ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Facultad de Ingeniería en Electricidad y Computación

ANÁLISIS PREDICTIVOS DE VENTAS DE CATEGORÍAS DE PRODUCTOS EN UNA EMPRESA DE RETAIL CON UN SISTEMA DE RECOMENDACIONES DE COMPRA PARA CLIENTES, EMPLEANDO TÉCNICAS DE VISUALIZACIÓN DE DATOS.

PROYECTO DE TITULACIÓN

Previo la obtención del Título de:

Magister en Ciencias de Datos

Presentado por:

JAIR ISRAEL VILLAO GÓMEZ

JONATHAN WASHINGTON BARRE ROMERO

GUAYAQUIL - ECUADOR

Año: 2022

DEDICATORIA

A mi tío y mis tías que me han apoyado desde pequeño tanto moral, emocional y psicológicamente.

A mi mami Martha (†) que me guía en las decisiones que tomo y con su ejemplo me motiva a no rendirme.

Jair Villao

A Dios, a mi madre y a mi familia quienes con mucho sacrificio y esmero me han permitido llevar a cabo una meta más, gracias por darme la fuerza para seguir adelante.

A las personas que me ayudaron de una manera desinteresada cuando más lo requerí, por extenderme su mano en situaciones difíciles, gracias por creer en mí.

Jonathan Barre

AGRADECIMIENTOS

A Dios que está presente en cada paso que doy. A mis familiares por el apoyo incondicional a lo largo de mi vida.

A mis amigos y compañeros que contribuyeron con enseñanzas para lograr mis objetivos. A la empresa donde laboro por facilitar la información del presente proyecto.

Jair Villao

A mi madre por haber sido mi mayor apoyo durante este proceso, gracias por ser el pilar fundamental en mi vida.

A mi tío Zenon por el apoyo que me ha venido brindando a lo largo de mi vida, gracias por ser el mejor ejemplo de hombre, padre y esposo.

A todos mis amigos y personas que ayudaron de alguna u otra manera para poder lograr el objetivo, quedaré eternamente agradecido con todos.

Jonathan Barre

DECLARACIÓN EXPRESA

“Los derechos de titularidad y explotación, nos corresponde conforme al reglamento de propiedad intelectual de la institución; *Jair Israel Villao Gómez y Jonathan Washington Barre Romero* damos nuestro consentimiento para que la ESPOL realice la comunicación pública de la obra por cualquier medio con el fin de promover la consulta, difusión y uso público de la producción intelectual"

Jair Israel Villao Gómez

Firmado electrónicamente por:
JONATHAN WASHINGTON BARRE ROMERO

Jonathan Washington Barre Romero

COMITÉ EVALUADOR

Firmado electrónicamente por:
SERGIO ALEX BAUZ OLVERA

Dr. Sergio Bauz

PROFESOR TUTOR

Firmado electrónicamente por:
JOSE EDUARDO CORDOVA GARCIA

Dr. José Córdova

PROFESOR EVALUADOR

RESUMEN

El sector retail utilizado básicamente para referirse al comercio minorista, provee una gran variedad de productos y servicios a distintos segmentos de consumidores (Personas). En el Ecuador uno de los giros de negocios de este tipo de industria comercializadora; es el especializado en la venta de electrodomésticos de diferentes marcas a nivel mundial. Las empresas ofrecen sus catálogos de productos a los clientes por distintos medios; tecnológicos o autoservicios, acorde a las compras realizadas en ventas pasadas de los clientes frecuentes y a las características demográficas (Edad, Genero, Estrato Social, Ingresos, etc.). Con ello surge la necesidad de establecer cuál es el mejor producto o segmento de producto que debería una empresa de retail ofertar para que el cliente reciba una atención personalizada por distintos canales de compra y así aumentar su rentabilidad. Otra problemática que surge es la rotación de sus productos que va de acorde a la demanda, y es que si no posees stock de una línea de producto que tiene una alta rotación pierdes oportunidad de venta y esto genera perdida en los ingresos de la empresa, debido a que no se establecen los mecanismo y metodologías necesarias para predecir correctamente la demanda de cada segmento de producto. El presente proyecto realizará un estudio para solventar ambas problemáticas “Demanda de Línea de productos” y “Recomendaciones de Sublíneas de Productos”, para aquello evaluaremos la información de una empresa comercializadora de electrodomésticos a la cual denominaremos “AAA” por temas confidencialidad de la información. Se mejorará los análisis en varios sectores de la empresa implementando técnicas de predicción y de Machine Learning acrecentando la fidelización de clientes recurrentes y sobre todo incrementado la participación en el mercado ecuatoriano. Se obtendrán predicciones de las categorías de productos representativas de la empresa mediante una combinación hibrida de series temporales (ARIMA) y de regresión múltiple, permitiéndonos tener predicciones ajustadas a la realidad con niveles de confianza óptimos para los encargados de dichas categorías, logrando abastecerse en determinadas épocas del año. Estableceremos un sistema de recomendaciones de categoría de producto para los clientes de la empresa, se buscarán clúster con perfiles de clientes semejantes para poder recomendar una categoría en base a comportamiento del cliente y así poder dar promociones focalizadas logrando la

fidelización del cliente. Comparamos los resultados de las metodologías implementadas versus las vigentes con la cual se comprobó que las predicciones mejoran en las líneas de producto. Las recomendaciones de categorías arrojadas por el modelo nos permitieron tener mayor concreción de venta al realizar campañas específicas a los clientes.

Palabras Clave: Pronóstico, Series de Tiempo, Sistema de Recomendaciones.

