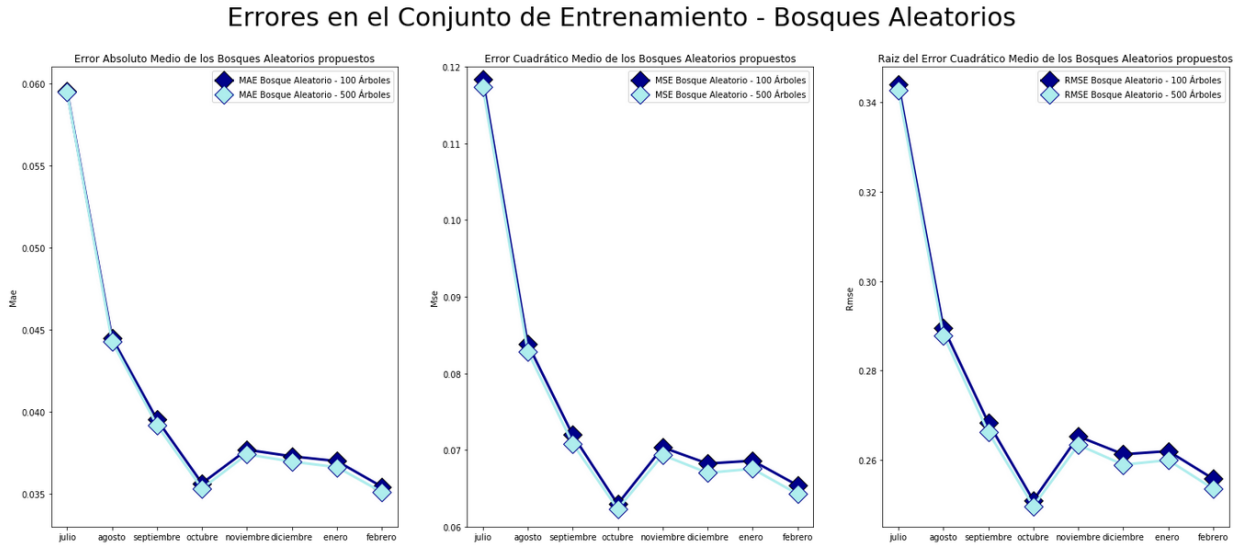


Figura 28: Errores de prueba (por mes) de los dos Bosques Aleatorios propuestos.

En las gráficas anteriores se puede observar que el comportamiento de ambos bosques es muy similar, y la diferencia que hay entre ellos es muy difícil de apreciar a simple vista, es por eso, que se incluye el siguiente conjunto de tablas:

Año	Mes	MAE		MSE		RMSE	
		RF 100 Árboles	RF 500 Árboles	RF 100 Árboles	RF 500 Árboles	RF 100 Árboles	RF 500 Árboles
2018	Julio	0.0595210	**0.0594550**	0.1184210	**0.1174170**	0.3441240	**0.3426616**
	Agosto	0.0444900	**0.0442530**	0.0838470	**0.0828420**	0.2895630	**0.2878229**
	Septiembre	0.0395140	**0.0391690**	0.0719580	**0.0708250**	0.2682500	**0.2661297**
	Octubre	0.0356360	**0.0353170**	0.0629380	**0.0622890**	0.2508740	**0.2495776**
	Noviembre	0.0376780	**0.0374000**	0.0703320	**0.0693290**	0.2652020	**0.2633040**
	Diciembre	0.0372770	**0.0369550**	0.0682220	**0.0670090**	0.2611930	**0.2588610**
2019	Enero	0.0369860	**0.0366200**	0.0685930	**0.0675390**	0.2619030	**0.2598827**
	Febrero	0.0354260	**0.0350980**	0.0653790	**0.0642420**	0.2556930	**0.2534601**

Figura 29: Errores de prueba (por mes) de los dos Bosques Aleatorios propuestos.

Con esto, se pude observar que en tanto al *error de entrenamiento* el mejor modelo es el bosque con 500 árboles.

Con respeto a los otros modelos propuestos, **árboles de decisión**, se presentan en seguida las gráficas que representan el desempeño de los 3 modelos propuestos en el conjunto de entrenamiento.

