Tecnológico de Monterrey

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

Proyecto Samsung
Entrega Final

Materia:

Laboratorio de Diseño y Optimización de Operaciones

Alumnos:

Alberto Montes Aguilar A01364419
Alejandro Galindo Merino A01021555
María Fernanda Ortega Ortega A01368137
Raúl Enrique Muñoz Garcia A01364332
Alejandro Alarcón Jaimes A01365420

Profesora:

Ana Luisa Masseto Herrera

Fecha de entrega:
19 de Noviembre de 2021

Introducción

Este proyecto fue realizado con el objetivo de lograr dar una correcta interpretación de los datos que se recopilaron sobre las ventas de los equipos de telefonía de la marca Samsung, en cada uno de las diferentes tiendas en la república mexicana, para de esta manera diseñar estrategias de merchandising, valorando los factores que intervienen en el proceso de attachment con el cliente y lograr conectar con el mercado meta, potencializar el rendimiento de los recursos y posicionar en el mercado a Samsung como la mejor alternativa de telefonía celular a lo largo de la república mexicana.

Dada la magnitud de datos con las que contamos se adoptó como recurso para el procesamiento de datos a información, la ciencia de datos, gracias a que la adaptabilidad de los modelos estadísticos permite que una vez procesada la información, se muestre de manera más gráfica y puntual, y ayuda a resolver las dudas que se tienen respecto al alcance y objetivos específicos a cubrir.

Finalmente se muestra en el reporte las estrategias de mejora y lecciones aprendidas a lo largo del proyecto para poder mostrar a los ejecutivos de los centros de distribución que alternativas se sugieren, dadas las condiciones actuales y futuras del mercado.

Etapa 1: Comprensión del negocio

Descripción de la situación actual.

Samsung es la marca que más productos tecnológicos abarca, desde smartphones hasta neveras pasando por monitores, relojes inteligentes, monitores y memorias RAM. Se trata del mayor grupo empresarial surcoreano, con numerosas filiales que abarcan negocios como la electrónica de consumo, tecnología, finanzas, aseguradoras, construcción, biotecnología y sector servicios.

Actualmente se cuenta con una base de datos que recopila las ventas de dicha compañía dentro de la República Mexicana, sin embargo se requiere transformar los datos a información para conocer cuales son los puntos de venta con más y menor ventas, verificar la rentabilidad de las sucursales y los equipos preferidos por los usuarios.

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

En una empresa tan grande como Samsung, cualquier mínimo error en el pronóstico de la demanda de sus productos puede significar consecuencias terribles, tales como:

- *Pérdida de clientes, provocando que se vayan con la competencia (pérdida de competitividad).*
- *Costos fijos y de materia prima sobre lo necesario o bajo lo necesario, propiciando problemas con la rentabilidad y liquidez de la empresa.*
- *Imposibilita la oportunidad de dominar el mercado.*

Por ello, es de suma importancia que se obtenga un pronóstico lo más acertado posible a la realidad para que permita hacer una correcta toma de decisiones, que permita establecer los grados de esfuerzo en las diferentes áreas de la empresa y que se logre el posicionamiento y éxito esperado en el mercado.

Entender y describir la problemática (en términos de ciencia de datos).

Para la toma de decisiones de una empresa tan importante como Samsung, es necesario que se apliquen metodologías y herramientas precisas para conseguir pronósticos de ventas más certeros. Mediante la ciencia de datos, se podrán utilizar los datos de ventas pasadas en la empresa para realizar las siguientes acciones:

- *Convertir datos en bruto en información funcional para la toma de decisiones de la empresa.*
- *Aplicación de métodos, procesos, algoritmos y sistemas científicos para extraer información.*
- *Obtención de conocimiento por medio de grandes volúmenes de datos*

(estructurados).

- *Aumento de la eficiencia, al tomar decisiones respaldadas por datos estadísticos.*
- *Establecer los parámetros adecuados para conocer las preferencias de los clientes y mejorar el nivel de servicio.*
- *Nuevas fuentes de ventaja competitiva frente a otros líderes tecnológicos.*

Plasmar los objetivos.

- *Conocer la empresa, entender la problemática, planificar y definir los objetivos.*
- *Definir el alcance del proyecto y establecer las variables, para comenzar con la limpieza de datos.*
- *Realizar la limpieza de datos y categorizarlos de acuerdo a las variables que le corresponden.*
- *Probar diferentes modelos estadísticos para realizar pronósticos, construir gráficas e interpretarlas.*
- *Contrastar los modelos, seleccionar el adecuado y tomarlo como referencia para la construcción del código.*
- *Diseñar un borrador del código, revisarlo y estructurarlo, para después codificarlo y correrlo con los datos que se deben analizar.*
- *Identificar los principales puntos de venta y los modelos más vendidos en las sucursales, gracias a las imágenes construidas en el análisis de datos.*
- *Evaluación de los resultados obtenidos graficar y comparar los modelos estadísticos empleados para ver cual es el adecuado para el pronóstico de venta*
- *Redacción de reporte y presentación ejecutiva, para que los directivos implementen las estrategias adecuadas para incrementar la utilidad.*

Estructurar el proyecto y hacer un plan preliminar.

El plan se muestra en un diagrama de Gantt que indica las tareas que engloba cada fase y aunque todos colaboramos en las etapas del proyecto, hay un responsable que coordina las tareas, designado de acuerdo a sus habilidades, tal como se observa en la ***Imagen 1. Diagrama de Gantt de las actividades.***

Etapa 2: Comprensión de los datos

Describir los datos crudos.

Los datos los obtuvimos a partir de un csv el cual contenía un total de 148,575 registros, dentro de las variables encontramos datos categorizados de manera errónea, por lo que decidimos usar la función **as type** para asignarle el tipo de

variable que le corresponde(chr,str,int,num ó factor).

Es necesario volver a estructurar el tipo de variable que es en especial las variables de tiempo con la finalidad de poder hacer una serie de tiempo, objetivo que nos permita utilizar el algoritmo predictor.

