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

Proyecto Samsung

Primera entrega

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

Arturo Flores Maldonado A01366160

Profesora:

Ana Luisa Masseto Herrera

Fecha de entrega:

6 de Septiembre de 2021

Etapa 1: Comprensión del Negocio

1) 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.

2) 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.

3) 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.*

4) 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.*

5) 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

1) 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 algoritmo predictor.

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

-**NUMERO 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

2) 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 T***abla 4. Justificación de cambio de variable.***

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:

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.

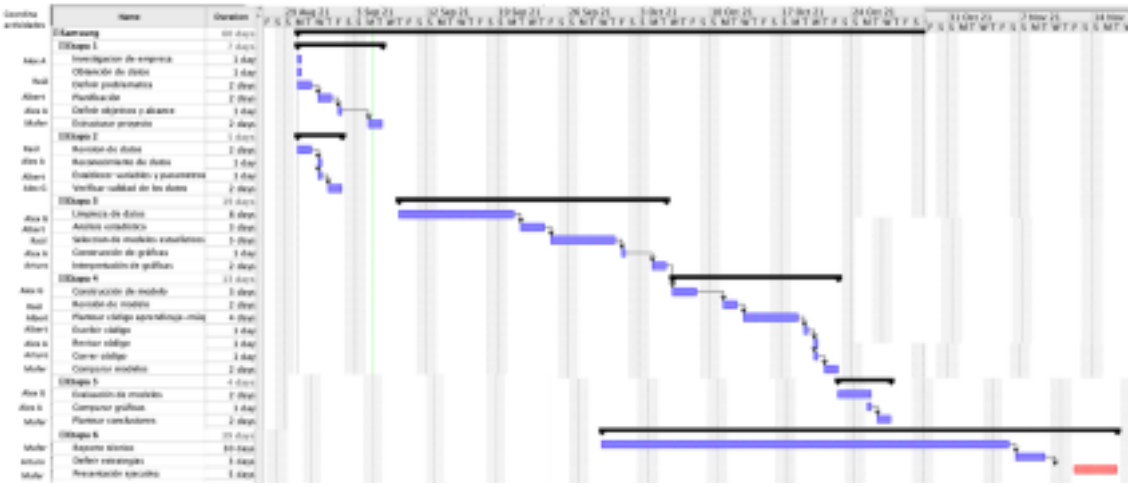

Imagen 1. Diagrama de Gantt de las actividades.

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