Universidad Nacional Autónoma de México

Diplomado en Ciencia de Datos

Por definir

Poyecto 1

Jessica Fernanda Rodríguez Mondragón

Septiembre 2020

Índice general

Resumen II

Introducción I

Aplicación al negocio I
0.1. Problemática . . . I
0.2. Solución . . . I

Calidad de datos I
0.3. Diccionario de datos . . . I
0.4. Ingenería de Variables . . . I
0.5. Visualización de datos . . . I
0.6. Outliers . . . I
0.6.1. Método Percentiles . . . I
0.7. Análisis de valores ausentes . . . II
0.8. Análisis descriptivo de Datos . . . III

Modelación Supervisada VI
0.9. Predicción de ventas . . . VI
0.9.1. K best . . . VI
0.9.2. Estandarización . . . VI
0.9.3. Regresión de Cresta . . . VI
0.9.4. KNN . . . VIII
0.10. Modelo de fuga de clientes . . . IX
0.10.1. Reducción de dimensiones . . . X
0.10.2. Transformación Entrópica . . . XI
0.10.3. Regresión Logística . . . XII
0.11. Sistema de recomendación . . . XIV

Resumen

Introducción

Aplicación al negocio

0.1. Problemática

Actualmente la vida que se conocía hasta hace unos meses ha dado un cambio radical en todo el mundo, debido a la pandemía generada por el COVID-19, han surgido nuevas normas y medidas sanitarias con las que se ha aprendido a vivir día a día. Debido al confinamiento, la economía mundial se ha visto afectada de tres maneras principales: afectando directamente a la producción, alterando la cadena de suministro y en el mercado, y por su impacto financiero en las empresas y los mercados financieros.

Si bien, la mayoría de las empresas han tenido que adapatarse a la tecnología para poder subsitir, trasladando su idea de negocio a internet, ya que las empresas que antes eran las líderes en el mercado financiero, han pasado a segundo plano, como lo son los hoteles, empresas de entretenimiento (conciertos y festivales), cines y más, mientras que las empresas que brindaban servicio por internet, como lo son food delivery, y tiendas online, han crecido exponencialmente, debido al confinamiento y a la reacción de la sociedad, ampliando su red de distribución.

Lo anterior, se debe a que acudir a tiendas departamentales, o supermercados casi desiertos, con líneas en el suelo dirigiendo el tráfico peatonal, y con más de dos metros de distancia entre cada comprador se ha convertido en algo con lo que se vive día a día, sin embargo no se vive la misma experiencia que hace unos meses, sin tomar en cuenta que aún tomando esas medidas, existe un riesgo de contraer COVID-19.

0.2. Solución

Actualmente casi no quedan personas que teniendo un negocio, no hayan pensado alguna vez en la posibilidad de crear tiendas online. Los cambios del mercado, la crisis, la evolución del marketing y el ahorro en los costes hacen que tener una tienda online sea una de laa mejores opciones ante la situación actual.

De esta forma, surge la idea de crear una tienda online, que permita dar una recomendación al cliente sobre los productos por categoría, este modelo estará basado en las relaciones que existen entre los productos, productos complementarios y sustitubilidad de un producto por otro. Además de presentar modelos de predicción de ventas y de fuga de clientes para un mejor entendimiento los ingresos percebidos y el comportamiento de los clientes, de la boutique.

Por otra parte, dependiendo del alcance del proyecto, se buscará realizar un modelo que permita al cliente subir una foto a la plataforma y probarse la prenda y/o accesorio que estén disponibles, ya que hoy en día, dada la situación, las tiendas departamentales no cuentan con este servicio de probadores.

Calidad de datos

0.3. Diccionario de datos

En esta sección, se presenta un diccionario de los datos, con el fin de entrar en contexto y presentar resultados. En la tabla siguiente se muestran las variables originales del conjunto de datos, de una boutique de ropa y accesorios de moda.

Variable	Tipo de dato	Descripción
ID _ ticket	String	Id de la orden de compra
Id _ cliente	String	Id único del cliente
Fecha	Date	Fecha y hora en la que la orden de compra se registró
VentaUnidades	Int	Número de productos incluidos en la orden de compra
ID_ producto	String	Id único del producto
VentaNeta	Float	Monto total de la compra
Store	Int	ID de la tienda física donde se hizo la compra
Departamento	String	Departamento del producto vendido
Subdepartamento	String	Subdepartamento del producto vendido
Categoría	String	Categoría del producto vendido
SubCategoría	String	Subcategoría del producto vendido
product_desc	String	Descripción del producto
NSE	String	Nivel socioeconómico del cliente
tipo_familia	String	Clasificación basada en las compras que ha realizado el cliente
share_of_waller	String	Clasificación del costo del cliente por producto

A continuación se presenta la clasificación de las variables de acuerdo a la siguiente taxonomía:

- Continuas: VentaNeta y VentaUnidades.
- Categóricas: 'tipo_familia', 'Store','NSE', 'Departamento', 'Subdepartamento', 'Categoría', 'share_of_wallet' 'SubCategoría'y 'product_desc'.
- ID: ID_cliente, ID_producto y ID_ticket
- Fecha: Fecha

0.4. Ingenería de Variables

De acuerdo al tipo de variables con las que se cuentan, se tienen ciertas variables que pueden derivar en otras variables. Al proceso de crear nuevas variables a partir de las ya existentes, se le conoce como Ingenería de Variables.
A continuación, se presenta un listado con las variables que dieron origen a nuevas:

- Fecha
    1. Año_compra : Año de la compra.
    2. Semana_compra : Número de la semana del año de la compra.
    3. Dia_semana: Día de la semana de la compra.
    4. Mes_compra: Año y mes de la compra.
- VentaNeta y VentaUnidades
    1. promoción: 1 si el artículo tenía descuento 0 en otro caso.
    2. precio_prod: precio del producto adquirido.
- VentaNeta
    1. devuelto : 1 si el artículo fue duevuelto, 0 en otro caso.
- product_desc:
    1. len_product_desc : Longitud de la descripción del producto.
    2. n_letters_desc: Número de letras de la descripción del producto.
- ID_cliente y ID_ticket:
    1. cont_cliente: 1 si el cliente es recurrente, 0 otro caso.

0.5. Visualización de datos

La visualización de los datos, permite conocer y comprender los fenómenos y el comprotamiento de los datos, sin embargo, para conocer la distribución de los estos, se requerire de pruebas y criterios estadísticos, para poder asignar un tratamiento correcto a los datos atípicos y a valores ausentes. Con el fin de conservar las distribuciones originales de los datos, únicamente se consideran los datos sin regsitros nulos, ya que al darle un tratamiento a los variables ausentes y posteriormente se muestra una visualización gráfica, la forma de la distribución se verá alterada.