-GAMMA:4 categorías de producto de acuerdo a su gama (baja, alta, premium y media)

-**NÚMERO DE VENTAS** de acuerdo a la gama es:

1. baja 109201
2. alta 20980
3. media 13032
4. premium 5362

-COSTO PROMEDIO: del total de acuerdo a la gama es:

1. baja 291209986.
2. alta 259226675.
3. premium 82977222.
4. media 73536512

-SKU: Contamos con un total de 36 modelos.

-CIUDADES: Para esta variable tenemos 225

-ESTADOS: 35 estados

-ANIO: del 2018 al 2019 con un total de 148,575 registros

-**SKU:** 36 códigos de teléfono.

-PUNTO DE VENTA: en esta variable se cuenta con un total de 1793 puntos

-**FECHA**:Se tienen un total de 301 fechas únicas.

-**MES**: 15 niveles, datos numéricos y strings.

-**MARCA**: Se cuenta con un total de 5 marcas.

-**ZONA**: Existen 9 zonas únicas registradas.

-**LATITUD**: Coordenadas de latitud, se cuenta con 148,575registro: 14.87414,14.87436 ,14.89408, 14.90445,16.09876,16.21713,16.69251,16.73140 122 16.73486 y 16.74634.

-LONGITUD: Coordenadas de longitud, se cuentan con 148,575registros:

-1.009514e+06,-1.171187e+02,-1.171153e+02,-1.171141e+02,-1.170599e+02,-1.170554e+02,- 1.170347e+02,-1.170291e+02,-1.170221e+02 y -1.170204e+02

*Toda la información de las variables se encuentra en la ***Tabla 1***, ***Tabla 2*** y ***Tabla 3*** en los anexos.

Detectar problemas de calidad.

Dentro de nuestros registros contábamos con variables que no estaban correctas en sus registros, nos encontramos con casos de palabras mal escritas, o meses que estaban en el formato de str (sep, oct, nov) en lugar de un formato numérico.

Estas son las variables que necesitan correcciones debido a problemas en cómo fueron capturados:

-PUNTO DE VENTA: Se tienen que cambiar algunos errores de registro por ejemplo: "ace aldamacentro"por "ace aldama centro, en esta variable solo tenemos 1793 puntos de venta.

-ANIO: En esta variable teníamos registros en formatos incorrectos por ejemplo año estaba como 19 en lugar de 2019

-MARCA: Solo teníamos una marca registrada por lo que también se contaron con algunos registros no con un formato homogéneo por ejemplo samsung en lugar de samsung o samsung-samsung.

-ZONA: Contamos con un total de 9 zonas debido a que se tenía un error de registro en "CENTRO OCCIDENTE" el cual se tiene que poner como "centro occidente" corregido esto se obtuvo un total de 8 zonas y están agrupadas por el número de ventas:

1. centro occidente 24546
2. centro sur 66682
3. golfo de méxico 8900
4. noreste 14560
5. noroeste 12662
6. norte 10633
7. pacifico sur 4209
8. peninsula de yucatan 6383

-ESTADO: Un total de 35 estados dentro de los estados contamos con algunos estaban en formato de municipio por lo que debemos corregirlos en la categoría que iban: ecatepec por estado de mexico, saltillo por coahuila, tehuacán por puebla. Por lo tanto al cambiar estos obtenemos 32 estados.

-MES: En el dataframe se tiene meses que no están con valor numérico los cuales son: ENERO, JUL,JUN, NOV Y SEPT. Debido a esto decidimos cambiarlos a formato int para que sean en total 12 que corresponde a un año.

-**LATITUD y LONGITUD**: Realizar correcciones numéricas en las coordenadas. 26447,25.5 y 26447,-100.9

Es necesario asignar realmente cual es el tipo de variable que es por ejemplo nuestra variable $ ciudad,$zona,$punto_de_venta y $estado está en el formato

Factor y este debe ser corregido a una variable str para facilitar la lectura al momento de que sea necesario agrupar variables esto también lo asignamos. Las variables que cambiamos fueron las siguientes, punto de venta(factor *string), fecha (factor*date), mes (factor*int), Sku (factor*char) , Estado (factor * string), Ciudad (factor * string), Zona (factor * string), para conocer la justificación puede verificarse en el anexo ***Tabla 4. Justificación de cambio de variable.***

Etapa 3: Preparación de los datos

Limpieza de datos

Durante esta primera etapa de limpieza de datos encontramos que nuestros registros tenían varios errores. Detectamos 5 puntos de venta mal escritos, también encontramos valores mal registrados por ejemplo números en formato de texto, los meses estaban en su gran mayoría escrito con número y no con nombre de mes y ese fue un error que también tuvimos que corregir, entre otras cosas.