ABSTRACT

The retail trade, basically used to refer minor commerce, provides a wide variety of products and services to different segments of consumers (people). In Ecuador, one of the business lines of this type of marketing industry; is specialized in the sale of home appliances of different worldwide brands. The companies offer their catalog of products to their client portfolio over different means; technological or self-services, according to the purchases made in past sales of frequent customers and to demographic characteristics (Age, Gender, Social Stratum, Income, etc.). With that said, it arises the need to establish which is the best product or product segment that a retail company should offer to the customer to receive personalized attention by different purchase channels and thus increase their profitability. Another problem that arises in the home appliance companies is the rotation of their products that goes according to the demand of these, and that is, if you do not have stock of a line of product that has a high rotation you lose an opportunity to sale and it generates loss in the company income due to the mechanisms and methodologies necessary to correctly predict the demand of each product segment are not established. This project will elaborate a research to solve both issues "Product line demand" and "sublimes products recommendation" , for that, we will evaluate the information of an appliance company, a.k.a. "AAA" as confidentially. Analysis will be improved in several areas of the company implementing prediction techniques and Machine Learning enhancing fidelity of recurrent clients and, especially, increasing the participation in the Ecuadorian market. It will be obtained predictions of the representative products categories of the company through a hybrid combination of temporal series (ARIMA) and multiple regression, allowing us to have predictions tied to the reality with optimal confidence levels for the persons in charge of those categories, managing to stock up at certain seasons in a year. It will be established a recommendation system of products category for the clients of the company, it will be searched for clusters with profiles of similar clients to recommend a selection category based on a behavior of the client in the way to provide focused promotions obtaining the client fidelity.

We will compare the results of the implemented methodologies versus current methodologies which proved that the predictions are improved in the product line. The recommendations of the categories given by the model allow us to have concretion of sales when carrying out specific campaigns to customers.

Keywords: forecast, time series, recommendation systems

ÍNDICE GENERAL

ABREVIATURAS

ESPOL	Escuela Superior Politécnica del Litoral
ARIMA	Autoregressive Integrated Moving Average
INEC	Instituto Nacional de Estadísticas y Censos
PIB	Producto Interno Bruto
IPC	Índice del Precio al Consumidor
VAR IPC	Variación de Índice del Precio al Consumidor
ENEMDU	Encuesta Nacional de Empleo, Desempleo y Subempleo
AIC	Criterio de Información Akaike
MS SQL	Gestor de Base de Datos Relacionales
IA	Inteligencia Artificial
ML	Machine Learning
SVD	Descomposición Singular de Valores
NFM	Matrices de Factorización no Negativa
KNN	K-vecinos más Cercanos

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

CAPÍTULO 1

1. PLANTEAMIENTO DE LA PROBLEMÁTICA

1.1 Descripción del problema

Las empresas de retail manejan diferentes líneas de productos que ofrecen a sus clientes para aquellos que les compran de forma recurrente y a los nuevos que ingresan a consumir su catálogo, para ofrecer una variedad de alternativas las empresas compran sus productos a diferentes proveedores de acuerdo con el histórico de venta que tienen registrados para poder cumplir con sus clientes. Básicamente en los últimos años han estado incursionando en análisis predictivos para determinar en qué fechas se venden determinados productos de acuerdo con el comportamiento del cliente.

Las empresas recomiendan a sus clientes artículos de acuerdo con las características de compra y de forma empírica en determinados meses del año. Por lo general promocionan productos en base a un evento local o mundial y de acuerdo con las festividades que se tienen a lo largo de los años.

En la empresa “AAA” se obtiene un presupuesto de ventas en base a un modelo de predicción empírico, teniendo como resultado estimaciones no confiables, originando molestias para los encargados de las negociaciones de compra ya que utilizan este presupuesto en los diferentes meses del año generando un bajo rendimiento en sus indicadores, afectando a todas las áreas en gestión de venta; se identificó algunas repercusiones como:

- Devaluaciones de productos dado a las malas predicciones iniciales.
- Perdida de ventas al no tener ciertas categorías en determinadas épocas del año.
- Disminución en la participación del mercado al no contar con suficientes productos en meses de alta compra.

La empresa gestionó un plan estratégico para tratar de ofrecer estos artículos a sus clientes recurrentes sin tener el efecto deseado ya que muchos de ellos ya

habían comprado recientemente alguno de los productos o por el simple hecho de que estaban interesando en otros, descubriendo otro punto débil en su modelo de negocio “no conocen a sus clientes recurrentes” por lo que se plantea resolver estas problemáticas y lograr mayor competitividad en el mercado fidelizando a sus clientes.

1.2 Justificación del problema

En la actualidad la empresa de retail “AAA” posee procedimientos definidos para la proyección de ventas, administrado por el departamento de planeación y presupuesto el cual es realizado por las diferentes categorías de producto teniendo como objetivo lograr una mayor participación en el mercado sin perder oportunidades de ventas. El departamento realiza estas proyecciones con metodologías empíricas sin una base estadística, manteniendo un gran margen de error en sus proyecciones de dichas categorías.

Por otra parte, los jefes de producto que trabajan en base a estos presupuestos de ventas en los diferentes meses del año reciben quejas continuas por las áreas de ventas, logística y crédito ya que no identifican en que meses del año deben tener más stock de los productos.

Otra de las áreas afectadas indirectamente por este insumo entregado por el departamento de planeación es el área de campaing que promociona artículos por diferentes medios electrónicos sin tener claro la característica del cliente para ofrecerle un producto con alta probabilidad de compra, ya que se guían por decisiones de “expertos” basadas en los totales de ventas mensuales por categorías sin un análisis estadístico, por lo que se debe hacer un estudio del perfil del cliente previo a ofertar cierto producto. Esto genera una disminución en la concreción de venta, perdiendo participación en el mercado.