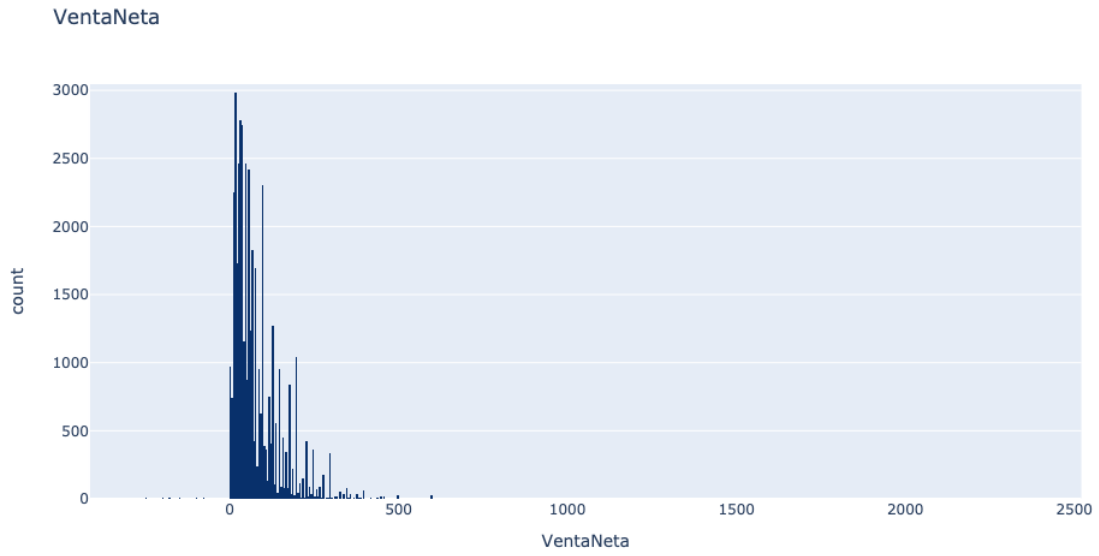


Figura 1: Histograma de Venta Neta

Como se observa, la mayor parte de las ventas son realizadas por un monto menor a 200 pesos mexicanos.

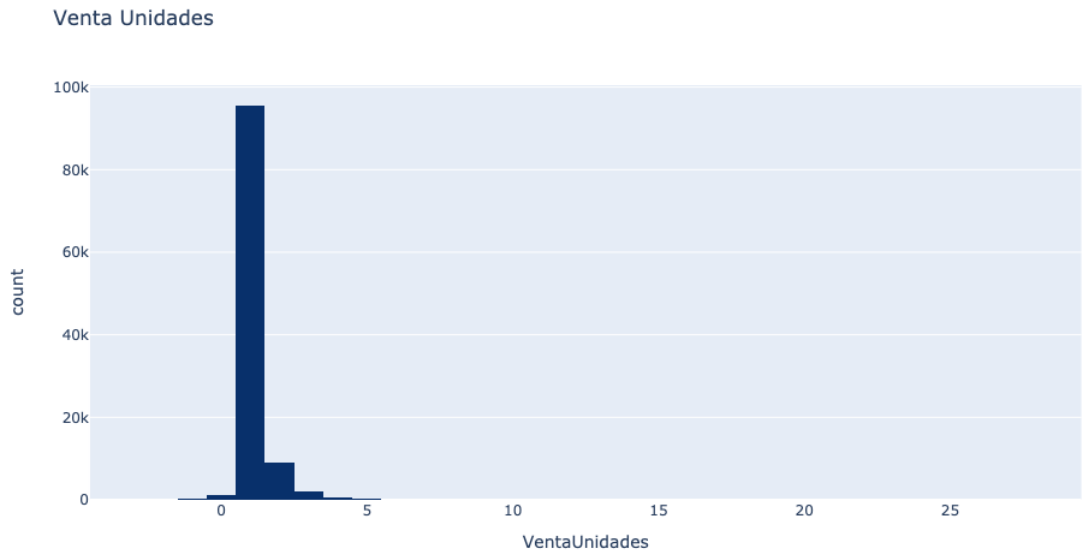


Figura 2: Histograma de Unidades por venta

Normalmente, la mayor parte de los clientes realizan compras de solo uno o hasta dos productos, hay una cantidad muy baja de clientes que realizan compras mayores a 4 productos.

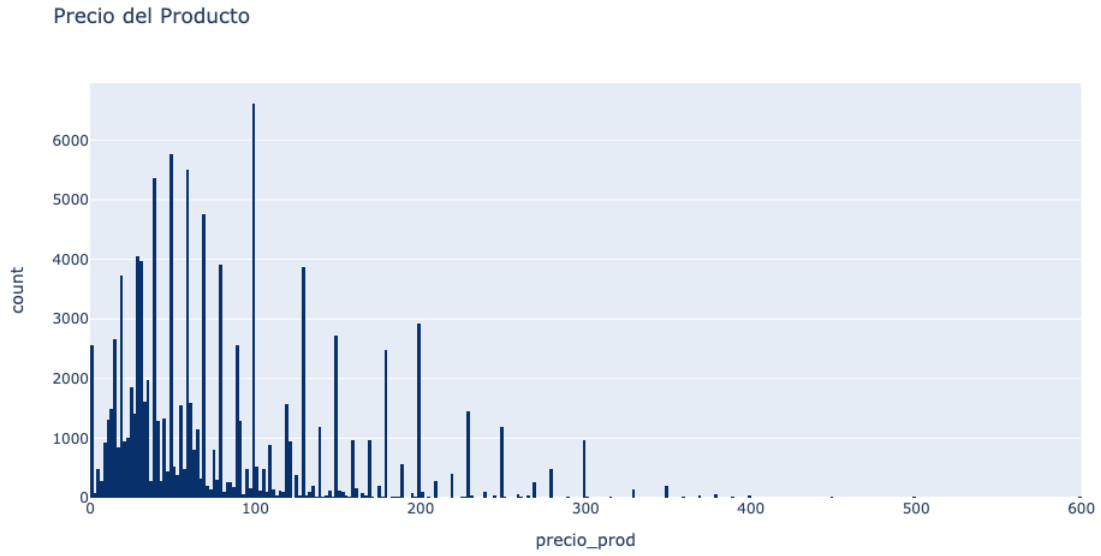


Figura 3: Histograma del precio de los productos

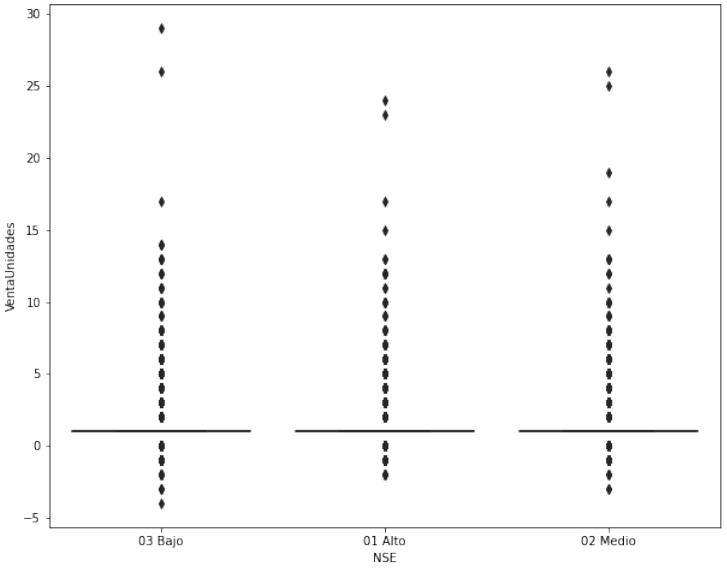
No hay productos con precios muy elevados, la mayor parte se encuentra debajo de los 200 mxn.