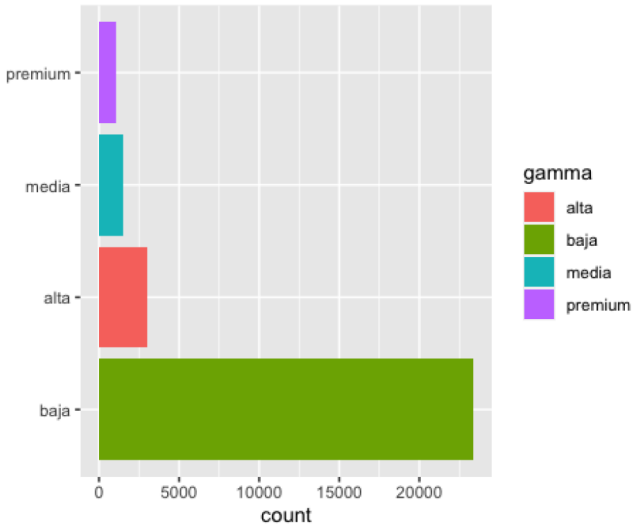
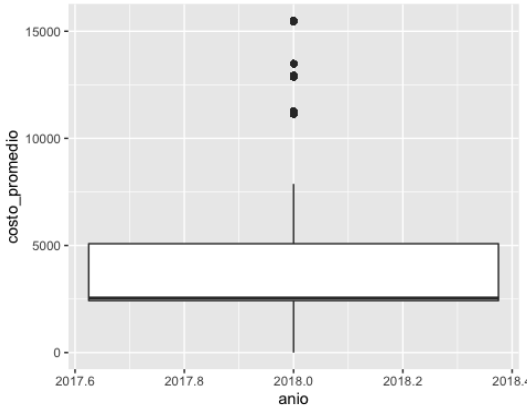
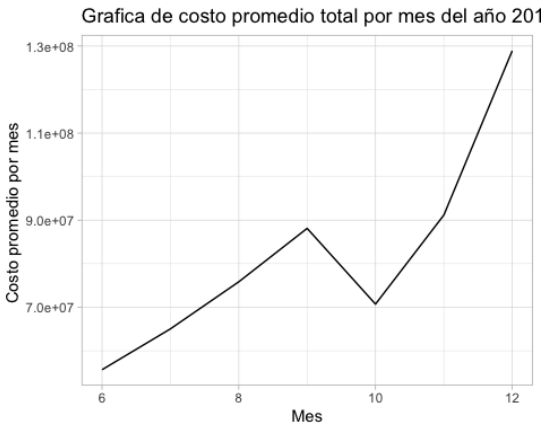
A lo largo de esta etapa tuvimos que estandarizar la información que nos fue entregada, revisando que esté completa, bien escrita, tratar de llenar los espacios vacíos con información que se tiene, y revisar que los datos estén dentro de un rango razonable.

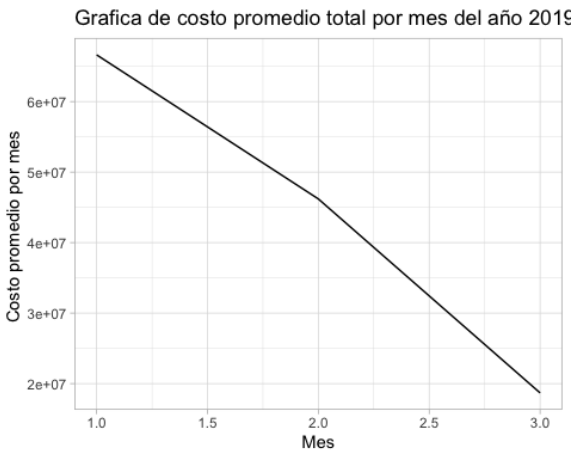
Un problema a destacar es el número de ceros que se obtuvo en la gama baja con un total de 113 valores con cero para el año 2018 y para el 2019 un total de 1759. Por lo que tenemos que revisar si estos ceros fueron debido a alguna promoción que se tuvo en algún punto de venta o tuvo que ver con otro factor.

Por último se tuvo que asignar las el tipo de variable que realmente le correspondía si era numérica, factor, fecha y character, debido a que no estaban asignadas en la categoría que les correspondía.

Análisis Exploratorio

Este análisis fue de mucha importancia debido a que se pudo apreciar de manera gráfica el total de ventas que se obtuvo de acuerdo a cada gama, pero algo a destacar es que si se tiene una diferencia notable entre el año 2018 al 2019 debido a que el año anterior tuvo un incremento notable a lo largo de los meses, pero por otro lado el 2019 obtuvo un comportamiento a la baja en el total de sus ventas por lo algunos posibles factores pueden ser que en esas fechas las personas por lo regular no suelen comprar teléfonos móviles.

	En esta gráfica de barras podemos apreciar el análisis de la demanda que se proyecta para los dispositivos de cada gama durante el año 2019, en donde podemos apreciar que aproximadamente un 80% de la demanda de los dispositivos es de gama baja, mientras que el resto se reparte en la gama alta en segundo lugar, seguidos por la gama media y la menos demandada es la premium. Los valores son los siguientes: • alta 3017 • baja 23391 • media 1495 • premium 1073
	En el diagrama de caja se observan datos atípicos en el año 2018 que están en un rango de entre 10,000 y 15,000 del costo promedio. La mayoría de los datos está en 5000, referente al costo de los dispositivos móviles.
	Se puede apreciar en la gráfica que se obtuvo un incremento considerable en el costo mínimo a partir del mes 10 en el año 2018 y que tiende a la alta durante el último periodo de ese año Posiblemente se debió a que son meses que por lo regular los consumidores adquieren un teléfono celular y como se analizó en las gráficas anteriores por lo regular son teléfonos de gama baja.

	Por otra parte en este gráfico que hace referencia al costo promedio mensual durante el 2019, y en referencia al 2018, se puede observar un comportamiento decreciente considerable, a partir del mes 3. Estos meses a detalle corresponden a enero, febrero y marzo por lo que se espera más ventas en meses como diciembre y enero, ya sea para regalos de navidad o alguna otra posible razón.

Ingeniería de características

Las variables seleccionadas fueron las siguientes al igual estas variables se tuvieron que cambiar con la función as.(tipo de variable):

Punto de venta	Character
Fecha	Date
Mes	Numeric
Año	Numeric
Sku	Character
Marca	Character

Estas variables fueron utilizadas para poder hacer mediante el joins, agrupamientos, promedios,contadores y sumatorias. Con la finalidad de poder visualizar las ventas total y generar nuestra variable respuesta la cual fue (y ventas siguiente mes). Las variables como: zona,estado,ciudad, latitud y longitud no se incluyeron debido a que esta información está implícita en la variable punto de venta.