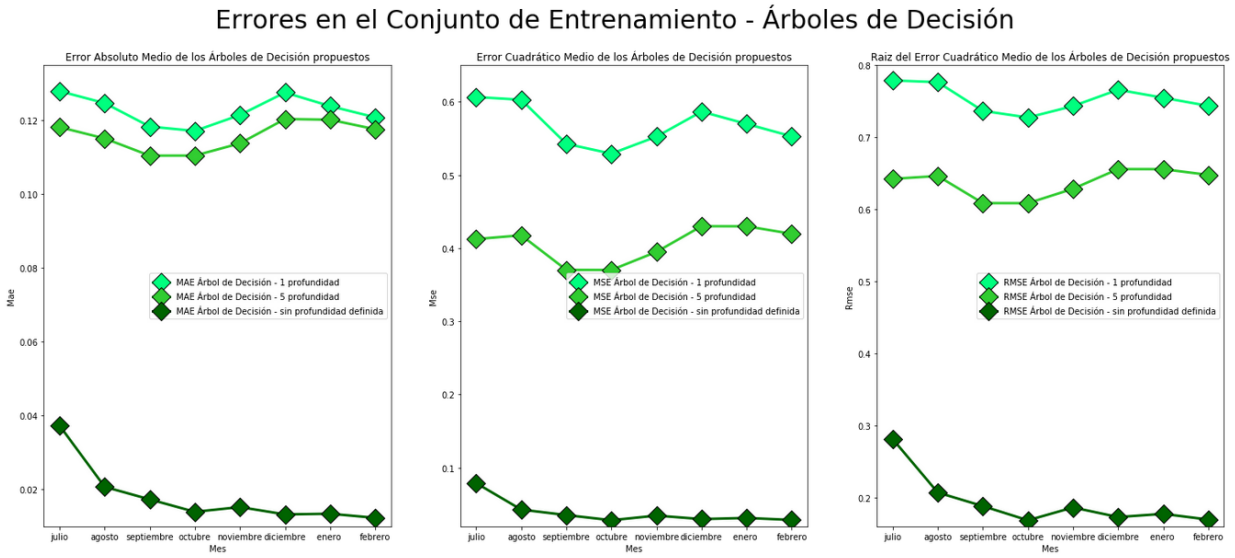


Figura 30: Errores de prueba (por mes) de los tres Árboles de Decisión propuestos.

En las gráficas anteriores, se puede observar que el mejor modelo corresponde al que no se le definió una profundidad máxima, sin embargo, al momento de probar el modelo con los datos de prueba, el resultado fue completamente diferente.

Como apoyo visual adicional a las gráficas anteriores, se presenta el siguiente conjunto de tablas.

Año	Mes	MAE			MSE			RMSE		
		DT 1 Profundidad	DT 5 Profundidad	DT sin profundidad	DT 1 Profundidad	DT 5 Profundidad	DT sin profundidad	DT 1 Profundidad	DT 5 Profundidad	DT sin profundidad
2018	Julio	0.127953	0.118202	**0.037272**	0.606496	0.412475	**0.078994**	0.778779	0.642242	**0.281059**
	Agosto	0.124720	0.115135	**0.020692**	0.602818	0.417577	**0.042806**	0.776414	0.646202	**0.206896**
	Septiembre	0.118360	0.110474	**0.017292**	0.542562	0.370329	**0.035357**	0.736588	0.608547	**0.188035**
	Octubre	0.117173	0.110490	**0.013928**	0.529097	0.370204	**0.028344**	0.727391	0.608444	**0.168357**
	Noviembre	0.121499	0.113808	**0.015218**	0.552389	0.394932	**0.034649**	0.743229	0.628436	**0.186142**
	Diciembre	0.127587	0.120408	**0.013223**	0.586575	0.430007	**0.029886**	0.765882	0.655749	**0.172876**
2019	Enero	0.123908	0.120221	**0.013410**	0.569746	0.429915	**0.031425**	0.754815	0.655679	**0.177271**
	Febrero	0.120863	0.117632	**0.012302**	0.552798	0.419590	**0.028691**	0.743504	0.647758	**0.169384**

Figura 31: Errores de prueba (por mes) de los tres Árboles de Decisión propuestos.

Con esto, se recalca la importancia de dividir los datos en conjunto de entrenamiento y de prueba, y de considerar el error de prueba como métrica final a evaluar en lugar de error de entrenamiento.

Referencias

[1] INEGI. (2019). COMUNICADO DE PRENSA NÚM. 179/19. 26 de septiembre de 2019, de INEGI Sitio web: https://www.inegi.org.mx/contenidos/saladeprensa/boletines/2019/OtrTemEcon/ENDUTIH_2018.pdf

[2] Santillán, O. (2018). Smartphone es el rey de los hogares: INEGI. 26 de semptiembre de 2019, de https://www.publimetro.com.mx/mx/noticias/2018/02/20/smartphone-rey-los-hogares-mexicanos-inegi.html

[3] González, L. (2018). Aprendizaje Supervisado: Proceso Generdor de Datos (Modelo teórico). 26 de septiembre de 109, de Github Sitio web: https://felipegonzalez.github.io/aprendizaje-maquina-mcd-2018/introduccion.html#aprendizaje-supervisado-1l

[4] González, L. (2018). Aprendizaje Supervisado: Predicciones. 26 de septiembre de 2019, de Github Sitio web: https://felipegonzalez.github.io/aprendizaje-maquina-mcd-2018/introduccion.html#predicciones

[5] Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and Regression Trees. CRC Press; 1984.