Por tanto, el presente proyecto plantea solucionar la problemática interna de la empresa de los siguientes hitos:

- Calidad de los modelos de predicción de categorías. - Se implementará modelos estadísticos para predecir el comportamiento de ventas de las categorías top, potenciando la planificación de los jefes de producto y la rotación de las categorías.
- Creación de recomendaciones de categorías a los clientes. - Nos enfocaremos en los perfiles de los clientes recurrentes de la empresa determinando el comportamiento de compra obteniendo recomendaciones personalizadas para cada uno de ellos.
- Implementar la herramienta de visualización. - Los usuarios tendrán un dashboard con pronósticos efectivos y sistema de recomendaciones eficiente sin necesidad de solicitar al departamento de analítica dicha información.
- Incremento de ventas. - Logrando pronósticos efectivos e incrementando nuestra concreción de ventas a nuestros clientes recurrentes con campañas personalizadas.

1.3 Objetivos

1.3.1 Objetivo General

Desarrollar modelos predictivos para pronósticos efectivos de ventas y sistema de recomendaciones de las principales categorías de productos para la empresa “AAA”.

1.3.2 Objetivos Específicos

- Analizar modelos predictivos con la finalidad de establecer tendencias en dos categorías top en base al comportamiento histórico.
- Sugerir recomendaciones de categorías de producto a los clientes con perfiles previamente analizados mediante analíticas.
- Mostrar visualizaciones interactivas de los resultados obtenidos mediante un dashboard que permita tomar decisiones optimas y oportunas.

1.4 Resultados esperados

La empresa “AAA” obtendrá un dashboard amigable e interactivo que estará alojada en un local host para el consumo de los usuarios. Las visualizaciones mostraran las predicciones de ventas de las principales categorías de producto con la mejor analítica probada para el set de datos con un buen nivel de confianza, con esto la empresa reducirá los errores presentados en las predicciones que se utilizaban en los últimos años, visualizando de mejor manera la tendencia de venta, mejorando las colocaciones y penetración en el mercado.

Los usuarios encargados de la toma de decisión (marketing – venta), tendrán un sistema de recomendación el cual mostrará la categoría de producto que el o los clientes consultados tienen más posibilidad de compra para que lo puedan gestionar de la manera más adecuada tanto por llamada, mensaje de textos o en el punto de venta.

1.5 Dataset

La empresa almacena todas las transacciones de venta (incluyendo las características del cliente) en una base de datos de MS SQL en diferentes tablas estructuradas. Actualmente por política de la empresa se tiene en el servidor de consulta información de venta desde enero del 2016; se seleccionó el dataset desde este año hasta junio del 2021 para ajustar los modelos propuestos.

En la data transaccional tenemos 19 variables que representan las características de los clientes, variables calificativas que se le da internamente en base a su perfil de riesgo e información de la venta como producto, método de pago, etc.

Previamente contamos 2.5 millones de registros, al depurar la data y eliminar en primera instancia las facturas que causaron notas de crédito, perdurando 2 millones de registro y 812.361 clientes únicos para realizar los análisis. A continuación, detallamos las variables transaccionales de venta que utilizaremos.

Tabla 1-1 Descripción Variables Transaccionales

VARIABLES	TIPO	DESCRIPCION
Fecha	**Date**	Fecha de compra
Cadena	**chr**	Cadena en la cual el cliente compro (para el análisis se va a considerar las 2 cadenas de electrodomésticos).
Ciudad	**chr**	Ciudad de compra
TipoVenta	**chr**	Si el cliente compra con Tarjeta, en efectivo o usa el crédito directo de la empresa.
Plazo	**num**	Numero en meses de financiamiento de la compra
Cantidad	**num**	Cantidad de artículos que compro el cliente
ValorFactura	**num**	Valor del articulo a crédito antes del financiamiento (Si es venta en efectivo o tarjeta esta variable es igual a valorContado).
sublinea	**chr**	Subcategoría del articulo
marca	**chr**	Proveedor del articulo (Indurama, Sony, LG, etc)
Tarjeta	**chr**	Nombre de la tarjeta con la cual realizo la compra
PromocionTarjeta	**chr**	Descripción del plazo de financiamiento de la compra con tarjeta
tipocliente	**chr**	Si es un cliente Nuevo o si ya nos compró anteriormente (Establecido), aparece cuando es compra a crédito
perfilCliente	**chr**	categoría del cliente, si está afiliado o no. Aparece cuando es compra a crédito
sexo	**chr**	Femenino o Masculino
edad	**int**	Edad del cliente
cargo	**chr**	Cargo del cliente
ingresos	**num**	Ingresos presuntivos del cliente en base al cargo
PuntajeScore	**chr**	Puntaje sobre el nivel de riesgo del cliente (este puntaje varía de acuerdo con los créditos que tiene)
Valor Cuota	num	Valor de la cuota que debe pagar el cliente por motivo de financiamiento

Se incluyeron ciertas variables económicas y de empleo que denominaremos variables externas para obtener ruido y mejores predicciones en los modelos. Se decidió incluir estas variables para determinar cuáles son significativas y poder mejor nuestras predicciones en los diferentes meses del año.

Tabla 1-2 Descripción Variables Externas

VARIABLES	DESCRIPCIÓN
IPC	Es una medida de la evolución temporal del nivel general de precios, en relación con el consumo de los hogares
Var IPC	Variación del IPC mensual
canasta familiar basica	imprescindibles para satisfacer las necesidades básicas del hogar
canasta familiar vital	mínimo alimentario que debe satisfacer por lo menos las necesidades energéticas y proteicas de un hogar
Precio barril de petroleo	Precio promedio del Barril de petróleo mensual
Desempleo abierto	Tasa de desempleo por periodos
Empleo Adecuado	Tasa de empleo adecuado por periodos
SUBEMPLEO INSUFICIENCIA INGRESOS	Tasa de subempleo por insuficiencia de ingresos.

El dataset interno de la empresa tiene la posibilidad de actualizarse de manera diaria (día caído) para el presente estudio las ventas serán agruparán por mes y categorías; en el caso del sistema de recomendaciones lo actualizaremos diariamente, las variables externas “Tabla 1-2” son calculadas al finalizar cada mes, pero para las predicciones de las ventas obtendremos un forecast de estas variables para poder incluir en el modelo de predicción.