Para poder hacer nuestro dataset se tuvo que codificar las variables de punto de venta a *pdv_id* (poniente=1, 5 de mayo zmm=2 , acayuca=3, etc). Este proceso se hizo similar para los meses creando un *mes_id* del 0 al 9, por último el sku de los 6 diferentes tipos N.SJ7PROD=1, N.SJ7PRON =2, N.SAMGA6PLA=3, N.SAMGJ4NG=4, N.SAMGJ6NG=5 y N.SAMGJ4LA =6. Esto se logró mediante la función nrow, que es un índice que nos proporciona R-studio.

Los contadores que se realizaron fue utilizando un *group by* de (sku & gamma) y con ayuda de mutate se puso una columna con el total de repeticiones con en las variables anteriores. En este apartado también se tuvo que hacer el uso del left joins para unir nuestro data set de índices con el original.

Un punto a destacar es que para poder obtener las ventas se tuvo que agrupar por punto de venta, sku y mes. Al agrupar así se puede ver el comportamiento mes con mes. Se utilizó summarise para que el número de ventas de acuerdo a la agrupación nos diera la variable ventas totales. Obteniendo un total de 72511 datos agrupados y 4 columnas.

La ingeniería de características nos funcionó debido a que al modelo se le agregaron las variables significantes que nos permitirán visualizar mejor la información al momento de hacer el forecast (como se puede observar en la ***tabla 5***). Con esto, podremos realizar un pronóstico para poder visualizar qué tan acertado es nuestro modelo de promedios móviles y después compararlo con uno de aprendizaje de máquina. Tomamos en cuenta que es necesario hacerlo una serie de tiempo por lo que se crearon 3 conjuntos y se utilizó un merge para unir mediante *pdv_id* nuestros datos.

Valores con Na se rellenan con ceros los cuales se generaron debido al uso de lags.

pdv_id <chr>	mes_id <chr>	sku_id <chr>	ventas_totales <dbl>	y_ventas_siguiente_mes <dbl>
1	0	1	1	0
1	1	1	0	0
1	2	1	0	0
1	3	1	0	0
1	4	1	0	0
1	5	1	0	0

6 rows | 1-5 of 24 columns

Tabla 5. Visualización de los índices creados en el conjunto de datos.

Etapa 4: Modelado

Promedios móviles

El método de los promedios móviles utiliza el promedio de los k valores de datos más recientes en la serie de tiempo como el pronóstico para el siguiente periodo. El término móvil indica que, mientras se dispone de una nueva observación para la serie de tiempo, reemplaza a la observación más antigua de la ecuación anterior y se calcula un promedio nuevo, en este caso se emplea la fórmula mostrada en los anexos como ***Figura 1***.
Como resultado, el promedio cambiará, o se moverá, conforme surjan nuevas observaciones. Yt : observación en el período t Ft : pronóstico para el período t

- Se promedian sólo las últimas observaciones
- El orden se determina a priori
- Un orden grande elimina los picos (suaviza)
- Un orden pequeño permite seguir muy de cerca los cambios de corto plazo

Para la serie de tiempo se emplearon los datos que se recolectaron durante los meses de Julio 2018 y hasta Marzo 2019. Se utilizaron diferentes promedios móviles los cuales fueron m=1,m=2 y m=3. Para ejecutar estos modelos se hizo uso de la función rolling la cual permite que de manera automática se agrupen los datos mediante el uso de **pdv_id** y **sku_id** para después utilizar el rolling() con la media de las ventas totales.

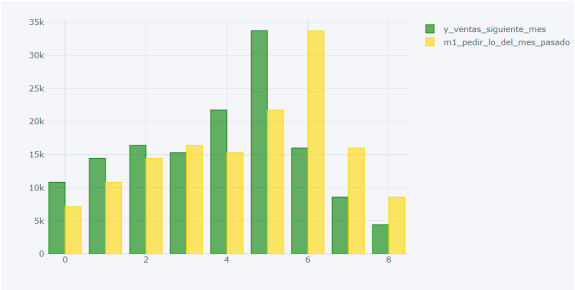
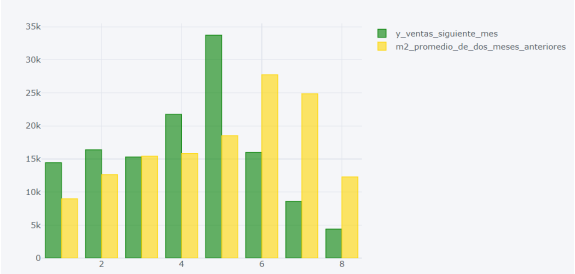
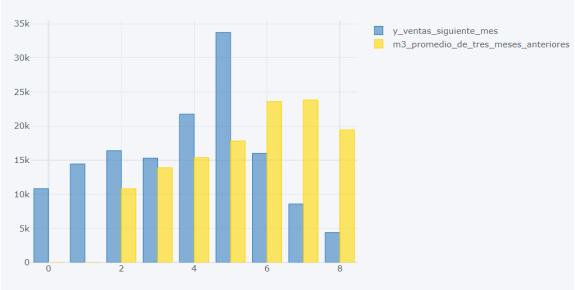
Al usar M1 pudimos ver que se podría apreciar mejor los patrones de los elementos irregulares con este enfoque de demanda reciente por lo que al intentar usar con M2 y M3 se perdía un poco la claridad del patrón en la gráfica ya que los datos están un poco más dispersos.

Para el análisis de factibilidad de los datos obtenidos por M1, M2 y M3 decidimos usar como métrica el MAE (Mean Absolute Error), el cual permite comprender que tan certeros fueron los pronósticos generados, restando la diferencia del valor pronosticado y el valor real, seleccionamos esta métrica debido a que deseamos conocer el accuracy del modelo, con los datos reales y si es funcional implementar el modelo para la planeación de asignación de recursos.

Dentro de los resultados obtenidos tenemos que utilizando el mes anterior el error se incrementa progresivamente hasta llegar a Enero en donde se obtuvo un decremento de este error de 0.507 a 0.272 en el mes de Febrero, tal como se muestra en la ***Tabla 6***, en donde podemos observar los valores puntuales de cada mes y su comportamiento para cada uno de los promedios móviles con los que se trabajó.