[6] Breiman, L. Random Forest. Machine Learning 2001, 45:5-32

[7] Sammut, C. (2019). Mean Absolute Error. 26 de septiembre de 2019, de Springer Link Sitio web: https://link.springer.com/referenceworkentry/10.1007%2F978-0-387-30164-8_525

[8] s.a. (2016). MAE and RMSE — Which Metric is Better?. 26 de septiembre de 2019, de Media Sitio web: https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d

[9] CONABIO. (2010). Regiones Económicas de México. 26 de septiembre de 2019, de CONABIO Sitio web: http://www.conabio.gob.mx/informacion/gis/layouts/recomgw.png

[10] Hyndman, R. (2016). Cross-Validation for time series. 14 febrero de 2020, de Sitio web: https://robjhyndman.com/hyndsight/tscv

[11] Cochrane, C. (2019). Time Series Nested Cross-Validation. 14 de febrero de 2020, de Towards Data Science Sitio web: https://towardsdatascience.com/time-series-nested-cross-validation-76adba623eb9

[12] González, L. (2018). ¿Qué es aprendizaje de máquina (machine learning)?. 26 de septiembre de 109, de Github Sitio web: https://felipegonzalez.github.io/aprendizaje-maquina-mcd-2018/introduccion.html

[13] González, L. (2018). Métodos Basados en árboles. 01 de marzo de 2020, de Github Sitio web: https://felipegonzalez.github.io/aprendizaje-maquina-mcd-2018/metodos-basados-en-arboles.html

[14] González, L. (2019). Bosque Aleatorios Regresión – Teoría. 01 de marzo de 2020, de Sitio web: https://ligdigonzalez.com/bosques-aleatorios-regresion-teoria-machine-learning

[15] Aler, R. (2015). DECISION TREE HYPER-PARAMETERS. TUNING DECISION TREES. 26 de septiembre de 2019, de Universidad Carlos III de Madrid Sitio web: http://ocw.uc3m.es/ingenieria-informatica/machine-learning-i/decisiontreeshyperparameters.html

[16] Handika, T. (2017). Practicing Regression Techniques on House Prices Dataset-Part 2. 26 de septiembre de 2019, de Media Sitio web: https://medium.com/@blazetamareborn/practicing-regression-techniques-on-house-prices-dataset-part2-16a78eec0df9

[17] Kaghazgarian, M. (2018). Decision Tree Regressor on Bike Sharing Dataset. 26 de Septiembre de 2016, de Kaggle Sitio web: https://www.kaggle.com/marklvl/decision-tree-regressor-on-bike-sharing-dataset/comments

[18] Koehrsen, W. (2018). Hyperparameter Tuning the Random Forest in Python. 26 de septiembre 2019, de Medium Sitio web: https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74

[19] Malik, U. (2018). Cross Validation and Grid Search for Model Selection in Python. 26 de septiembre de 2019, de Stackabuse Sitio web: https://stackabuse.com/cross-validation-and-grid-search-for-model-selection-in-python/

[20] scikit-learn developers. (2019). Decision Tree Regression. 26 de septiembre de 2019, de scikit-learn Sitio web: https://scikit-learn.org/stable/auto_examples/tree/plot_tree_regression.html#sphx-glr-auto-examples-tree-plot-tree-regression-py

[21] scikit-learn developers. (2019). sklearn.tree.DecisionTreeRegressor. 26 de septiembre de 2019, de scikit-learn Sitio web: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html

[22] Ruiz, D. (2018). predicting sales 1c. 26 de septiembre de 2019, de Github Sitio web: https://github.com/Druizm128/predicting_sales_1c

[23] Orellana, J. (2018). Arboles de decision y Random Forest. 01 de marzo de 2020, de Bookdown Sitio web: https://bookdown.org/content/2031/arboles-de-decision-parte-i.html

[24] Russell. (2018). Creating and Visualizing Decision Trees with Python. 01 de marzo de 2010, de Medium Sitio web: https://medium.com/@rnbrown/creating-and-visualizing-decision-trees-with-python-f8e8fa394176