CAPÍTULO 2

2. MARCO TEÓRICO Y ESTADO DEL ARTE

2.1 Marco teórico

2.1.1 Series de tiempo

Una serie de tiempo es una secuencia de observaciones, medidos dentro de varios periodos de tiempo, los cuales a menudo son del mismo tamaño (Céspedes Urrutia, 2017).

Para (J. Villavicencio, 2010) una serie temporal es un conjunto de observaciones medidas a través del tiempo, ordenadas cronológicamente y espaciadas uniformemente, de modo que los datos dependan unos de otros. Teniendo como objetivo principal el de realizar pronósticos.

En otras palabras, las Series de tiempo son utilizadas en el ámbito estadístico con el fin de verificar y analizar el comportamiento de una o varias variables a lo largo del tiempo, variables ordenadas cronológicamente en un periodo establecido con el objetivo de ajustar dichos datos de la serie temporal a modelos teóricos ya establecidos.

2.1.2 Componentes de Series de Tiempo

Entre las terminologías prescritas en las series temporales se encuentran varias componentes esenciales e indispensables para poder entender el comportamiento descriptivo y predictivo de las series.

Para (J. Villavicencio, 2010) el análisis tradicional de series temporales se compone en el supuesto de que los valores que toma la variable de observación son consecuencia de tres componentes, cuya acción conjunta es traducida a valores medidos, los componentes son:

- **Componente tendencia. -** Se define al comportamiento de la serie temporal que produce un cambio a largo plazo desde el nivel medio de la media. La tendencia se podría representar con una línea recta de forma creciente o decreciente.

- **Componente estacional. -** Muchas series de tiempo tienen cierta periodicidad es decir una variación constante en ciertos períodos de tiempo (anual, mensual, etc.). Por ejemplo, las ventas de un negocio aumentan en los meses de noviembre y diciembre durante las promociones del viernes negro y festividades navideñas. Estas características son sencillas de detectar por sus oscilaciones periódicas, a su vez pueden ser incluidas o excluidas del set de datos y así contemplar ambos escenarios.

- **Componente aleatoria. -** Las series de tiempo poseen características aleatorias en su conjunto de observaciones, este comportamiento debe ser aislado y estudiado posteriormente bajo modelos probabilísticos que describan dicho comportamiento.

2.1.3 Clasificación descriptiva de las series temporales

Siguiendo con la ideología de (J. Villavicencio, 2010) establece dos tipos de clasificación de series temporales:

- **Estacionarias. -** Muchos autores describen a una serie de tiempo como estacionaria cuando es invariable o inalterable a lo largo del tiempo, es decir cuando poseen una media y una varianza constantes en el tiempo evaluado. Esto se traslada gráficamente al hecho de que las series tienden a oscilar alrededor de una constante.

- **No estacionarias. -** Son series cuya media y/o variabilidad son alteradas respecto al tiempo, es decir varían al transcurrir el tiempo. Estas variaciones en la media determinan una tendencia positiva o negativa a largo plazo.

2.1.4 Modelo Regresión Múltiple

La regresión lineal múltiple, según (Cruz Trejos, 2011) es una metodología matemática y estadística que conforma la relación existente entre la variable explicada "Y" con las distintas variables explicativas o regresoras "Xi" y el término del error aleatorio ε. Se detalla los objetivos de aplicar los modelos de regresión:

- Comprobar si la variable dependiente esta correlacionada con las distintas variables regresoras.

- Dados valores de las distintas variables regresoras, predecir el dato de la variable dependiente.

2.1.5 Supuestos Regresión Múltiple

Para (Uriel, 2013) la realización de estos modelos de regresión necesita cumplir con los supuestos que se detallan a continuación:

- **Linealidad. -** Este supuesto estable que la relación que se determina entre la variable dependiente e independiente debe ser lineal; es decir, que deben mantener una correlación lineal. Una forma posible de verificar que existe linealidad es mediante diagramas de dispersión o mediante el indicador de Pearson oscilante entre -1 y 1.

- **Homocedasticidad. -** La homocedasticidad ocurre cuando la varianza de los errores es constante; es decir, no varía para ninguna variable independiente.

$$Var(\mu t) = \sigma 2 \; t=1, 2,..,n. \quad (1)$$

- **Normalidad. -** El supuesto de normalidad indica que todos los errores poseen un comportamiento ajustable a una distribución normal con media cero. El comportamiento de estos errores hace factible el cálculo de intervalos de confianza.

$$\varepsilon \sim N(0,\sigma 2) \quad (1)$$

  Una forma de probar el supuesto de normalidad es con un histograma o realizando pruebas de hipótesis de normalidad.

- **No colinealidad. -** La no colinealidad básicamente implica que las variables independientes o explicativas no tengan ningún tipo de correlación lineal. Mientras que la multicolinealidad ocurre cuando existe dependencia alguna entre las variables explicativas.