Este comportamiento es muy similar al momento de emplear **m=2** y **m=3**, pero la que obtiene un mejor ajuste en el modelo fue el **promedio móvil 1** debido a que es más constante en cuanto a sus valores y no se muestran cambios drásticos a comparación de los promedios móviles 2 y 3, tal como se muestra en la ***Figura 2***, puesto que se considera como un modelo factible aquel en el que el valor del MAE sea mas cercano 0 y como un valor ideal menor a 0.15.

Considerando la ***Figura 2***, podríamos decir que el M ideal es el 3, ya que es el que muestra un comportamiento no tan volátil, sin embargo, es necesario tener los valores del MAE tal como se muestra en la ***Tabla 6***, para poder visualizar el accuracy de los datos mensuales y finalmente asegurar que el M1 es el modelo que mejor se adapta al proyecto dado que su valores son los más cercanos a 0.

	Al utilizar la métrica de accuracy se obtuvo un total del 83.3 por ciento utilizando y_ventas_siguiente_mes
	En estas no se pudo poner el accuracy debido a los valores nulos que se generaron, pero como se puede observar se tiene un buen ajuste en cuanto al valor real de la predicción,
	En el m=3 se observa que si contamos con un cambio importante en cuanto a los datos ya que no se tiene como tal un ajuste bueno, por ende esto genera que en nuestra métrica de error de valores grandes.

En conclusión, para poder verificar que nuestro modelo realmente tiene una predicción acertada, va ser necesario comparar con algún algoritmo de Machine Learning el cual será el árbol de decisión.

Esta comparativa es importante ya que se tomarán en cuenta las distintas métricas de medición de errores para determinar el modelo que mejor se adapte a nuestro proyecto y los objetivos que se desean cumplir, lo cual permitirá crear las estrategias ideales para dar solución a las áreas de oportunidad identificadas en este proyecto.

Modelo de aprendizaje de máquina

El algoritmo de machine learning que se utilizó fue el de árboles de decisión, el cual consiste en la combinación de algoritmos estadísticos. Este algoritmo pretende predecir el valor de una variable respuesta. El Árbol de decisión cuenta con una estructura de nodos internos, ramas y nodos finales.

Al utilizar este modelo de aprendizaje de máquina podemos cambiar los diferentes hiper parámetros con la finalidad de mejorar su predicción, información obtenida de la librería de ***SKLEARN*** de la ***Tabla 7***.

Antes de aplicar el modelado de aprendizaje de máquina, empleamos el método de validación cruzada para poder entrenar a nuestro modelo. Con esto, pudimos realizar el conjunto de entrenamiento y prueba para 8 particiones. El set de entrenamiento pretende que crezca conforme transcurren los meses, en cambio el set de prueba solo corresponde al mes próximo, por lo que el objetivo de etapa fue poder calcular el error de ambos sets.

Los datos utilizados para nuestro set de entrenamiento fueron desde el mes de Junio a Marzo y set de prueba el mes próximo para crear la partición , con la métrica MAE para realizar el cálculo del error (***tabla 8***), el cual nos permitirá determinar el modelo óptimo.

Metrica	Conjunto	Mes
mae	entrenamiento	Julio
mae	entrenamiento	Agosto
mae	entrenamiento	Septiembre
mae	entrenamiento	Octubre
mae	entrenamiento	Noviembre
mae	entrenamiento	Diciembre
mae	entrenamiento	Enero
mae	entrenamiento	Febrero
mae	prueba	Agosto
mae	prueba	Septiembre
mae	prueba	Octubre
mae	prueba	Noviembre
mae	prueba	Diciembre
mae	prueba	Enero
mae	prueba	Febrero
mae	prueba	Marzo

Tabla 8. Tabla del conjunto de entrenamiento y de prueba.

Cálculo del modelo

Primero se construyó el modelo con la función de tree.Decision Tree Regressor y utilizamos el la profundidad de uno y tres, ya construido se realizó un entrenamiento a este con los datos de nuestro del set de entrenamiento X y Y , con el propósito de realizar la predicción de nuestra variable X la cual se comparará entre el valor real al de la predicción. La finalidad de esto sería obtener el error y poder hacer una evaluación entre un modelo tradicional a uno de ML.Los resultados del error MAE del modelo de árboles de decisión con profundidad de 1 y 3 se muestran en la ***Tabla 9***.

Metrica	Conjunto	Mes	dt_1_profundidad	dt_3_profundidad
mae	entrenamiento	Julio	0.201776	0.201497
mae	entrenamiento	Agosto	0.194739	0.184803
mae	entrenamiento	Septiembre	0.206256	0.195333
mae	entrenamiento	Octubre	0.208248	0.19794
mae	entrenamiento	Noviembre	0.23036	0.236703
mae	entrenamiento	Diciembre	0.276251	0.279987
mae	entrenamiento	Enero	0.276183	0.285721
mae	entrenamiento	Febrero	0.262277	0.274972
mae	prueba	Agosto	0.260432	0.269485
mae	prueba	Septiembre	0.236886	0.233321
mae	prueba	Octubre	0.217324	0.233321
mae	prueba	Noviembre	0.327753	0.328807
mae	prueba	Diciembre	0.505704	0.510913
mae	prueba	Enero	0.334418	0.457062
mae	prueba	Febrero	0.164931	0.233476
mae	prueba	Marzo	0.115978	0.193096

Tabla 9. Tabla del conjunto de entrenamiento y de prueba.