0.6. Outliers

Un outlier, o en español, dato atípico, es un registro con un valor completamente distinto al de todo el conjunto, es decir, se comporta de distinta manera a los demás. Este tipo de registros u observaciones se encuentran en los extremos de la distribución de los datos, en el caso de que estos sean continuos, por lo que una de las técnicas para identificarlos es el método de percentiles.

0.6.1. Método Percentiles

Un valor atípico, se caracteriza por estar en los extremos de las distribiciones de las variables, por lo que el método de Percentiles excluye a todos aquellos que quedan menores al 1 % y al 99 % de los datos. A continuación, en la siguiente gráfica se observan datos atípicos en la variable de VentaUnidades por los diferentes sectores económicos:



Debido a lo anterior, este comportamiento puede afectar al desempeño del modelo, por lo que optamos por quitar aquellos datos atípicos, perdiendo el 2 % de información de los registros.

0.7. Análisis de valores ausentes

Se cuenta con la tabla inicial, en la que se encuentran todas las variables originales, por lo que se crea una visualización sobre los valores ausentes en los datos.

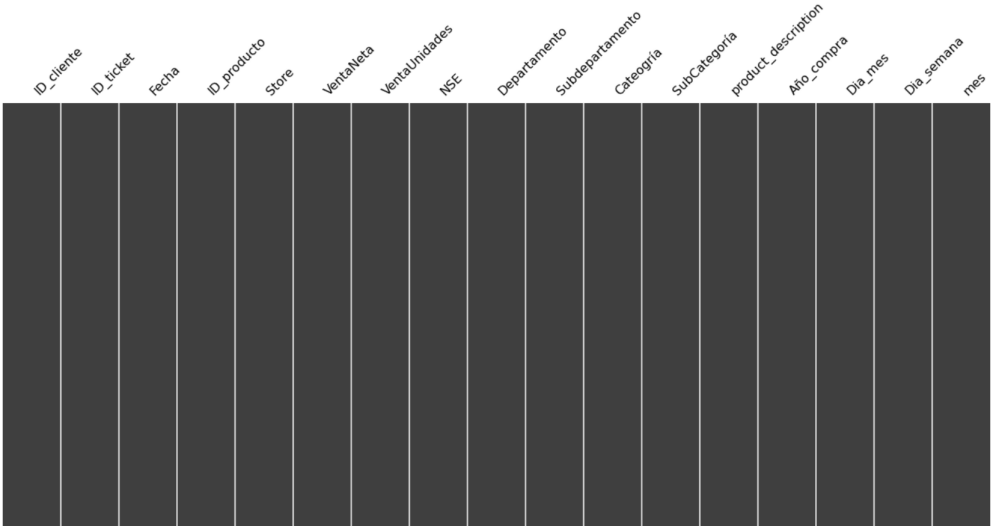
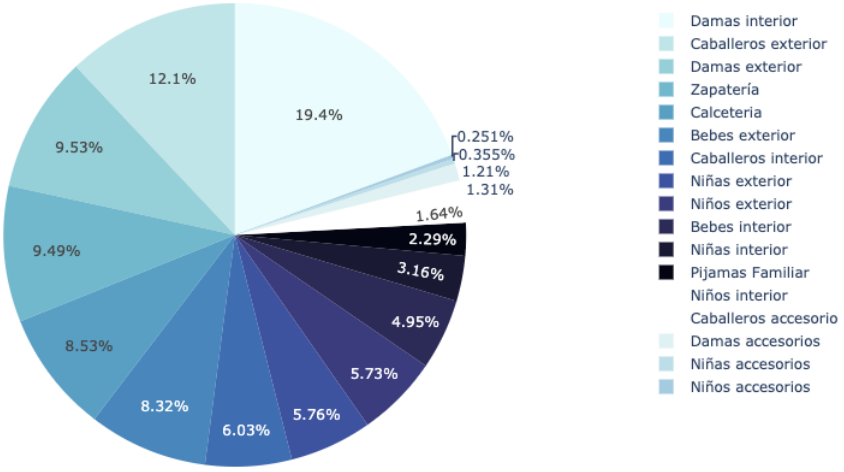


Figura 4:

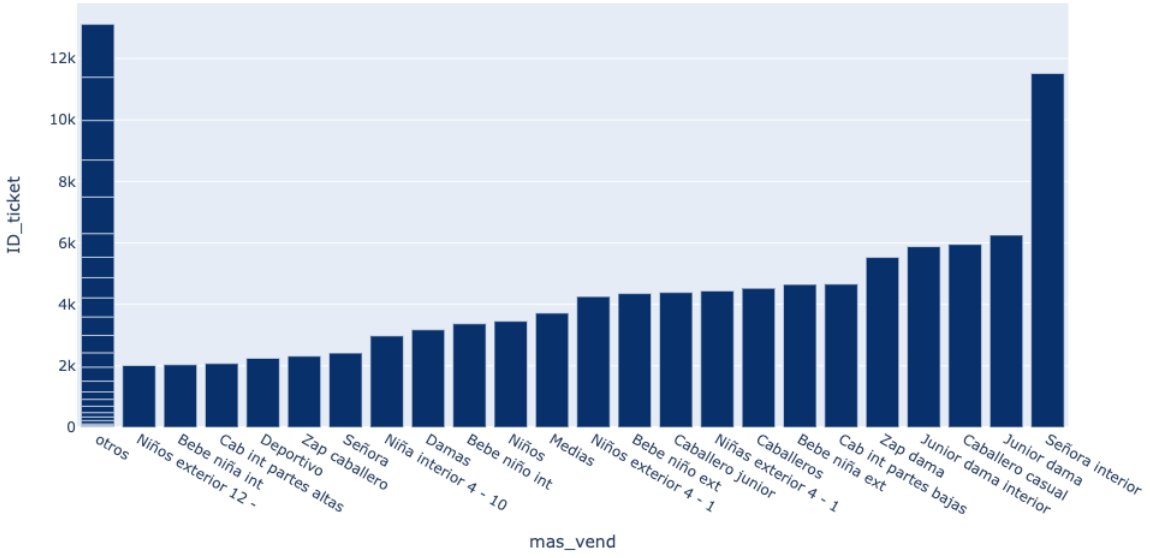
Como se observa, no se tienen valores ausentes, por lo que no se requiere algún tratamiento especial o adicional.