- **Independencia. –** La hipótesis de independencia indica que la diferencia entre los valores reales y los pronosticados (residuos) no deben estar correlacionados entre sí, es decir que sean independientes. Para corroborar esta hipótesis se puede realizar el diagrama de dispersión o pruebas de independencia.

$$E(\varepsilon t|Xt)=E(\varepsilon t)=0 \; t=1,2,\ldots..,n \; (1)$$

2.1.6 Modelos ARIMA

En estadística un modelo autorregresivo integrado de media móvil o ARIMA (Carrillo, 2019) es una método estadístico plasmado en procesos dinámicos que utilizado para series temporales, variaciones y regresiones con el fin de encontrar patrones y predecir valores futuros en base a observaciones y comportamientos pasados de alguna característica o variable modificada a través del tiempo. Se lo denota como:

$$Y't=\delta+\varphi 1Y't-1+\cdots+\varphi pY't-p-\theta 1at-1-\cdots-\theta qat-q+at \quad (1)$$

$$Y't-\varphi 1Y't-1-\cdots-\varphi pY't-p=-\theta 1at-1-\cdots-\theta qat-q+at \quad (2)$$

$$(1-\varphi 1L-\cdots-\varphi pLp)(1-L)dY't=(1-\theta 1L-\cdots-\varphi qLq)at \quad (3)$$

Donde:

$Y't$: es la variable a predecir diferencial

p: parámetro autorregresivo

q: parámetro medio móvil

d: diferencias

L: retardo

at: es el término del error

Este modelo también permite modelar datos estacionales, utilizando los siguientes términos: ARIMA (p,d,q) (P,D,Q)s, donde P es el parámetro autorregresivo del proceso estacional, Q el parámetro de media móvil, D es el parámetro de diferenciación la parte estacional y s representa la periodicidad de la serie.

Uno de los criterios que son de útiles para determinar el orden del modelo es el Criterio de información Akaike (AIC).

2.1.7 Modelo ARMA(p,q)

Se trata de un modelo mixto con componente autorregresiva y otra de media móvil. el modelo autorregresivo se basa en la idea de que la observación actual puede explicarse con valores pasados, estos modelos se denotan como AR(p) siendo p el número de muestras que explican la muestra actual.

El modelo de medias móviles consiste en realizar una aproximación a la serie utilizando únicamente ruido blanco. La expresión general de los modelos MA(q), juntando ambos términos obtenemos los modelos ARMA (p,q).

2.1.8 Modelo ARIMA(p,d,q)(P,D,Q)

Para obtener la estacionalidad de una serie se tiene I(d) que es la forma de denotar el concepto de esta integración, siendo d el número de diferencias necesarias para lograr estacionalidad. Las siglas (P, Q, D) hacen referencia a un ARIMA estacional, se lo conoce también como SARIMA.

Los términos del primer paréntesis (p, d, q) son para modelar la parte regular, es decir, la dependencia asociada a observaciones consecutivas, y los del segundo paréntesis (P, D, Q) la estacionalidad, que está asociada a observaciones separadas por n periodos.

2.1.9 Test de los modelos

- **Test ADF. –** La prueba de Dickey-Fuller aumentada identifica si la serie presenta tendencia estocástica mediante un contraste de hipótesis.

- **Test KPSS.** – La prueba de Kwiatkowski-Phillips busca determinar si la serie temporal es estacionaria en base a contraste de hipótesis.
- **Test Shapiro-Wilks. –** La prueba de Shapiro-Wilks busca contrastar si la serie tiende a una distribución Normal, este test es aplicable cuando se tiene una muestra mayor a 50 elementos.

2.1.10 Modelo Regresión con Errores ARIMA

El modelo de Regresión de errores ARIMA permite la combinación de datos históricos e información considerable sobre variables predictoras. Su metodología es similar a un modelo de regresión estándar, con una diferencia que radica en términos del error.

$$Yt=\beta 0+\beta 1X1,t+\beta 2X2,t+\cdots+\beta kX\mathrm{n},t+at \quad (1)$$

Yt es una función lineal de las n variables predictoras

$$(X1,t,X2,t,\ldots,Xk,\mathrm{n}) \quad (2)$$

Considerando que, en modelos de regresión estándar, el término del error es ruido blanco; es decir, no existe correlación alguna entre sus errores, en la metodología de regresión con errores ARIMA permite que los errores contengan autocorrelación. Al reemplazar el término at por εt en la ecuación, la serie temporal de regresión con errores nt asume que el término del error procede de un modelo ARIMA. (Sampín, 2021)

2.1.11 Sistema de Recomendaciones

Los sistemas de recomendaciones son herramientas que utilizan algoritmos de inteligencia articficial (IA) y de Machine Learning (ML) para ofrecer recomendaciones en tiempo real como un mecanismo de aprendizaje automático. Las sugerencias pueden variar para cada usuario dependiendo de la cantidad de datos y características que se desee incluir, el tiempo de respuesta dependerá de la tecnología usada.

Cuando se diseña un mecanismo personalizado para cada usuario, se utiliza datos enfocadas a las características de cada uno y sus preferencias para poder tener recomendaciones relevantes recopilando información de manera constante facilitando el proceso de toma de decisiones.

La eficacia de las sugerencias está directamente ligada a la cantidad de los datos obtenidos y su depuración previa. Así, cuantos más datos haya de una y otras personas parecidas a él, mejores y personalizadas podrán ser las sugerencias, avivando el interés de este impactando en la toma de decisiones de manera oportuna (d'Arc, s.f.).

Existen diferentes sistemas de recomendaciones desde películas, músicas, publicad y de productos personalizados de acuerdo con la necesidad del negocio, así como el del usuario final obteniendo buenos resultados.

2.1.12 Filtrado Colaborativo - Descomposición en valores singulares (SVD)

Se denomina Descomposición Singular de Valores (SVD) a la técnica de factorización de matrices que nos ayuda a descomponer y convertir una matriz Q en otras matrices W, X, Y.

En el enfoque de filtrado colaborativo manejamos una matriz de clasificación de usuarios e ítems, consideramos la matriz como fuente principal para aplicar SVD y conseguir la factorización de matrices reduciendo a k dimensiones aplicando el teorema de Eckar – Youn esta propiedad nos permite disminuir el tiempo de procesamiento del cálculo y del uso de memoria.

2.1.13 Filtrado Colaborativo matrices de factorización no negativa (NFM)

Las matrices de factorización no negativa consisten en aproximar una matriz cuyos valores son positivos mediante la multiplicación de dos submatrices de elementos no-negativos, se busca aproximar una matriz X, dado el carácter de aproximación se introduce una matriz E de error de las mismas dimensiones, de modo que pueda hablarse de igualdad.