Etapa 5: Evaluación

	Metrica	Conjunto	Mes	Modelo_base	dt_1_profundidad	dt_3_profundidad
0	mae	entrenamiento	Julio	NaN	0.201776	0.201497
1	mae	entrenamiento	Agosto	NaN	0.194739	0.184803
2	mae	entrenamiento	Septiembre	NaN	0.206256	0.195333
3	mae	entrenamiento	Octubre	NaN	0.208248	0.197940
4	mae	entrenamiento	Noviembre	NaN	0.230360	0.236703
5	mae	entrenamiento	Diciembre	NaN	0.276251	0.279987
6	mae	entrenamiento	Enero	NaN	0.276183	0.285721
7	mae	entrenamiento	Febrero	NaN	0.262277	0.274972
8	mae	prueba	Agosto	0.2490	0.260432	0.269485
9	mae	prueba	Septiembre	0.2521	0.236886	0.233321
10	mae	prueba	Octubre	0.2327	0.217324	0.233321
11	mae	prueba	Noviembre	0.2883	0.327753	0.328807
12	mae	prueba	Diciembre	0.5047	0.505704	0.510913
13	mae	prueba	Enero	0.4515	0.334418	0.457062
14	mae	prueba	Febrero	0.3848	0.164931	0.233476
15	mae	prueba	Marzo	0.1964	0.115978	0.193096

Tabla 10. Registro del error MAE para MA, DT-1 y DT-3.

Se evaluaron los tres diferentes modelos: moving average, árboles de decisión (depth =1) y árboles de decisión (depth = 3). Los resultados que se obtuvieron de la métrica MAE, determinan en cuánto difiere el modelo de predicción con el valor real. En la ***Tabla 10***, se muestran los errores de cada uno de los modelos empleados. Para poder tomar una decisión de qué modelo se le propondrá al punto de venta, se

tuvo que graficar los errores MAE de cada modelo, con el propósito de observar que cual mostraba un mejor comportamiento a lo largo de nuestra serie de tiempo.

Esta gráfica se puede visualizar en la ***figura 4***, utilizando el promedio móvil de 1 me, se comparó con el árbol de decisión que es un modelo de machine learning, al cual se le dieron unos hiper parámetros como su profundidad los cuales fueron max_depth que es la profundidad del árbol los valores (1 y 3). Si bien el modelo de moving average con m=1 es bueno, cuando lo comparamos con los de ML nos percatamos que en realidad tenía un mayor error en cuanto a la métrica de MAE. Al seguir las tres gráficas podemos visualizar cómo de Agosto a Diciembre son muy similares los modelos, pero cuando se pretende predecir para meses posteriores a Diciembre se muestra una variación en estos.

Por esto, se determinó que el mejor modelo fue el de **árbol de decisión con profundidad 1**, ya que genera un menor error entre el valor real con la predicción. Por lo tanto, si se requiere calcular cuántos teléfonos se venden en los 1793 distintos puntos de venta será necesario implementar este modelo para maximizar las ganancias en el número de ventas y saber específicamente cuáles y cuántos de los 36 distintos teléfonos se venderán mensualmente en el mes de abril.

La asertividad y eficiencia del proyecto la evaluamos contrastando el modelo seleccionado y aplicado, con otros modelos y haciendo las pruebas necesarias para asegurar que la predicción del modelo es funcional a las necesidades del proyecto.

Sin duda alguna el uso de Business Intelligence y herramientas de machine learning permite que las empresas tengan una alternativa asequible y funcional para la toma de decisiones basada en conocimiento, una práctica que día con día es adoptada en toda la industria.

Cabe resaltar que el modelo fue construido con base en las características de las variables con las que contamos y las preguntas a las que se les busca dar respuesta, para de esta manera transformar los datos en información y generar propuestas de mejora, que permitan a los managers de las sucursales comprender el comportamiento de las variables y tomar las decisiones correctas para el cumplimiento de sus objetivos.

Referencias

Dominguez, G. (2020). Así es como gana dinero Samsung: Ni Tvs ni smartphones son la base de su negocio. Recuperado 7 de septiembre de 2021, de TECNICANET sitio web: https://www.tecnicanet.com.ar/en_US/blog/our-blog-1/post/asi-es-como-gana-dinero-samsung ni-tvs-ni-smartphones-son-la-base-de-su-negocio-28

Anexos (Etapa 1 y 2)

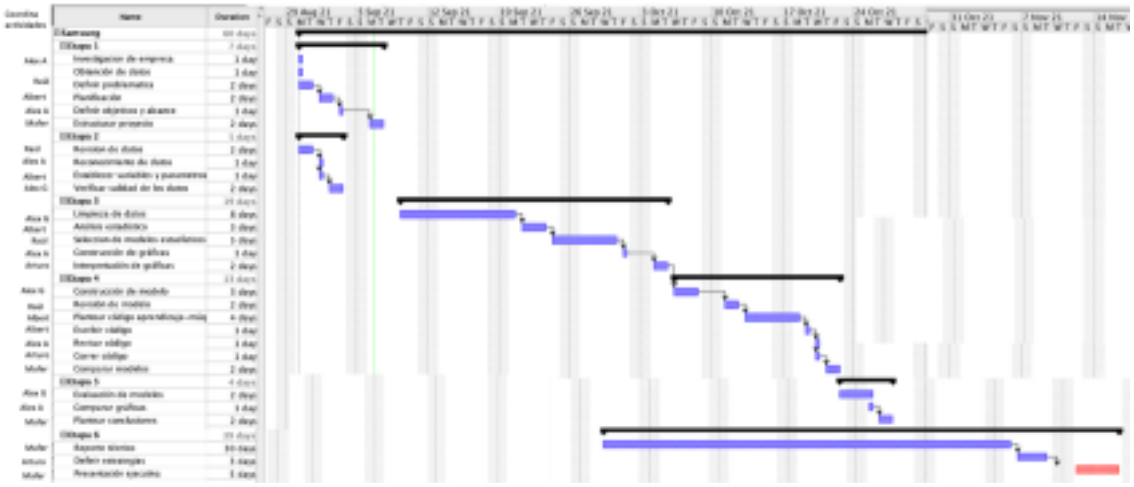


Imagen 1. Diagrama de Gantt de las actividades.

