Curso:

Laboratorio de diseño y optimización de operaciones.

Reporte Final del Proyecto

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

Profesora:

Ana Luisa Masetto Herrera

Integrantes:

Joanna Priscilla Torres A01367209

Luis Ángel Ramírez A01363601

ITESM, Campus Toluca

Fecha de entrega: 20 de Noviembre del 2020

Introducción

El análisis de datos es una herramienta y rama de la estadística que nos permite comprender, procesar y analizar datos que llegan a diversas bases de datos y nos dan acceso a soluciones, planes de acción y proyecciones para el mejor aprovechamiento de los recursos de diversas empresas.

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 móvil, sector que ha presentado una enorme demanda de producto. 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.

A pesar de que la necesidad de tener un dispositivo móvil de la mayoría de las personas, no todas buscan las mismas características en estos. Es por eso, por lo que las compañías enfocadas a la venta de estos productos 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 diversos problemas, como lo son costos logísticos excesivos o insatisfacción de clientes .

Apple es una compañía líder en cuanto a la venta de celulares inteligentes, enfocado a un nivel socioeconómico de rango medio-alto, el celular iPhone en sus diferentes presentaciones ha sido un pionero en cuanto tecnología y lo ha llevado a estar dentro de los más vendido en México, incluso teniendo un precio alto. En el siguiente proyecto se tomarán en cuenta las ventas de la compañía Apple dentro de la República Mexicana, se hará un análisis estadístico con herramientas enfocadas a ello. El uso de la herramienta llamada CRISP para un análisis adecuado de los datos que nos otorgarán de cada punto de venta guiará al proyecto para obtener los resultados deseados al finalizar el proyecto. Con ayuda de R Studio y de Python, se sabrá cómo se encuentran sus ventas y con dichos datos, hacer un pronóstico para los puntos de venta y con esos resultados contestar la pregunta vital de este proyecto: ¿Cuántas unidades de cada producto de Apple, se van a vender en todos los puntos de venta, al siguiente mes de registro?

Etapa 1: Comprensión del Negocio

Descripción de la situación actual

Apple se coloca en Agosto de 2020 como la empresa más valiosa del mundo y siendo la segunda empresa en alcanzar el valor de dos billones de dólares en la bolsa de Nueva York, marcando así un registro histórico. Pudieron duplicar su valor a pesar del mal año que representó 2018 para sus ventas. Las acciones de la firma se situaron en los 467.77 dólares

Cada trimestre, Apple hace un comunicado general de toda la compañía para reportar sus indicadores financieros. Para el primer trimestre del año 2020, Apple hizo un comunicado de prensa mostrando que obtuvo un récord de ingresos netos de $22,200 millones de dólares y generó un flujo de efectivo operativo de $30,500 millones de dólares.

El sector norteamericano es de los más importantes para Apple, siendo Estados Unidos, uno de sus mayores compradores, pero el sector mexicano es también gran parte de este cliente, las personas que tienen un ingreso superior a los 100,000 pesos al año, es generalmente un cliente predilecto para la marca, ya sea en Estados Unidos, Canadá o México, gracias a su sistema operativo amigable.

Es importante realizar proyectos de análisis con datos reales para poder verificar la realidad de las acciones y situación económica de una empresa, en el caso de Apple podemos notar que la realidad de las ventas de Apple va mejorando y se ha consolidado cómo la empresa más valiosa en el mercado a pesar de que hay tiempos donde la demanda fluctúa.

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

La problemática a la que nos estamos enfrentando es poder determinar la demanda que se necesitará para cada punto de venta en todo méxico, poder analizar, comprender y explotar los datos arrojados por una inmensa base de datos que nos ayudarán a dar solución al suministro de teléfonos Apple en la república mexicana. Es importante realizar este tipo de proyectos ya que de esta manera se reducen costos de producción, envío, movilización.

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

La venta de celulares a nivel mundial sigue en crecimiento desde sus inicios y junto con la tecnología de desarrollo de móviles, la demanda y complejidad del mercado también.

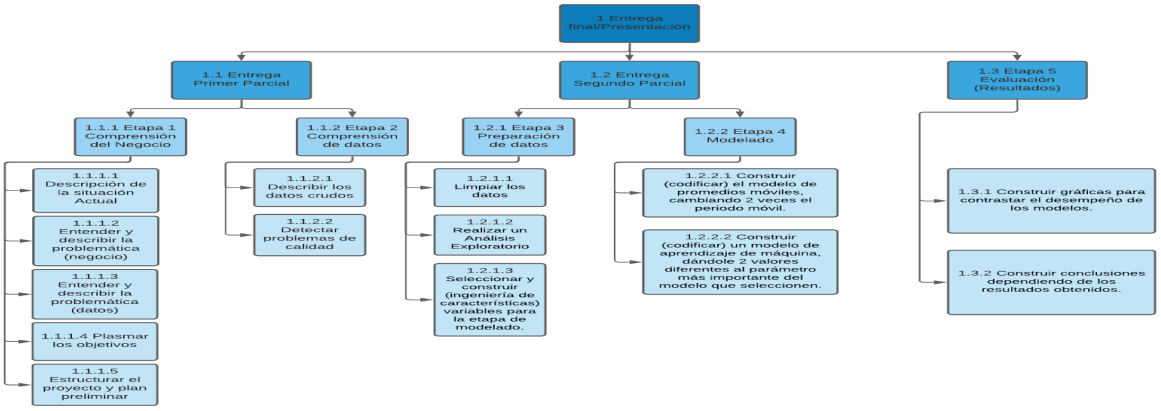
¿Cuántas unidades de cada producto, se van a vender en cada punto de venta, en el siguiente mes de registro?
Se necesita conocer una demanda aproximada de los teléfonos que serán vendidos para poder así mandar a fabricar un número aproximado al valor real de la demanda, como consecuencia, se pueden reducir costos de inventario, de producción, desperdicio y la fuga de capital.