El modelo NMF calcula factores los cuales son una forma simplificada de los usuarios e ítem, luego de obtener estos valores minimizando los errores extrapolaremos los factores para calcular los rankings no conocidos.

Elaborar sistemas con factorización de matrices tiene varios beneficios simplificando el problema y logrando que el ruido que tiene los rankings no conocidos no despinte el perfil del usuario o el ítem.

2.1.14 Filtrado colaborativo utilizando k-vecinos más cercanos (KNN)

Estos métodos de filtrado se basan en la descripción de un elemento y un perfil de las opciones preferidas del usuario. Para un sistema basado en contenido se utiliza los perfiles de usuario para indicar los gustos de cada uno de estos describiendo los elementos. En otras palabras, los algoritmos intentan recomendar productos que sean similares a los que le han gustado a un usuario en el pasado.

Este algoritmo que se basa en los k vecinos más cercanos (KNN) no hace ninguna suposición sobre la distribución de datos subyacente, es un algoritmo de aprendizaje automático para encontrar grupos de usuarios similares. Cuando un KNN hace una predicción sobre un artículo, calculará la "distancia" entre el artículo de destino y todos los demás artículos de su base de datos.

2.1.15 Modelo Lightfm

Utiliza algoritmos de perdida WARP (Weighted Approximate-Rank Pairwise) cuya finalidad es crear tuplas entre usuario e item con el fin de estimar preferencias desconocidas. Cuando se multiplican las representaciones producen calificaciones para cada elemento de un usuario siendo más probables que elemento con alta puntuación sean interesantes para dicho usuario, estimando incrustación por cada característica y estas se suman para llegar a representar a los usuarios e item.

Aplicaremos el test_interaction para evaluar el modelo con técnica de validación cruzada.

2.2 Metodología

El presente ítem plantea describir las metodologías en las que se basa nuestro estudio para la aplicación de los modelos estadísticos y de machine Learning que nos permitan calcular los pronósticos de las ventas por categorías para el caso de los modelos predictivos y un sistema de recomendaciones para clientes con perfiles previamente analizados.

Aplicaremos diferentes modelos ARIMA para las predicciones de las series de tiempo para la categoría “Video” y “Refrigeración” y para el sistema de recomendaciones utilizaremos filtrado colaborativo basado en KNN, con herramientas de visualización iterativa, eficaz y de fácil manejo para una correcta toma de decisiones en el negocio.

En relación con el análisis se utilizó una serie de datos histórico-mensuales, que contiene las transacciones de las ventas totales por categoría de producto desde enero 2016 a abril 2021, teniendo un total de 64 periodos para nuestro análisis. Los datos de las ventas son registrados diariamente mediante las transacciones ejecutadas y almacenadas en el proceso de venta de algún producto, dichas ventas fueron sumadas y agrupadas por mes y por categoría respectivamente.

Como complemento a estas variables históricas de ventas, se obtuvo información de las publicidades que se realizaron en determinados meses del año por diferentes motivos o eventos, por ejemplo, en el mundial las empresas se enfocaron en la venta de televisores o aparatos tecnológicos. Esto nos ayudara a poder discriminar en que fechas se venden determinadas categorías de productos por dichos acontecimientos, ayudando al modelo que se propone.

(Céspedes Urrutia, 2017) Adicional se agregaron variables externas a los modelos para tener percepción de los factores externos a la empresa que puedan estar influyendo en el comportamiento del consumidor, estos datos fueron obtenidos por el Instituto Nacional de Estadísticas y Censos de la encuesta nacional de empleo y subempleo la cual mide mediante indicadores mensuales la tasa de empleo, subempleo y desempleo a nivel nacional. El valor del PIB y del IPC fueron otros de los indicadores de los datos externos adjuntados a los modelos; ambos indicadores como los de empleo y desempleo fueron tomados en el mismo periodo de tiempo de la serie histórica de ventas de la empresa, por tanto, para los meses en los que por motivo de la pandemia mundial (2020) cesaron de producir estos indicadores y se decidió analizar tendencias, inferencias y promedios móviles para el cálculo de estos valores en particular.

2.3 Fundamentos del problema

En la actualidad las empresas buscan estar un paso delante de sus principales competidores, tratando de disminuir costos y aumentar los ingresos siendo un punto importante el desarrollo de estudios predictivos mediante técnicas estadística, manteniendo un conocimiento de sus clientes al determinar su comportamiento de compra.

Predecir la demanda para las organizaciones se ha convertido en un rol importante para la toma de decisiones a altos niveles jerárquicos, esto ayuda a las empresas a tener una mejor planificación de su presupuesto, liquidez e inversión con la finalidad de reducir sus costos incrementado sus ingresos. Si bien

es cierto la gran mayoría de las industrias se han adaptado o están en el proceso de adaptación de tener una cultura de datos para tomar decisiones en base al comportamiento de estos mismo; de aquí surgen las metodologías estadísticas implementadas por autores como:

(Carrillo, 2019) Propone utilizar series temporales para predecir las ventas de una cervecería ubicada en Galicia – España implementado modelos clásicos como lo son ARIMA y ET´S los cuales detalla que por sí solos estos modelos no son lo suficientemente robustos proponiendo en segunda instancia modelos avanzados como la regresión dinámica inclusive aplicar series jerárquicas.

Otro estudio realizado en la ciudad de Cuenca – Ecuador por (Paul Paucar, 2020) propuso pronosticar el consumo de energía eléctrica en la Universidad Politécnica Salesiana y hacer un comparativo con el consumo de energía en la Universidad Politécnica de Valencia mediante la aplicación de métodos estadísticos como la regresión lineal simple, múltiple y modelos ARIMA obteniendo resultados significativos para distintos sets de datos (agrupados por estaciones).