0.8. Análisis descriptivo de Datos

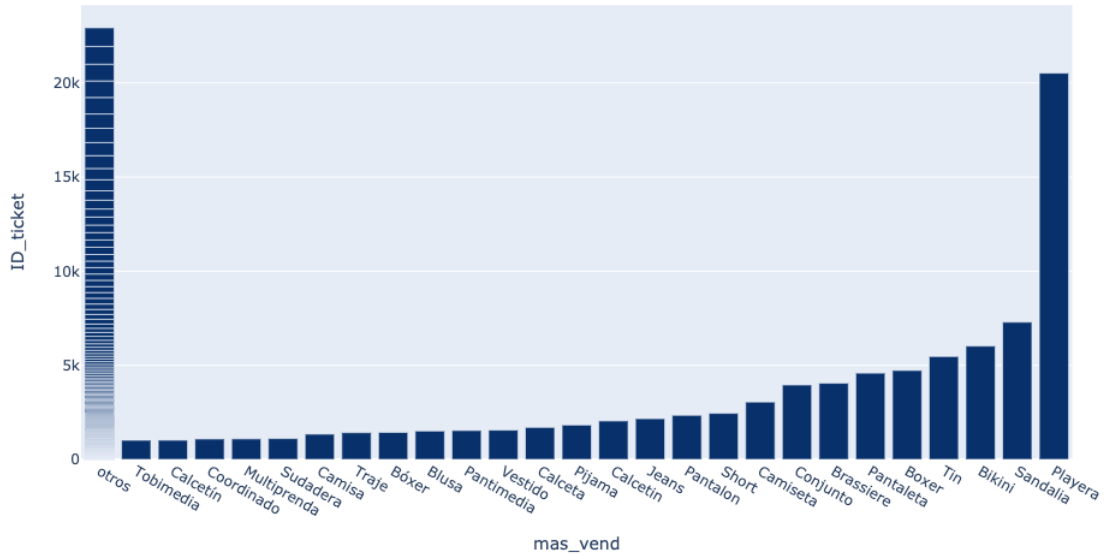
En este apartado, se busca realizar un análisis sobre los datos, para entender su comportamiento, por lo que se presenta una serie gráficos con el fin de una mejor interpretación.
El departamento de interés a analizar de la boutique es Ropa, el cual está conformado por 17 subdepartamentos:



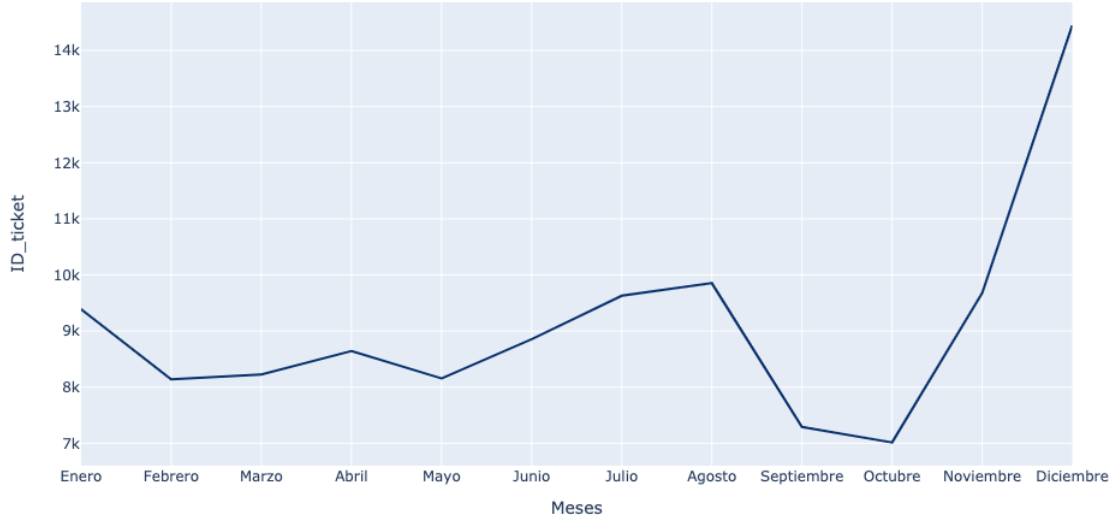
Los subdepartamentos con mayores ventas son Damas interior, Caballeros exterior, Damas exterior, y Zapatería, siendo estos el 50 % de las ventas totales de la tienda. Por otra parte, existen 48 categorías de productos, las más vendidas son las siguientes:



Sin embargo, la categoría de Señora interior es muy amplia, por lo que a continuación, a nivel general se muestran los productos más vendidos:

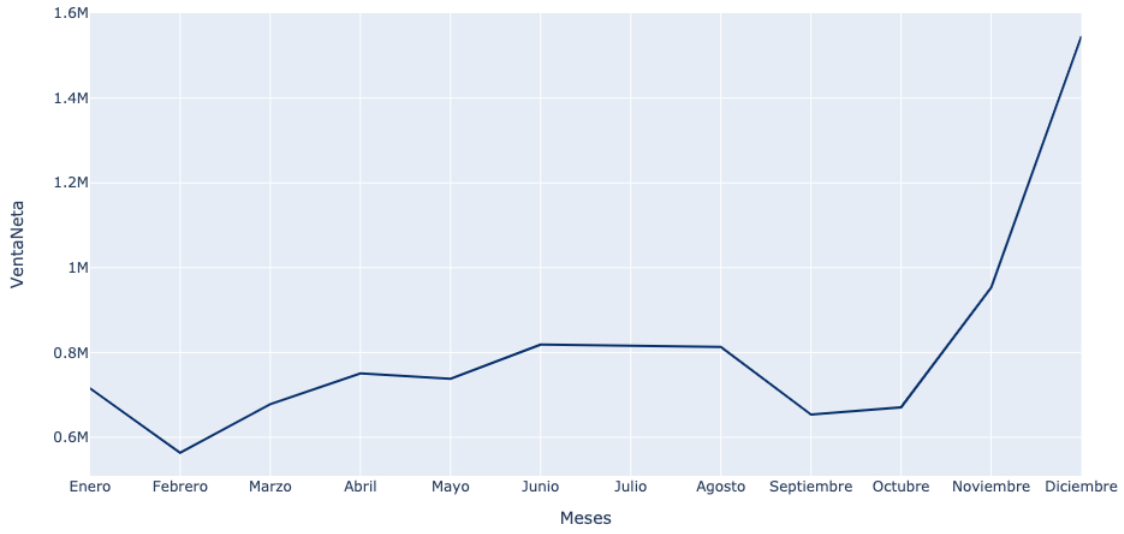


Se observa que el producto más vendido es playera, este producto puede estar dentro de las diferentes categorías, ya que existen playeras para dama, juniors, caballeros y bebés. Por otra parte, uno de los principales objetivos de este proyecto, es estimar las ventas, por lo que a continuación se muestra un seguimiento de las ventas mensuales de 2019 de todas las categorías:

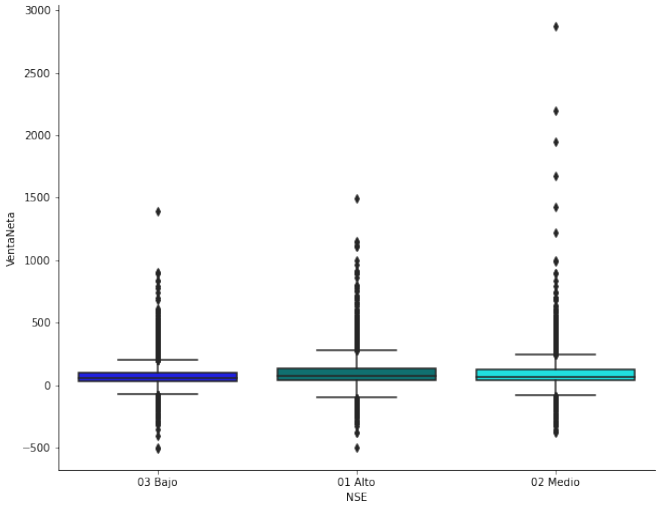


Se observa que las ventas en el mes de diciembre son mayores a diferencia del resto del año, los meses más bajos respecto a ventas son Septiembre y Octubre.

Ahora, este comportamiento se sigue manteniendo de acuerdo a la cantidad de productos vendidos mensualmente, lo cual es de esperarse, ya que entre más productos vendidos, existe un mayor flujo en la ganancia para la Boutique:



Otra característica de interés con la que se cuenta es el nivel socioeconómico, dependiendo de la compra que el cliente realizó, el cual se divide en 3, 01 Alto, 02 Medio y 03 Bajo, a continuación se muestra el monto de las compras por nivel socioeconómico:



De las compras que pertenecen al nivel socioeconómico Alto, la cantidad neta por cada compra es mayor a la de los demás sectores.

Modelación Supervisada

0.9. Predicción de ventas

Haciendo uso de modelos de regresión, se estimarán las ventas semanales por subcategoría de la tienda de ropa, ajustando el mejor modelo, para esto se cuenta con una dimensionalidad de 51 columnas, las cuales son el numero de semana del año 2019, con 593 columnas, que refieren a las funciones de agregación (suma, media, desviación estándar, máximo, mínimo y conteo para las diferentes categorías, así como para las variables previamente creadas), debido a esto, haremos uso de la técnica de selección de variables K Best.

0.9.1. K best

Esta técnica devuelve las mejores variables basadas en su puntuación. Con fines de ralizar la estimación de las ventas, y reducir la cantidad de variables que se tienen, bajo esta técnica se usarán las 45 mejores variables para entrar al modelo, las 10 primeras son las siguientes:

- 'sum_monto_total_Calceteria',
- 'sum_monto_total_Damas exterior',
- 'sum_monto_total_Damas interior',
- 'sum_monto_total_Niñas exterior',
- 'sum_monto_total_Niños exterior',
- 'sum_monto_total_Zapatería',
- 'mean_monto_total_Niñas exterior',
- 'mean_monto_total_Niños exterior',
- 'mean_monto_total_Zapatería',
- 'P1_sum_monto_total_Niñas interior',

0.9.2. Estandarización

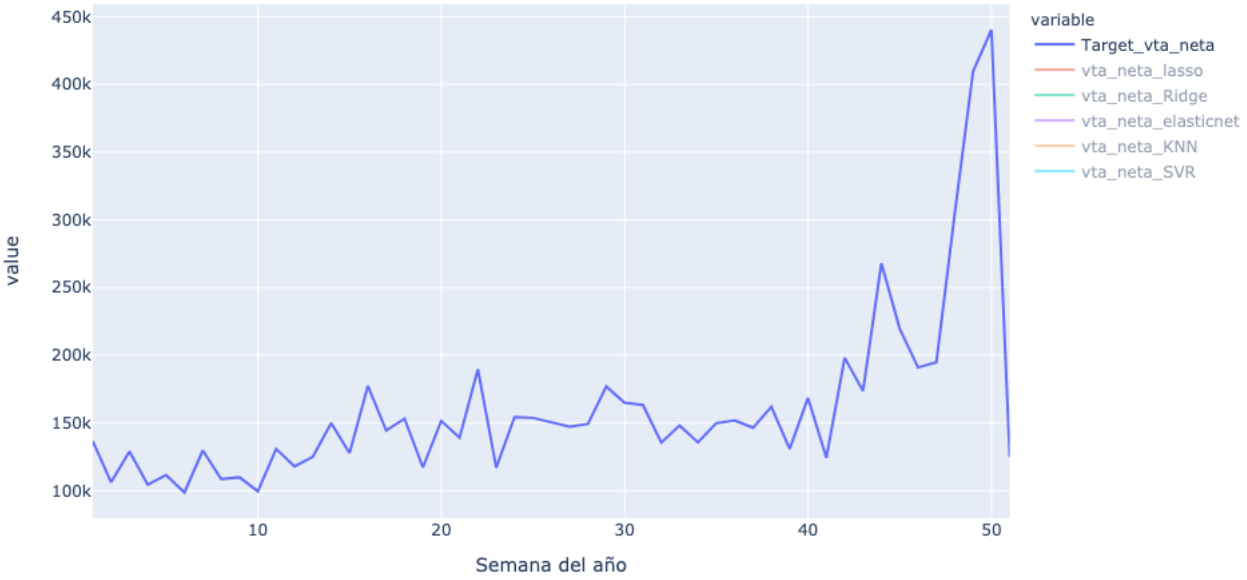
Dado que se tienen variables de entrada de diferentes maginitudes, haremos una estandarización de las variables, llevando todo el conjunto a un mismo espacio de dimensión reducida. Haciendo uso de Standard Scaler, el cual asume que los datos se distribuyen normalmente para cada variable, centrando su distribución alrededor del 0 con una desviación estándar de 1.

0.9.3. Regresión de Cresta

La regresión de cresta es utilizada para modelar fenómenos logrando coeficientes ajustados con menor varianza, dando estabilidad a la predicción del modelo.