Plasmar los objetivos.

- ➔ Etapa 1: Comprender los antecedentes y la situación actual de la compañía a través de diferentes medios, analizar información, describir el problema y estructurar el proyecto.
- ➔ Etapa 2: Recolectar datos de demanda de iPhone en México, reportar la descripción de los datos, explorarlos y hacer un reporte de la calidad de los datos.
- ➔ Etapa 3: Seleccionar, limpiar y construir datos en función al proyecto, posteriormente integrar y dar formato a los datos.
- ➔ Etapa 4: Seleccionar técnica de modelado, generar un test de diseño, construir el modelo, evaluar el modelo.
- ➔ Etapa 5: Evaluar resultados, revisar proceso, determinar los siguientes pasos.
- ➔ Etapa 6: Desplegar, monitorear y dar mantenimiento al plan, producir un reporte final y dar revisión final al proyecto.
- ➔ Objetivo del proyecto: Desarrollar un proyecto de ciencia de datos enfocado a resolver un problema de construcción de portafolios de productos (predicción de demanda),de Apple ,para los diferentes puntos de venta de la empresa,empresa de la industria de telecomunicaciones,enfocada al sector de la telefonía celular.

Estructurar el proyecto y hacer un plan preliminar.

Se hizo un cronograma y WBS para tener una excelente presentación del proyecto, a continuación se presentan los gráficos.



Etapa 2: Comprensión de los datos

Describir los datos crudos.

Los datos crudos son a groso modo, la materia prima de nuestro proyecto.

- ➔ "punto_de_venta": Lugar donde se venden los teléfonos celulares.
- ➔ "fecha": día en que se vendió el producto (iPhone).
- ➔ "mes": hace referencia al mes de “fecha”.
- ➔ "anio": hace referencia al año de “fecha”.
- ➔ "num_ventas": número de ventas que obtuvo dicho punto de venta.
- ➔ "sku": Stock Keeping Unit, número de referencia único del teléfono.
- ➔ "marca": marca del teléfono, es este caso Apple.
- ➔ "gamma": hace referencia al nivel del producto, alta, media o baja, dependiendo del precio. dependiendo del precio. Trabajamos con gamma alta.
- ➔ "costo_promedio": costo promedio del teléfono celular.
- ➔ "zona": zona del país donde se vendió el producto.
- ➔ "estado": estado donde se vendió el producto.
- ➔ "ciudad": estado donde se vendió el producto.
- ➔ "latitud": punto geográfico (latitud) del punto de venta.
- ➔ "longitud":punto geográfico (longitud) longitud del punto de venta.

Detectar problemas de calidad.

Punto de venta: Hay 5 puntos de venta escritos de manera errónea.
Fecha: Todos los registros están limpios.
Mes: Esta variable es numérica. Hay valores mal registrados (en lugar de numero, son letras).
Año: La variable año debe de seguir el formato de un valor numérico de 4 dígitos.
Número de ventas: Todos los registros están limpios.
Sku: Todos los registros están limpios
Marca: Hay 5 marcas que están escritas de forma errónea
Gamma: Todos los registros están limpios.
Costo: Todos los registros están limpios.
Zona: Hay 1 zona que está mal escrita.
Estado: Hay 3 estados más de los que en realidad existen.
Ciudad: Todos los registros están limpios.
Latitud: Hay 1 valor fuera de rango.
Longitud: Hay 1 valor fuera de rango

NOTA: Ver Anexo 1 para la visualización de PDF con WBS y archivo excel.

Etapa 3: Preparación de los datos

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

¿Cómo se hizo la limpieza de datos?

Se recibieron los datos en un archivo de Excel para poder analizar los diferentes puntos de venta que Apple tiene. Como el número de datos pasaba los límites de la herramienta de Excel, se requirió hacer uso de la herramienta de R Studio que ayudó a tener un mejor entendimiento en cuanto a los datos que se enfrentaba el equipo y hacer uso de sus herramientas para limpiar los datos obtener gráficas. Los pasos que se hicieron para la limpieza de datos fueron los siguientes:

- Se descargaron los datos de Apple y se hizo una breve introducción RStudio sobre el equipo y la marca que analizamos.
- Se llamó a la librería tidyverse para poder hacer uso de diferentes herramientas, dentro de RStudio, que nos ayudaría a tener un mejor entendimiento de nuestros datos.
- Leímos los datos sucios, del archivo de excel, en R Studio con ayuda de la herramienta de read.csv.
- Hicimos un análisis general de los datos, como ver la dimensión de los datos a los que se enfrentaba el equipo, ver el nombre de las variables, un resumen de todos éstos para tener un conocimiento general de las variables y ver el encabezado de 20 de ellas para así saber cómo se encontraban registrados los datos. El análisis nos ayudó a obtener un mejor entendimiento sobre cómo se encontraban los datos y darnos una idea sobre cómo hacer la limpieza de datos.
- El paso anterior nos ayudó a detectar problemas de calidad como fueron los puntos de venta que se encontraban mal escritos, el mes, el año, marca, zona, estado, latitud y longitud.
- En punto de venta el equipo se encontró con 5 puntos de venta más escritos como lo fue el punto de venta 5 de mayo zm, 5 de mayo zmm, arsa cty shops dl valle, ARSA periSUR, acr ATLIXCOCENTROPUE y cotzacoalcos. Para los meses, se encontraron diez meses escritos en número, para el año se encontraron dos errores con 19 y 202019, para la marca se observó que hubo 6 errores al estar mal escrita y no tener el formato “apple”, en la zona hubo un error en “NRTE”, en estado se encontraron 3 errores, dos veces con “cancun” y otro con “tepic” y en latitud y longitud hubo 1 error en cada uno.
- Para la limpieza de datos se corrigieron los problemas de calidad con la herramienta de str_replace, que fue de gran ayuda al querer corregir palabras mal escritas y después se rectificaba que los errores habían sido eliminados con la herramienta de select.
- Al final de la limpieza de datos, se volvieron a observar los datos con la herramienta de summary y vimos que los datos se encontraban correctos.

2) Realizar un Análisis Exploratorio que permita entender mejor la situación actual (gráficas,conteos, etc).

Para un mejor análisis de los datos obtenidos de los diferentes puntos de venta, se decidió realizar tres gráficas que nos ayudarían a entender mejor la situación de la compañía. Los datos que fueron más relevantes para el análisis fue el año, el número de ventas, la zona que mejor ventas tiene e histórico de ventas.

La primera pregunta fue: *¿Qué año tuvo más ventas?* y obtuvimos que el año 2018 fue el año con mayor ventas pasando por poco un número de 15,000 ventas. Para el 2019 apenas y se alcanzan las 5,000 ventas. Analizando los resultados obtenidos por la gráfica podemos decir que las ventas del año 2018 fueron mayores por la situación económica del país, el dólar aún no llegaba a los 20 pesos. *Ver Anexo 2 para la visualización de la gráfica obtenida por R Studio.

La segunda pregunta que se planteó fue: *¿Cuál es la zona donde se vende más Apple?*, cuando se analizaron los datos se tomó en cuenta cuál de todas las zonas es la que da mejores números en cuanto a ventas. Como resultado se obtuvo que el Centro Sur de la República Mexicana cuenta con el mayo número de ventas mientras que en la Península de Yucatán carecen de ventas. Analizando los resultados podemos concluir que la zona centro sur es en donde se encuentra Ciudad de México, Estado de México. El hecho de que se encuentra la ciudad dentro de la zona da indicios de que habrá más ventas ya que se encuentran los corporativos, la vida socioeconómica de las personas es mejor en comparación de los otros estados junto con el Estado de México. *Ver Anexo 3 para la visualización de la gráfica obtenida por R Studio.

Para el tercer análisis se observó que *histórico de ventas de los años registrados*. Por lo que se puede observar en la gráfica que las ventas a lo largo de los 12 meses es constante, pero también se observa que hay más ventas a principio de año, a mitad de año, en verano, hay una caída y para los meses que siguen vuelve a subir aunque en diciembre hay una caída. Se puede analizar que en enero hay descuentos en cuanto a celulares por varias fechas festivas a principios de año. A mitad de año puede haber una baja de ventas por ser verano y por que las familias hacen otro tipo de gastos como las vacaciones y finalmente hay un rebrote de ventas por inicio de clases, donde hay varios descuentos en electrónica, celulares incluidos, el segundo punto importante es el pico derivado de los descuentos por el “Buen Fin”, después de eso viene la caída de ventas por Navidad al saber que las familias hacen otro tipo de gastos para las festividades. *Ver Anexo 4 para la visualización de la gráfica obtenida por R Studio.