La era digital está en su apogeo en donde los negocios evolucionan constantemente para lograr captar y fidelizar a sus clientes, una de las técnicas implementadas son los sistemas de recomendaciones los cuales se encuentran en cualquier actividad diaria que realizamos.

La compañía (River, 2021) tuvo un desafío interesante al diseñar un motor de recomendación de libros para maestros de Scholastic, al explorar las bibliotecas obtuvieron miles de libros e información recopilada de casi 100 años sin previo análisis ofreciendo recomendaciones basadas en la experiencia de sus colaboradores o algún producto de marketing, las cuales fueron tomadas con un elemento para el desarrollo del sistema.

Para los autores (Matías Bavera, 2018) tener información de los clientes ya sea datos demográficos o sobre el comportamiento de compra es fundamental para

elaborar un sistema de recomendación, conocer al cliente en base a su comportamiento aumenta los aciertos de las ofertas realizadas por la empresa; sin una buena calidad de datos no se logra recomendaciones óptimas.

(Tulasi K. Paradarami, 2017) encontraron procesos que sobrecargan de información a los sistemas debido a la gran cantidad de datos de entrada reduciendo la calidad y eficiencia de las decisiones, muchas veces estas sobrecargas causen problemas omitiendo algo importante o cometiendo un error, los autores manifiestan que un sistema de recomendación está diseñado para enfrentar estos problemas de sobrecarga proporcionando sugerencias optimas de artículos a los clientes.

(Keunho Choi, 2011) subrayan la importancia de los sistemas de recomendaciones personalizadas en una era de sobre carga de información, permitiendo a los compradores encontrar lo buscando sin perder tiempo y a su vez el vendedor ofrecerá un producto con alta probabilidad de compra preexistiendo una estrategia de ganar/ganar con beneficios mutuos.

(Loepp B., 2014) enfocan su problemática a la baja interactividad y control que tienen los usuarios al utilizar los sistemas de recomendaciones comunes; enfatizando que los usuarios no comprenden el porqué de la sugerencia de ciertos elementos, desconfiando del sistema por tal motivo buscan mayor participación en las decisiones al interactuar directamente con el cliente se produce una sinergia que aumentaría la precisión y diversidad en las recomendaciones.

Toda esta problemática que enfrentamos se enfoca en entender las preferencias de los clientes para así formar recomendaciones relevantes como nos indica (Ghobar, 2017) “Recomendar, sugerir e influir en las opiniones de los usuarios forma parte de la nueva realidad que extrae valor de las informaciones de modo más potente a través de los recursos de Machine Learning”.

2.4 Soluciones de analítica y aprendizaje relacionadas al problema

Los sistemas de recomendación brindan comodidad al cliente evitando el proceso de búsqueda de algún artículo permitiéndole obtener a simple vista lo que necesita, tal es el caso de Netflix que recomiendan películas en base a calificaciones y de distintos usuarios mediante enfoque supervisado como clasificación y regresión y enfoque no supervisados como la disminución de dimensionalidad y agrupación utilizando 107 algoritmos para predecir una única salida, este sistema ha ahorrado a la empresa más de mil millones al año descritas en la página de (Pramodh, 2020).

Para (Keunho Choi, 2011) realizar un sistema de recomendación con Filtrado neuronal ofrece un mejor rendimiento entre usuario y elemento obteniendo mejores recomendaciones, pero (Jinkun Wang, 2016) indica que obtener una lista de recomendación diversificada es óptima para dar alternativas al usuario implementado algoritmo evolutivo multiobjetivo MOEA mediante técnica de Pareto.

2.5 Librerías y software a utilizar

En el presente trabajo se realizarán los modelos predictivos en Rstudio utilizando las librerías data.table, dplyr que son óptimas para el manejo, depuración y modelamiento de los datos, para gráficos de las series y demás utilizaremos las librerías ggplot2, gpmisc que nos permiten realizar graficas de diferente índole y por último la librería forecast para el estudio de series temporales.

Para el sistema de recomendación seguiremos lo planteado por (Matías Bavera, 2018) con el lenguaje de programación de Python con una laptop corei7 con memoria RAM de 16GB y procesador x64 de sistema operativo.

La librería Surprise la utilizaremos para plantear varios modelos de sistema de recomendación como es el caso de Knn (Vecino más cercano), SVD (Descomposición en valores singulares) y NFM (Matrices de factorización no

negativa) esta librería también nos ayudara para la comparación de estos modelos mediante medidas de desempeño y cross validation.

Otra librería que se utilizará para comparar modelos de recomendación basado en contenido es Lightfm para modelar estos sistemas de acuerdo con las características de las categorías, luego realizaremos un modelo hibrido utilizando la función de WARP obteniendo métrica para evaluación del modelo.

CAPÍTULO 3

3. DISEÑO E IMPLEMENTACIÓN

Para la elaboración e implementación de los diferentes modelos de series de tiempo analizaremos la data de manera descriptiva concretamente de las categorías de producto y escogeremos las dos con mayor volumen de venta. Por otra parte, en el sistema de recomendaciones utilizaremos las variables que describen a los clientes (edad, sexo, cantidad de compra, categoría de producto).