Variables de tiempo	Fecha, mes, año
Variables de tipo de producto	gamma, marca
Variables de localización	Punto de venta, zona, estado, longitud y latitud
Variables financieras	costo promedio,número de ventas
Variable identificadora	Sku

Tabla 1. Tabla de variables

Resumen de variables

```
'data.frame': 148575 obs. of 14 variables:
 $ punto_de_venta: Factor w/ 1793 levels "1 poniente","5 de mayo zmm",..: 1 1 1 1 1 1 1 1 1 1
... $ fecha : Factor w/ 301 levels "01/01/2019","01/02/2019",..: 291 291 235 6 86 126 176 176 285
293 ...
 $ mes : Factor w/ 15 levels "1","10","11",..: 7 7 8 9 9 9 9 9 9 9 ...
 $ anio : int 2018 2018 2018 2018 2018 2018 2018 2018 2018 2018 ...
 $ num_ventas : int 1 1 1 1 1 1 1 1 1 1 ...
 $ sku : Factor w/ 36 levels "N.SAMA8N","N.SAMA8PN",..: 32 33 4 7 9 7 6 6 7 7 ... $
 marca : Factor w/ 5 levels "samsung","Samsung",..: 1 1 1 1 1 1 1 1 1 1 ... $ gamma
 : Factor w/ 4 levels "alta","baja",..: 2 2 3 2 2 2 2 2 2 2 ...
 $ costo_promedio: num 4184 4196 5109 1815 2505 ...
 $ zona : Factor w/ 9 levels "centro occidente",..: 3 3 3 3 3 3 3 3 3 3 ...
 $ estado : Factor w/ 35 levels "aguascalientes",..: 22 22 22 22 22 22 22 22 22 22 ... $ ciudad :
Factor w/ 225 levels "acambaro","acapulco",..: 178 178 178 178 178 178 178 178 178 178 ...
 $ latitud : num 18.5 18.5 18.5 18.5 18.5 ...
 $ longitud : num -97.4 -97.4 -97.4 -97.4 -97.4 ...
```

Tabla 2. Resumen de variables

Variable	Tipo de variable	Diccionario de variables
punto_de_venta	str	Variable donde obtenemos el lugar donde se realizó la transacción
Fecha	date	Variable de tiempo que está desde el 2018 al 2019
Mes y año	int	Variables para agrupar de acuerdo a la fecha
Num_Ventas	int	Variable que indica el número de ventas de un tipo de teléfono en específico.
Sku	chr	Variable identificador del modelo que se vendió sirve para el inventario
Gamma	factor	Variable tipo factor está en (premium, alta, media y baja)
Costo promedio	num	Variable de costo que indica el celular que se vendio
Estado	str	Variable que indica el estado donde se ubica la tienda .
Ciudad	str	Variable que indica la ciudad en donde se vendió el dispositivo
Longitud	num(float)	Variable de posición geográfica
zona	str	variable que agrupa de acuerdo a zona geográfica.
Latitud	num(float	Variable de posición geográfica

Tabla 3. Diccionario de variables.

Variable	Tipo de variable	Justificación de cambio de variables
punto_de_venta	str	Esta variable se decidió cambiar de factor a string debido al alto número de puntos de venta los cuales son 1793
Fecha	date	La variable fecha estaba como factor, pero decidimos utilizar el as type date con la finalidad de ponerlo como serie de tiempo.
año	int	Se cambió de factor a numérica para poner agrupar de acuerdo al número de mes.
Sku	chr	Para esta variable es mejor tomarlo como carácter ya que será más fácil el agrupamiento ya que solo son 36 modelos.

Estado	str	Fue necesario cambiarlo de factor a str ya que eran demasiados estados por lo que sera mas facil tomarlo como palabras al momento de los análisis.
Ciudad	str	Al igual que en el caso de la variable estado es necesario tomarlo como string en caso de algún agrupamiento.
zona	str	Al solo contar con 8 zonas, se podía tomar como factor, pero decidimos como string para facilitar.

Tabla 4. Justificación de cambio de variable

Anexos (Etapa 4)

$$F_{t+1} = \frac{Y_t + Y_{t-1} + \ldots\ldots + Y_{t-k+1}}{k}$$

Figura 1. Fórmula de modelo estadístico de promedios móviles

	Mes	mae_pedir_anterior	mae_promedio_2_meses_anteriores	mae_promedio_3_meses_anteriores
0	Julio	0.200412	NaN	NaN
1	Agosto	0.236204	0.249016	NaN
2	Septiembre	0.254108	0.252155	0.265346
3	Octubre	0.236359	0.232740	0.240854
4	Noviembre	0.286861	0.288303	0.295247
5	Diciembre	0.529219	0.504759	0.505420
6	Enero	0.507580	0.451505	0.417085
7	Febrero	0.272523	0.384866	0.373646
8	Marzo	0.151538	0.196452	0.297655

Tabla 6. Valores de MAE para cada promedio móvil.

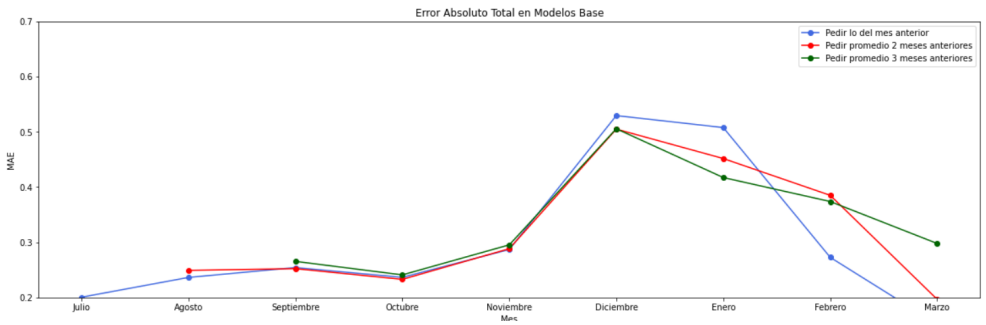


Figura 2. Comportamiento de la predicción utilizando el método estadístico de Promedios Móviles, para poder obtener el MAE.