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

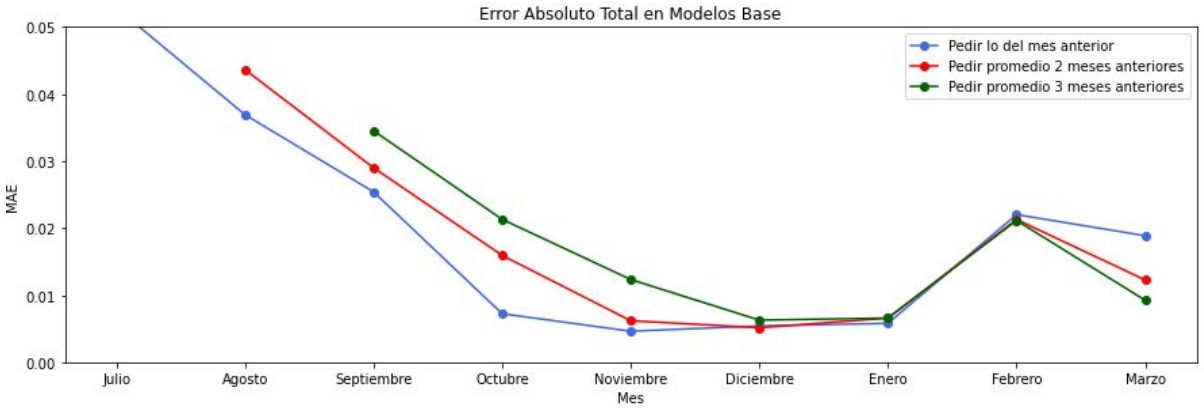
- Se llamó a la librería tidyverse para poder hacer uso de diferentes herramientas, dentro de RStudio, que nos ayudaría a tener un mejor entendimiento de nuestros datos.
- Después descargamos los datos de Apple Limpios y los leímos con read.csv.
- En el siguiente paso se hizo el análisis de los datos con las herramientas de head, dim y summary.
- Después de haber observado cómo se encontraban nuestros datos con la herramienta de summary, se procedió a hacer una recolocación de datos siendo character, date, factor o numérico.
- El siguiente paso fue crear índices por separado, esto con el objetivo de tener un manejo más sencillo de las variables cualitativas más importantes. Primero fue el índice referente a los puntos de venta, la fecha porque lo que nos interesaba es el pronóstico por mes, para el mes fue un índice distintivo por periodo de registro, después se siguió con el SKU, la marca no fue considerada porque solo hay una para el estudio, tampoco fue considerada la gamma, el costo, la zona, el estado, la ciudad, latitud, longitud y ventas totales ya que esa información venía implícita en el punto de venta o porque no era una variable cualitativa.
- El segundo paso fue agregar nuevas columnas con índices en nuestros datos.Después se analizó si había una venta del mismo producto, en el mismo punto de venta, en la misma fecha (mes), para así proceder a agrupar nuestras ventas totales, para este paso tuvimos que quitar fecha para así poder hacer un análisis por mes.
- El punto tres se completó la serie de tiempo y se construyó tres conjuntos nuevos con índices. Se crearon datasets con combinaciones.
- Para el punto cuatro se analizó la variable de respuesta que serían las ventas del siguiente mes. Nos ayudamos de la herramienta de dyplr.
- En el paso cinco, se crearon nuevas características, primero creando las características de ventas promedio por mes tienda, producto y ventas totales. Se hicieron conteos y promedios por duplas de características y como resultado se obtuvo una tabla que nos podía responder cuánto se obtuvo de venta total y el promedio de venta en cierto mes y en cierto punto de venta. Como segundo paso se incluyeron variables en los datos completos con ayuda de left_join y como tercer paso se crearon regazos de tres tiempos de ventas totales, de ventas totales por tienda y mes, de ventas promedio por tienda y mes, de ventas totales por tienda y sku y ventas promedio por tienda y sku. Por último se crearon regazos con NA, con la librería de zoo se realizó lo mismo pero para NA.
- La dimensión final de nuestros datos es de 14 columnas con 19,890 datos.

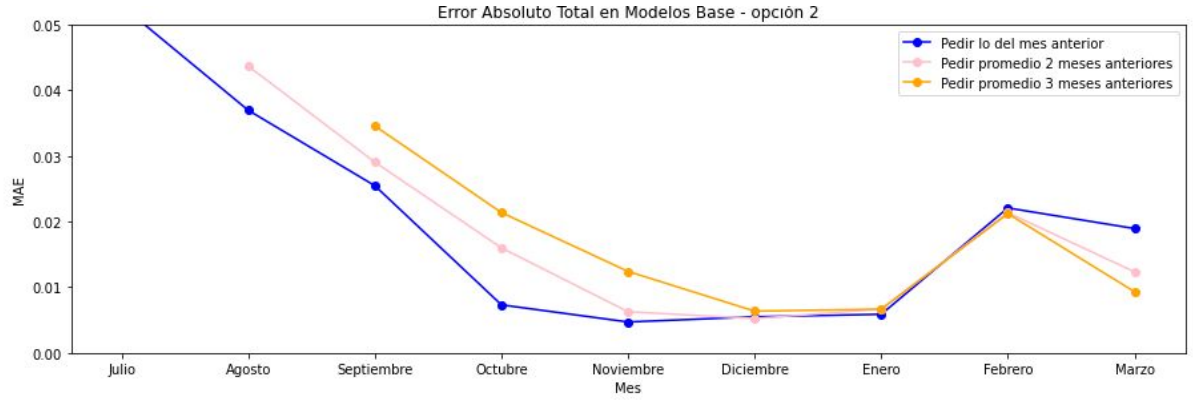
Etapa 4: Modelado

Para hacer el modelado en la etapa 4 del proyecto usamos los promedios móviles, que son indicadores de tendencias que se usan para realizar análisis de datos pasados tales como precios, ventas, puntos de venta, etc, que tiene cómo finalidad hacer un pronóstico y formar una serie de datos nuevos que nos ayuden a predecir datos futuros.

En realidad, realizar este método es bastante sencillo, tiene la facilidad de poder hacerse en diversas plataformas y entregarte resultados más o menos confiables, sobre todo cuando hablamos de predicciones. Una de las pocas desventajas o limitaciones que encontramos dentro de nuestra etapa de modelado fue que está conformado por muchos pasos sencillos pero repetitivos lo que puede llegar a confundir al usuario y de no hacerse bien podríamos llegar a un resultado que aparentemente está bien pero en realidad puede contener información errónea.

Después de hacer el cálculo de errores y dividir los conjuntos de datos por mes, creamos data frames de los errores, graficamos los datos analizando.