La regresión de cresta consiste en generar una línea (aproximación) que se ajuste a los valores reales, pero evitando el sobreajuste, es decir, agrega un porcentaje de ruido a la estimación con la

finalidad de reducir el error estándar de éstos, es por esto que se utilizará para estimar las ventas semanales por categoría.
A continuación, se muestra la tendencia de las ventas semanales reales que percibió la boutique en 2019.

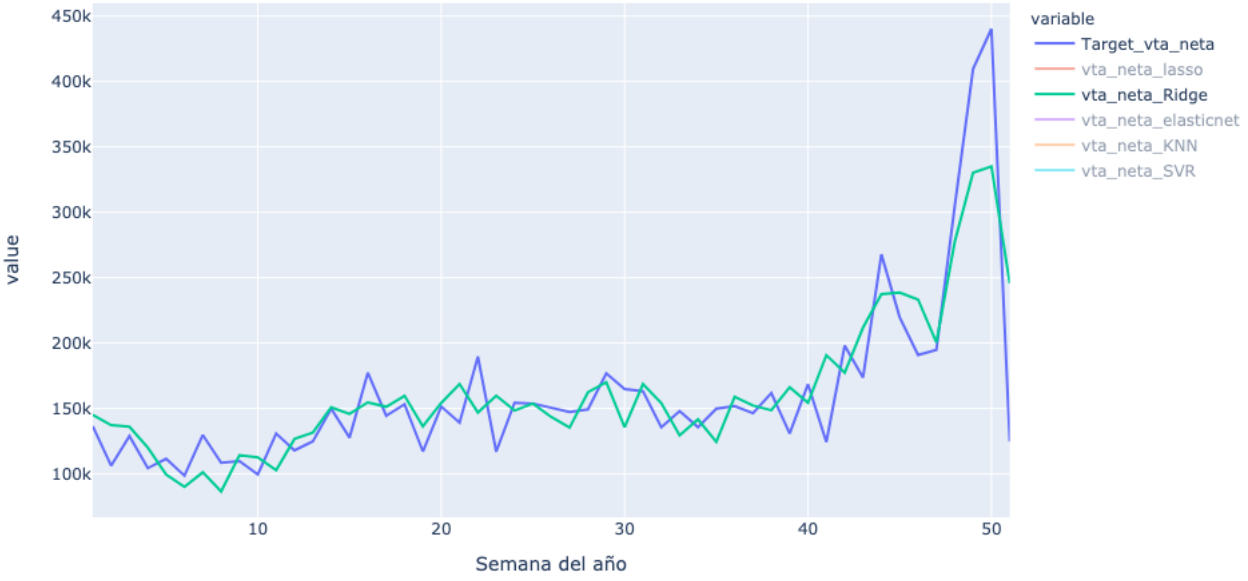


Haciendo uso de la regresión de cresta y teniendo en cuenta las 45 mejores variables a entrar al modelo, se entrena el modelo con solo el 70 % de los datos totales. Para evaluar la precisión del modelo haremos uso de el coeficiente de determinación o R^2, el cual es una medida de la proporción de la variabilidad en las variables de respuesta explicadas por las variables predictoras o de entrada al modelo, por lo que se buscará que el R^2 sea lo más cernano a 1.

Dado lo anterior, se muestran los resultados obtenidos:

Conjunto	Métrica	Resultados
Entrenamiento	R^2	30 %
Test	R^2	25 %

El R^2 obtenido en el conjunto de entrenamiento es mayor relativamente al de test o prueba, por lo que nos habla de un pequeño sobreajuste del modelo, quedando de la siguiente forma las predicciones:



Como se puede observar, la predicción sigue la tendencia creciente de las ventas hacia el final del año, sin embargo no es tan exacta como se esperaba, ya que el Error Absoluto Medio (MAE) equivale al 10 %, por lo que se propone un modelo diferente.

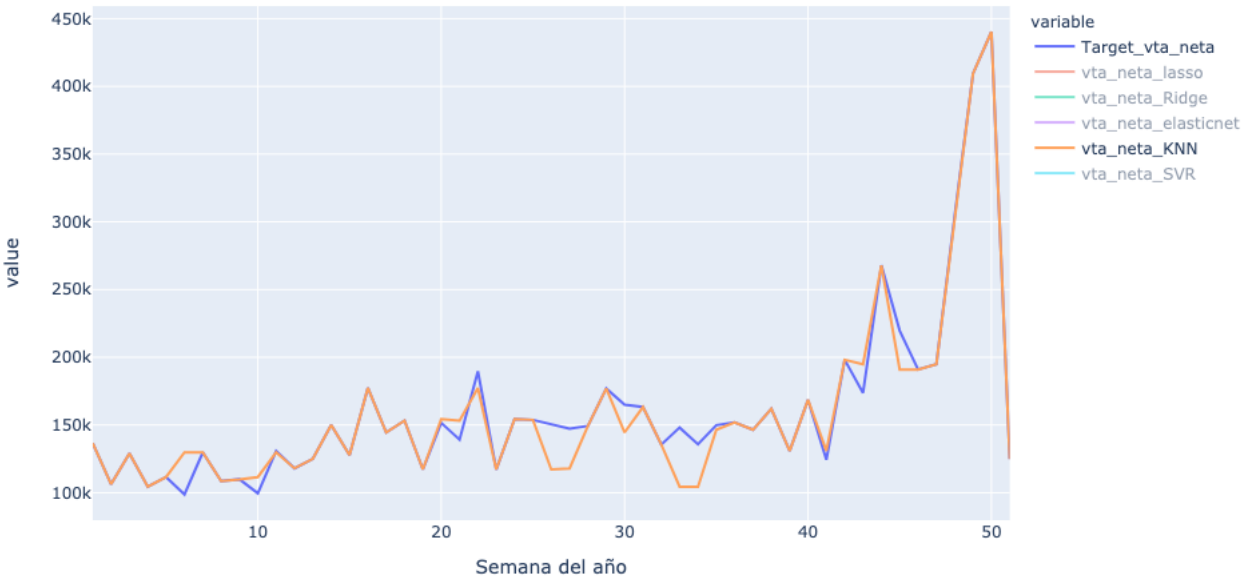
0.9.4. KNN

El algorítmo k-vecinos más cercanos o bien KNN por sus siglas en inglés, es un algorítmo de aprendizaje automático supervisado que se usa para resolver problemas de clasificación y regresión, en este caso, se utiliza como un problema de regresión. KNN, parte de la idea de que las cosas que están cerca se parecen, ya que modela simulaciones similares basadas en distancia.