HIPERPARAMETROS
max_depth : int, default=None La profundidad máxima del árbol. Si Ninguno, los nodos se expanden hasta todas las hojas son puras o hasta que todas las hojas contienen menos de **min_samples_split** : int o float, default=2 El número mínimo de muestras necesario para dividir un nodo interno. **min_samples_leaf** : int o float, default=1 El número mínimo de muestras requeridas para estar en un nodo hoja. **min_weight_fraction_leaf** : float, default=0.0 La fracción ponderada mínima de la suma total de pesos (de todos las muestras de entrada) requeridas para estar en un nodo hoja. **max_features** : int, float o {"auto", "sqrt", "log2"}, default=None El número de características a considerar al buscar la mejor división: **random_state** : int, instancia de RandomState o None, default=None Controla la aleatoriedad del estimador. Las características son siempre permutado aleatoriamente en cada división, incluso si "splitter" está configurado en ""mejor"". Cuando "max_features < n_features", el algoritmo seleccione "max_features" al azar en cada división antes de encontrar la mejor dividido entre ellos. **max_leaf_nodes :** int, default=None Cultiva un árbol con "max_leaf_nodes" de la mejor manera. Los mejores nodos se definen como la reducción relativa de la impureza. Si Ninguno, entonces un número ilimitado de nodos de hoja. **min_impurity_decrease** : float, default=0.0 Un nodo se dividirá si esta división induce una disminución de la impureza. mayor o igual que este valor.

Tabla 7. Hiperparametros para el árbol de decisión

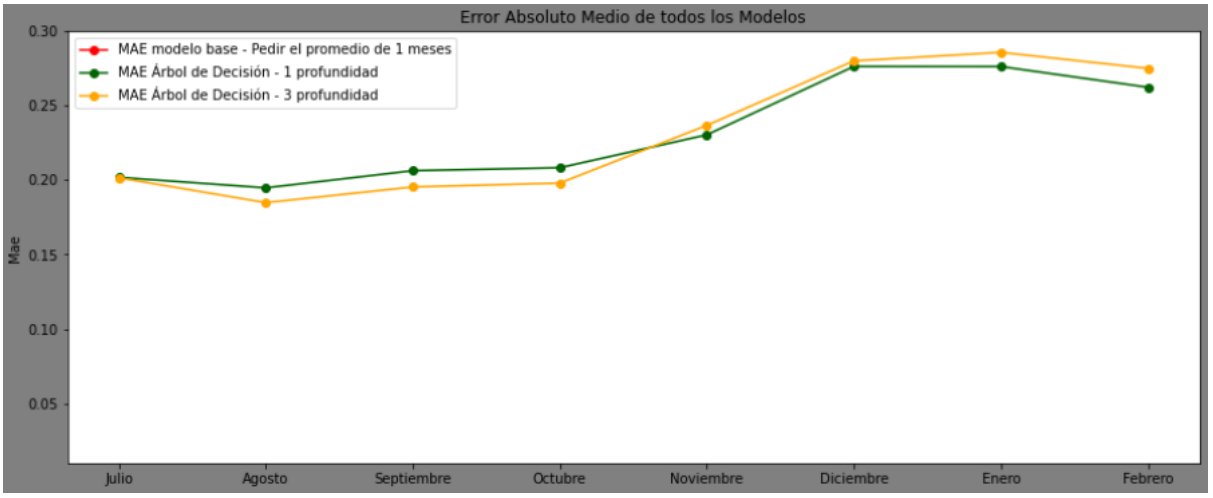


Figura 3. Comportamiento del error Absoluto Medio para el Árbol de decisión con una profundidad 1 y 3

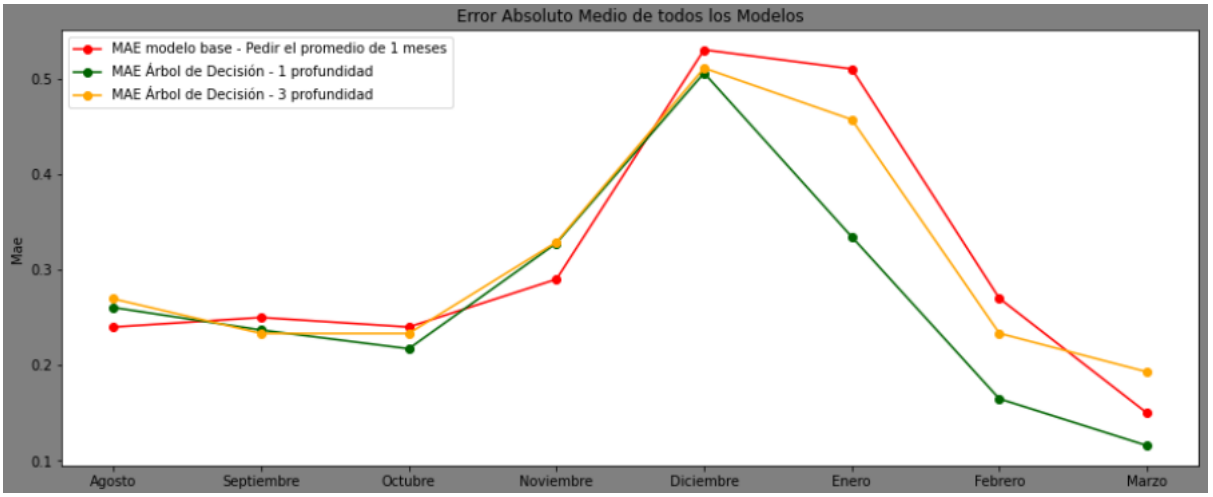


Figura 4. Comportamiento del error Absoluto Medio para el Árbol de decisión con una profundidad 1 y 3 junto con el modelo base de promedio móvil m=2