Graficamos los errores absolutos medios para los promedios móviles tomando en cuenta uno, dos y tres meses anteriores para hacer los pronósticos y revisar qué tanta variación podemos encontrar entre trabajar con valoraciones de promedios móviles distintos. Al cambiar dos veces el promedio móvil, se puede observar que

los resultados son muy similares, puede haber un cambio muy pequeño en cuanto a su exactitud pero es muy leve, por lo que se concluye que no hay mucha diferencia entre las dos opciones, en cuanto a saber cuántos meses nos conviene para tener un mejor pronóstico, se analizó y se concluyó que tomar un promedio de dos meses es mejor para tener una mejor exactitud ya que no tiene tanta variación en sus resultados y así no tiene tendencia a errores en el siguiente periodo.

Árbol de decisión y Random Forest

Árbol de decisión

El equipo seleccionó árboles de decisión para poder generar un resultado de desempeño de los datos obtenidos. El árbol de decisión es un modelo predictivo que divide el espacio de los predictores agrupando observaciones con valores similares para la variable de respuesta o dependiente. Son útiles para entender la estructura de un conjunto de datos y sirven para resolver problemas tanto de clasificación, como de regresión (como es el caso de este proyecto).

El tipo de problema a resolver dependerá de la variable a predecir, en este caso, nuestra variable es continua por ser un problema de regresión. La estructura de un árbol de decisión contiene los elementos: Raíz - Inicio, Condicional - Condición de cambio, Nodo - Decisiones y Hojas - Clasificación. La idea principal de este modelo es buscar puntos de corte (c) 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 y). En lugar de buscar puntos de corte o interacciones a mano, con los árboles de decisión intentamos encontrarlos de manera automática.

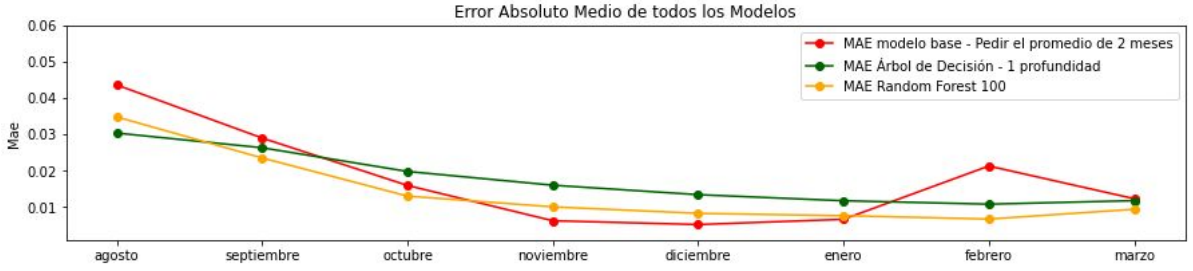
Las ventajas de un árbol de decisión es que son fáciles de construir y visualizar, selecciona variables más importantes y en su creación no siempre se hace uso de los predictores, hace predicciones tempranas, permite relaciones no lineales entre las variables explicativas y la variable dependiente y sirven tanto para variables dependientes cualitativas como cuantitativas, predictoras o independientes numéricas y categóricas.

Las limitantes dentro de los árboles de decisión son las observaciones de un nodo (mínimas) para poder hacer una partición, las observaciones en un nodo terminal y la profundidad del árbol. Esta característica dentro de un árbol de decisión corresponde a la longitud máxima que puede alcanzar una rama dentro del árbol.
El equipo hizo la lectura de datos, después se hizo la validación cruzada (preparación de los datos), donde se dividían los datos entre preparación y prueba, se generaban índices y se dividían ambos conjuntos, esto fue hecho en cada partición (en total 8). Posteriormente, se hizo el modelado para cada partición, que primero se recurrió a construir el modelo, entrenarlo, hacer predicciones, calcular máximos y mínimos de predicción, redondear y ajustar valores, juntar datos para

poder calcular el error y por último calcular los errores. El siguiente paso fueron los resultados donde el equipo registró manualmente sus datos en la herramienta de Excel y en un script nuevo de python se leyeron los registros de los errores, después de haberlos leído, se filtraban por entrenamiento y prueba, obteniendo el siguiente resultado:

	Metrica	Conjunto	Mes	Modelo_base	dt_1_profundidad
0	mae	entrenamiento	julio	NaN	0.030390
1	mae	entrenamiento	agosto	NaN	0.026347
2	mae	entrenamiento	septiembre	NaN	0.019825
3	mae	entrenamiento	octubre	NaN	0.016043
4	mae	entrenamiento	noviembre	NaN	0.013434
5	mae	entrenamiento	diciembre	NaN	0.011761
6	mae	entrenamiento	enero	NaN	0.010826
7	mae	entrenamiento	febrero	NaN	0.011788
8	mae	prueba	agosto	0.043629	0.030390
9	mae	prueba	septiembre	0.029020	0.026347
10	mae	prueba	octubre	0.015978	0.019825
11	mae	prueba	noviembre	0.006261	0.016043
12	mae	prueba	diciembre	0.005218	0.013434
13	mae	prueba	enero	0.006653	0.011761
14	mae	prueba	febrero	0.021328	0.010826
15	mae	prueba	marzo	0.012327	0.011788

Como se puede observar en la tabla anterior los errores se dividieron en entrenamiento y prueba, las particiones con profundidad de 1.



Tras analizar los resultados obtenidos en la gráfica comparativa de errores absolutos medios podemos concluir que el árbol de decisión, a pesar de ser un mejor modelo que los promedios móviles (los cuáles fueron con los que se inició el análisis) no cuentan con un mejor desempeño que los bosques aleatorios porque nuestro siguiente paso fue hacer la comparación entre dos modelos de bosques y así decidir cuál sería una mejor recomendación para la predicción del próximo mes de venta.