Siguiendo la misma idea del modelo anterior (45 mejores variables y partición en entrenamiento y prueba), los resultados son los siguientes para KNN:

Conjunto	Métrica	Resultados
Entrenamiento	R^2	50 %
Test	R^2	47 %

Se obtuvo una mejor presición, ahora se muestra la predicción respecto a las ventas reales:



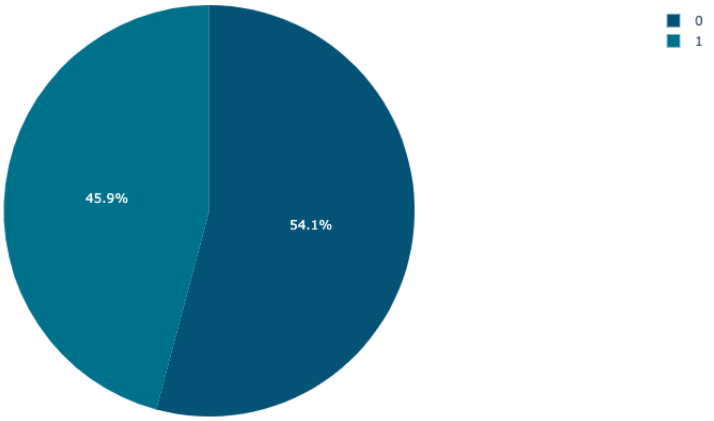
Como se observa, este modelo predice bien el comportamiento que se da en las ultimas semanas del año, sin embargo cuando se equivoca el error es mayor, tal es el caso de las semanas del 20 al 35, obteniendo un MAE equivalente al 18 %.

0.10. Modelo de fuga de clientes

En la actualidad, existe una gran competencia en las tiendas online, si bien, el éxito de tiendas conocidas se basa en los productos únicos que ofrecen y en la publicidad que hacen por medio de redes sociales. Sin embargo este panórama indica que la administración de la retención de clientes es una clave importante para mantener y/o mejorar la posición de la tienda de ropa en el mercado.

Debido a lo anterior y desde una perspectiva de inteligencia de negocio, se realizó un modelo de fuga de clientes bajo dos perspectivas, la primera conocer la situación de la cartera de compradores de la Boutique, es decir, predecir una potencial fuga de estos, y la segunda es aplicar medidas preventivas para evitar la fuga, las cuales pueden ser campañas de marketing o descuentos a los productos, para los clientes con una mayor probabilidad de irse.

Para este análisis, se define como cliente de la Boutique a aquella persona que ha comprado más de dos veces en la tienda, y se toman como clientes inactivos a aquellos que no realizaron compras en los últimos dos meses del 2019, por lo que la distribución de clientes activos (0) e inactivos(1) es la siguiente:

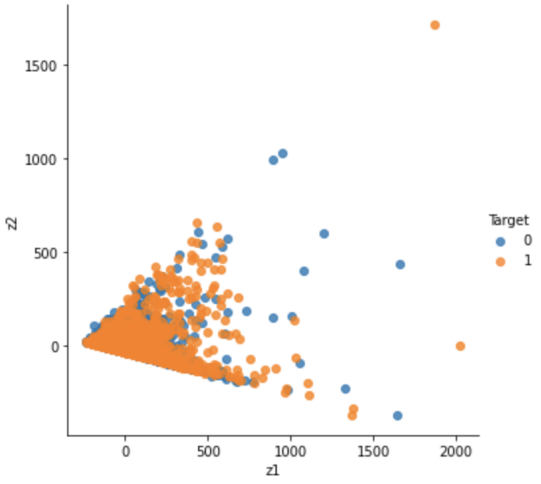


0.10.1. Reducción de dimensiones

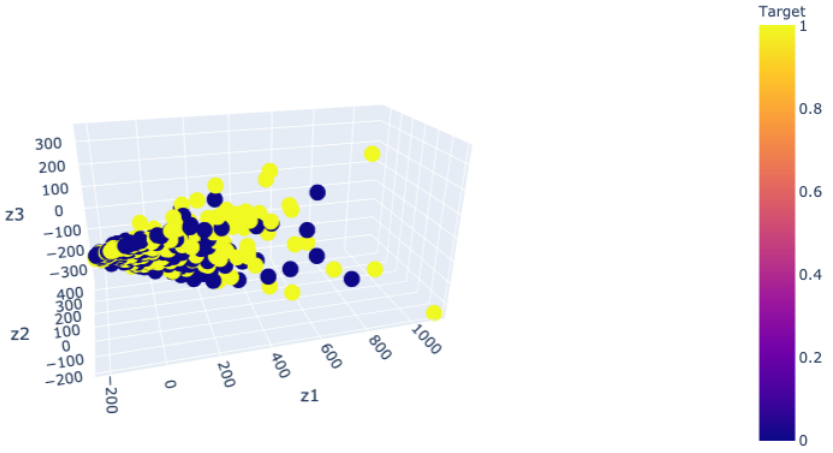
PCA: Principal Component Analysis

El análisis de componentes principales es una técnica multivariada que deforma el espacio del cuál provienen los datos, de forma que genera componentes que serán ortogonales, es decir, reduce la dimensión con la cual se trabaja.

Para poder aplicar esta técina, se requiere que las variables cumplan ciertas características, es decir, se requiere que sean Normales (estandarizadas), con media 0 y varianza 1, además de ser independientes, es decir, lo que comunmente se ve en estadística como ï.i.d N(0,1)".
Tomando en cuenta la variable de clientes inactivos y activos creada, se obtiene una visualización en 2 dimensiones, la cual se muestra a continuación:



A primera vista, en la tabla anterior, las observaciones no se aprecian linealmente separables, cabe mencionar que la varianza explicada es del 80 %,por lo que llevándolo a una dimensión superior queda de la siguiente forma:



0.10.2. Transformación Entrópica

WOE:Weight of Evidence

El WOE, indica el poder predictivo de una variable independiente en relación con la variable dependiente. Una vez utilizado el WOE, para medir la potencia de cada variable respecto a la target, se usa la medida conocida como Information Value o IV, el cual se calcula de la siguiente forma:

$$IV = \sum \left[\mathbb{P}\left(NoEvento\right) - \mathbb{P}\left(Evento\right)\left(WOE\right)\right]$$

Dado lo anterior, el IV por cada variable es el siguiente:

1. tipo_familia : 0.005
2. share_of_wallet :0.008
3. NSE: 0.006
4. Store :0.16

Las variables predictivas que influyen en la permanencia de un cliente son la variable de tipo familia, el share of wallet, el nivel socioeconómico y la tienda en donde realizó la compra, siendo esta la que más aporta información.