Random Forest

El equipo seleccionó como segundo modelo de aprendizaje de máquina los bosques aleatorios, que tiene como base los árboles de decisión que si se aplica de manera iterativa (n veces) el algoritmo de árboles de decisión con diferentes parámetros sobre los mismos datos,se obtiene un bosque aleatorio de decisión.
Ventajas de usar este modelo:

- Se pueden usar tanto en clasificación como en regresión.
- Maneja los valores perdidos y mantiene la precisión con la falta de datos.
- No sobreajusta el modelo.
- Maneja grandes conjuntos de datos con mayor dimensionalidad.
- Genera predicciones más robustas.

Desventajas o limitaciones:

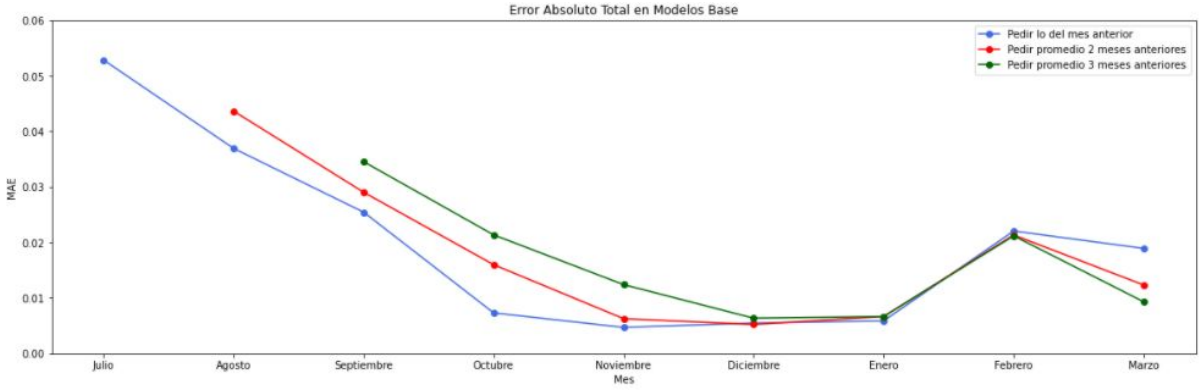
- No predice más allá del primer dataset (afortunadamente trabajamos sólo un un data set en nuestro proyecto)
- No hay mucho control sobre las acciones que realiza el modelo.
- Ya que creo muchos clasificadores encontrar el o los mejores se convierte en un proceso tedioso.

A simple vista no podemos saber cuál de los dos modelos es mejor, ya que las líneas están prácticamente una sobre otra, y aunque podemos ver que algunos meses de color verdes por debajo de los meses de color amarillo y visceversa, así que se hizo un analisis directo en la base de datos y con un muestreo y conteo de MAE's por modelo, pudimos observar que el MAE del random forest 100 tiene mejores errores absolutos más bajos. Aunque podemos elegir indistintamente entre los dos ya que ambos nos regresaran un resultado casi igual.

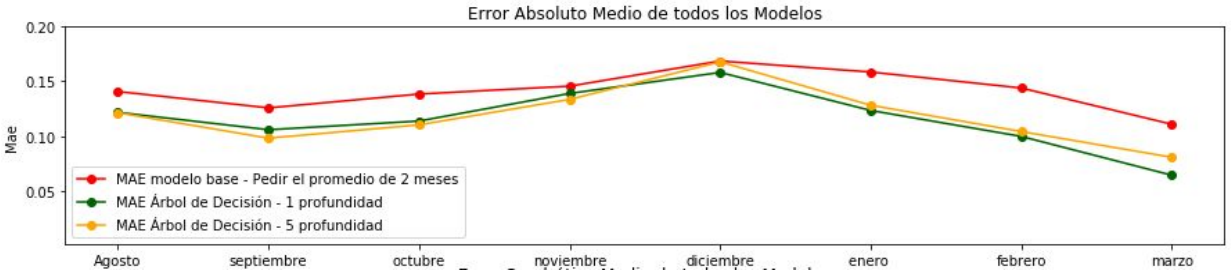
Etapa 5: Evaluación (Resultados - Conclusiones)

Empezamos este proyecto en el mes de Agosto con altas expectativas por parte de nuestros profesores y hemos podido llevar a cabo, lo que el equipo considera, un excelente trabajo, este proyecto ha sido fruto del gran esfuerzo de los integrantes y el compromiso de aprender y entregar un trabajo de calidad, tomado como fundamentos los conocimientos adquiridos en la clase. Mediante varios métodos de análisis de datos, se analizaron datos “crudos” sobre ventas de la compañía Apple en diferentes puntos de venta y fechas dentro de la república mexicana, gracias a programas tales como RStudio, Anaconda y python, se logró transformar la materia prima (los datos) a información útil para hacer un análisis de demanda de teléfonos celulares.

Gracias a las diversas maneras de poder analizar los datos que se vieron en clase, hicimos tres tipos de modelado para los data frames con datos limpios de celulares Apple. Por lo que comenzamos por hacer el modelado con las propuestas dadas por Apple (Ana), que era con promedios móviles, dentro de esta misma propuesta, se hicieron comparaciones entre el número de meses que se utilizaron para hacer la predicción de la venta de teléfonos, para propósito de nuestro proyecto, fueron usados datos del mes anterior para la primera predicción, dos para la segunda y tres para la última, dándonos como resultado la siguiente gráfica.

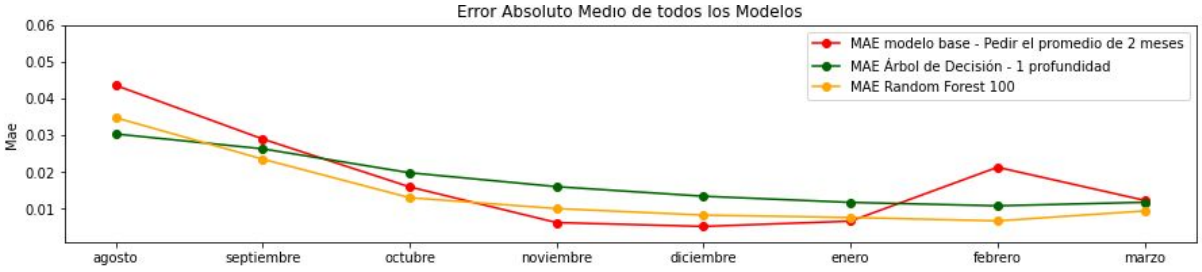


El análisis un poco más completo de esta gr´fica se encuentra en el apartado cuatro, pero la decisión tomada fue elegir el modelo de dos meses ya que ofrece un error menor al de tres meses y una mayor estabilidad en la predicción de los datos, y aunque las pruebas de errores hayan sido menores que las de dos meses, se nota una desviación mayor lo que podría representar un riesgo al momento de medir datos futuros ya que podríamos tener una alta incertidumbre, lo que con el modelo dos podemos evitar. Así que, promedios móviles de dos meses fue nuestro modelo base para poder hacer las comparaciones con los siguientes modelos.

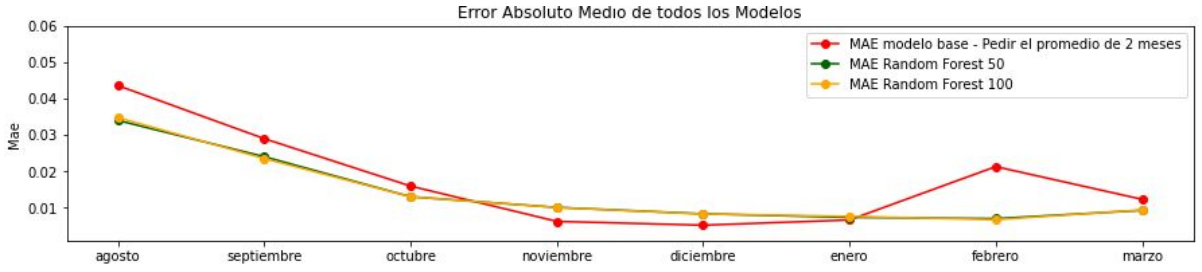


Comparamos el modelo base con uno nuevo, dos árboles de decisiones, el mismo análisis de error absoluto medio para los tres modelos y podemos observar que claramente estos dos nuevos modelos mejorarían el análisis de ventas; así que el segundo paso fue seleccionar árboles de decisión como nuevo modelo para predicción.

En la cuarta etapa se hizo la comparación entre un último modelo que fue random forest, contra árboles y el modelo base. Esto con la finalidad de conocer y ver si un random forest nos daría una mejor respuesta.



Al terminar este análisis, es un poco confuso elegir, porque en esta gráfica se observa que promedios móviles y random forest tienen un mejor desempeño que el árbol de decisión. Y yendo un poco más adelante en la interpretación, vemos que la desviación del random forest es menor por lo que decidimos usar este modelo para hacer la nueva y última comparación.



Finalmente vemos lo bien que se comportan los random forest así que después del análisis que se encuentra un poco mejor detallado en en la etapa 4, la decisión es tomar como nuevo modelo base los random forest (100 o 50 funcionan prácticamente igual). ya que con ellos tenemos una mayor certeza de que la predicción de las ventas de iPhones en México será lo más acertada posible.

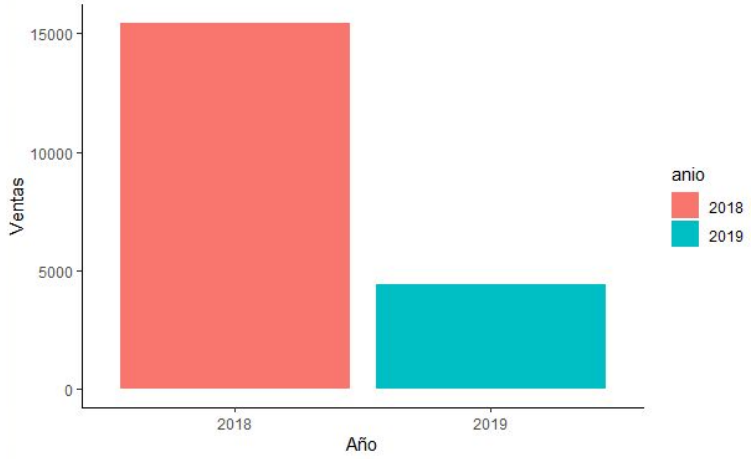
El proyecto enriqueció de manera extraordinaria nuestros conocimientos en cuanto a bases de datos, manejo de los mismo y su procesamiento, realmente es una rama de la estadísticas que es en extremo útil, que sirve para cualquier tipo de negocio, para cualquier tipo de proyecto y que bien aplicado puede generar grandes ahorros y maximizar ganancias y beneficios.