0.10.3. Regresión Logística

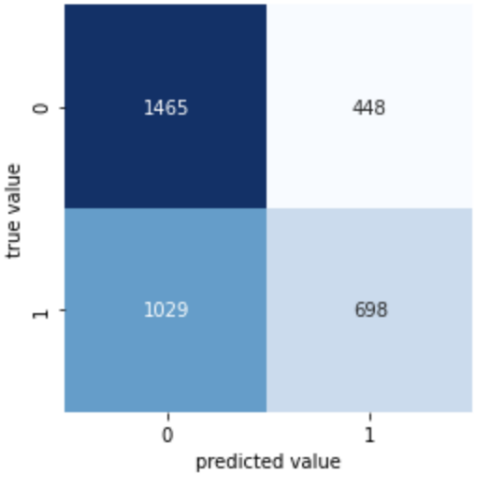
La regresión logística, al igual que otras técnicas estadísticas multivariadas, es muy utilizada en la industria para modelar la relación entre una variable de respuesta binaria, en este caso la variable target, y un conjunto de variables predictoras. Teniendo en cuenta las variables agrupadas a nivel cliente, se entrena el modelo con el 70 % de los datos, dejando el resto para prueba del modelo.

Una vez entrenado el modelo, las métricas para comprobar que el modelo generaliza correctamente y evaluar su desempeño son las siguientes:

1. Accuracy score : Proporción de las clases correctamente clasificadas.
2. Curva ROC : Representa la sensibilidad del clasificador binario, es decir, en dónde es más probable que el modelo se equivoque.

Los resultados del entrenamiento del modelo son los siguientes:

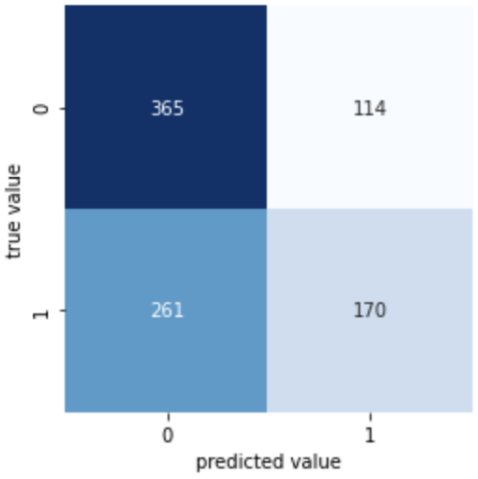
	Resultado
ROC	63 %
Accuracy Score	60 %



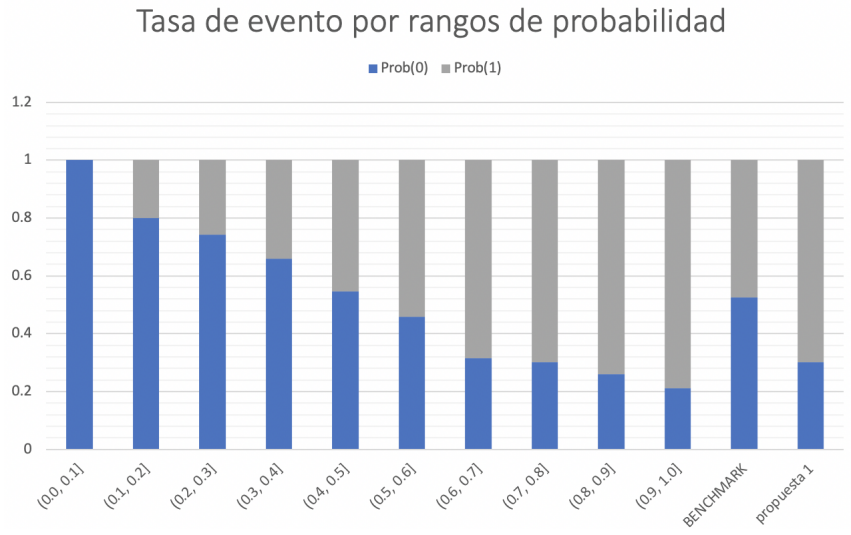
En especial buscaremos mantener el Score igual para el conjunto de entrenamiento y el de test, ya que como se observa, al entrenar el modelo se esperaría que si el modelo indica que el cliente no se va y realmente sea fuga, se estaría perdiendo la oportunidad de retenerlos, por lo que se busca mantener los falsos positivos relativamente bajos.

Por otra parte, los resultados para el conjunto de prueba son los siguientes:

	Resultado
ROC	62 %
Accuracy Score	59 %

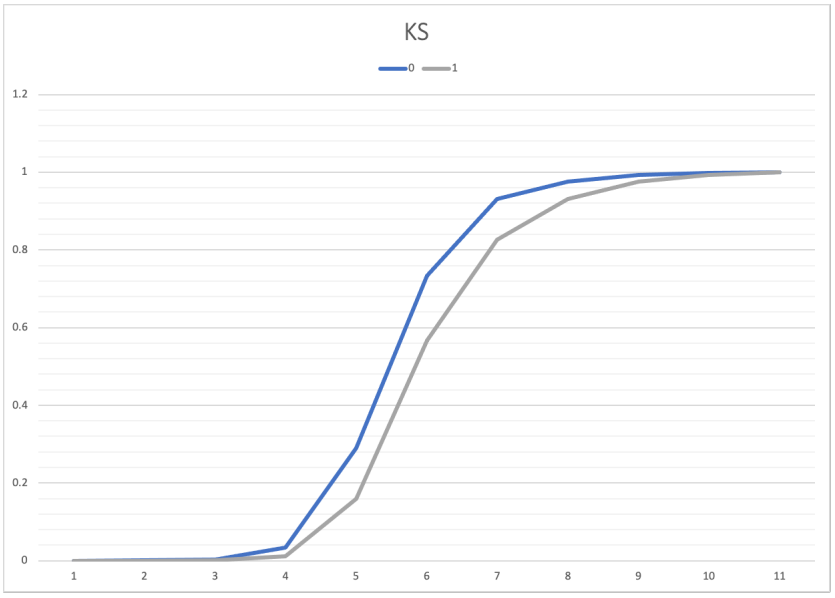


Lo anterior indica un buen comportamiento del modelo, ya que generalizó de manera correcta, sin embargo veremos su desempeño. En la siguiente grafica se muestra la tasa de evento por diferentes rangos de probabilidad:



Como se observa, el modelo aprendió un patrón dentro del comportamiento de abandono de los clientes, por lo que el Benchmark indica la tasa de fuga que se tiene si no se toma ninguna acción para retener a los clientes, la cual es 47 %. Lo que se propone es que tomando en cuenta a los clientes que tienen una probabilidad mayor al 6 % de fuga, se les otorgue algún cupón de descuento o alguna promoción de compra que aplique en la subcategoría o para productos similares a los que ha comprado, así logrando la permanencia los que están más probables a irse y limitando los recursos monetarios disponibles para campañas de publicidad o de descuentos.

En la siguiente gráfica, se muestra la distribución de los eventos 0 y 1, de lo cual se espera que estas distribuciones estén lo más separadas posibles, ya que es de interés que ambas se distribuyan diferente.



0.11. Sistema de recomendación