Anexos:

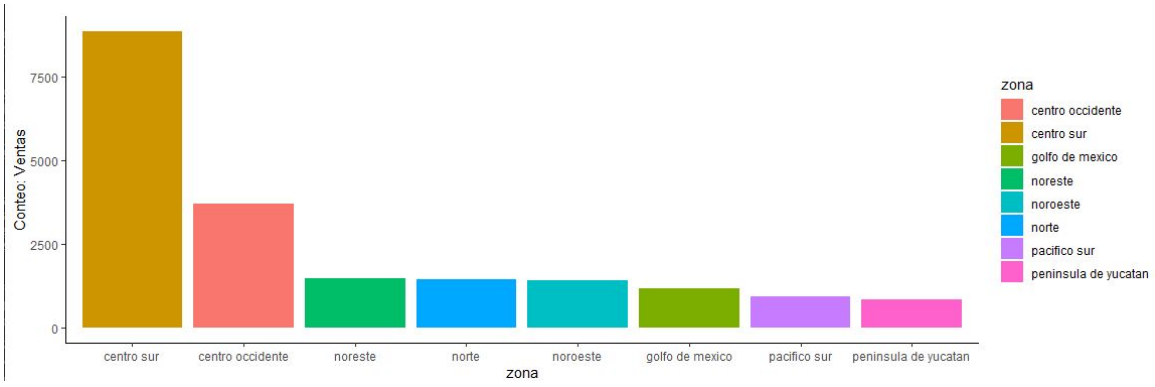
Anexo 1

Adjunto el PDF con el WBS y archivo excel con el cronograma (Gantt):
WBS:
https://docs.google.com/spreadsheets/d/1WBNJEfsyOOQq4YKEGbsyxawuQfVLkzjExfd4Twc84ws/edit?usp=sharing
GANTT:
https://app.lucidchart.com/invitations/accept/5e3d93a7-1c63-4a9f-b70a-675841b4f947

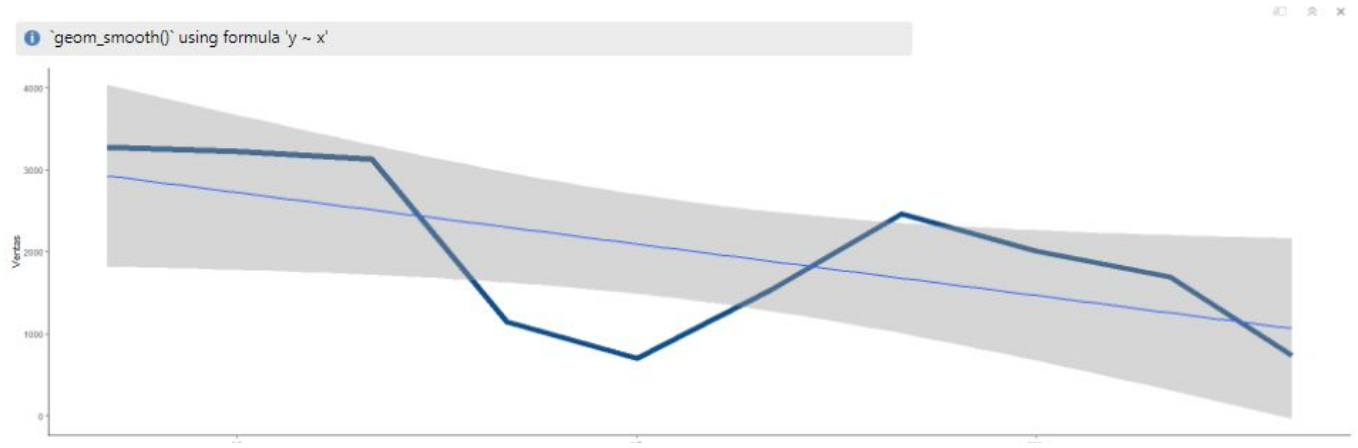
Anexo 2



Anexo 3



Anexo 4



Referencias:

Forbes Staff. (2018). Demanda de iPhone, a la baja: otro proveedor recorta previsiones. 05/09/2020, de Forbes Sitio web: https://www.forbes.com.mx/demanda-de-iphone-a-la-baja-otro-proveedor-recorta-previsiones/

EFE. (2020). Apple alcanza los dos billones de dólares de valor en bolsa. 05/09/2020, de INFORMADOR.MX Sitio web: https://www.informador.mx/economia/Apple-alcanza-los-dos-billones-de-dolares-de-valor-en-bolsa-20200819-0051.html

CNN. (2015). ¿Quién compra más productos Apple?. 05/09/2020, de CNN Sitio web: https://cnnespanol.cnn.com/2015/10/31/quien-compra-mas-productos-apple/

BELLOSTA,C. (2018). R PARA PROFESIONALES DE LOS DATOS: UNA INTRODUCCIÓN. https://www.datanalytics.com/libro_r/arboles-de-decision.html

VILLALBA, F. (2018). CAPÍTULO 8 BOSQUES ALEATORIOS DE DECISIÓN | APRENDIZAJE SUPERVISADO EN R. https://fervilber.github.io/aprendizaje-supervisado-en-r/bosques.html

KANJEE, R. (2017). RANDOM FOREST -FUN AND EASY MACHINE LEARNING [VIDEO]. RETRIEVED FROM https://www.youtube.com/watch?v=d_2lkhmjcfy