El gráfico a continuación muestra las ventas totales de todas las categorías por cada año de estudio:

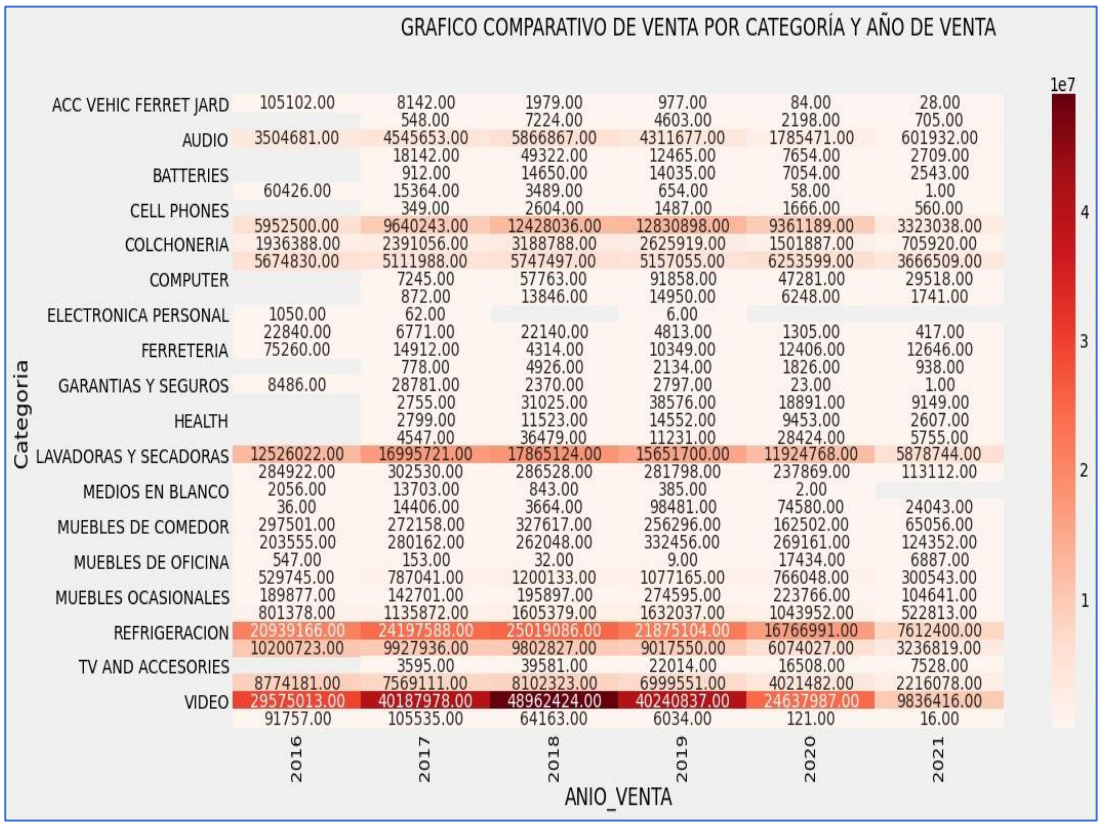

Ilustración 3-1 Ventas totales por Categoría y Año

En la ilustración 3-1 observamos en un heatmap el total de ventas de las categorías por cada año de estudio y verificamos que las categorías de video y refrigeración son las que mayor venta posen en cada año.

A continuación, en la ilustración 3-2 visualizamos que ambas líneas constituyen aproximadamente el 50% de las ventas totales de la empresa en cada año, por lo que serán las dos categorías de análisis para las series de tiempo y predicción de nuestro estudio.

Ilustración 3-2 Participación Ventas totales Categorías Refrigeración y Video por Año

3.1 Exploración y validación de datos y fuentes

En esta sección exploraremos las series de ventas de las categorías de producto seleccionadas, se analizará la estacionalidad y tendencia de las series; de la misma forma se analizará las series de variables económicas y de empleo.

3.1.1 Exploración de los datos de Venta

En la Ilustración 3-3 evidenciamos el pico más alto de la serie a finales del 2017 precisamente en el mes de noviembre en donde se dan las promociones de Black Friday en los meses de mayo también se presentan picos de venta por la época del día de la madre. Existe una tendencia creciente hasta comienzo del 2020, luego esto cambió debido a la emergencia sanitaria de COVID19 que atravesó el país lo cual conllevo a una caída significativa de las ventas en los meses de marzo a mayo 2020.

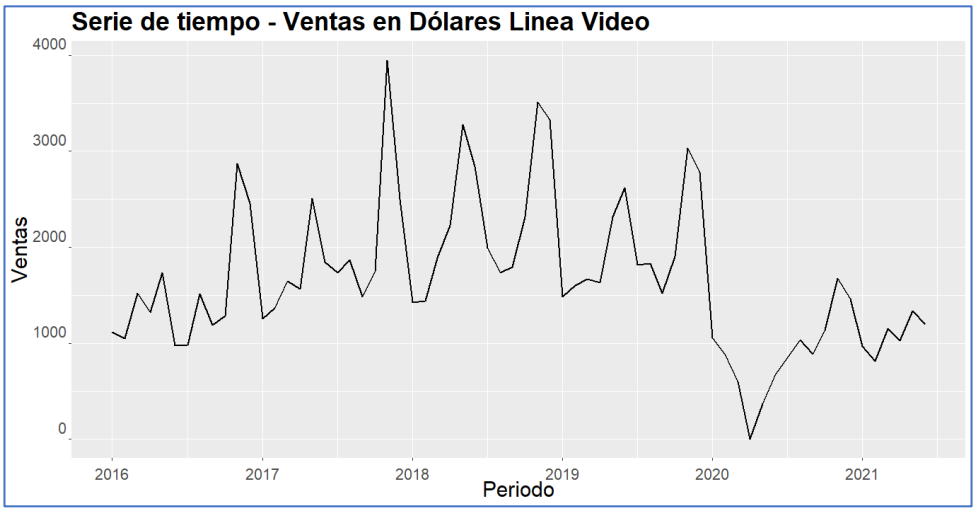

Ilustración 3-3 Serie Ventas Categoría Video

Se puede apreciar en la ilustración 3-4 que en los meses de noviembre y diciembre hay un considerado incremento de las ventas en comparación con los demás meses del año, esto debido al Black Friday, la época navideña y Fin de año. Para los meses de mayo y junio donde también se evidencia un incremento de las ventas de la categoría video tiene como consecuencia clara el día de la madre y del padre, así como también el mundial de futbol realizado en el año 2018.

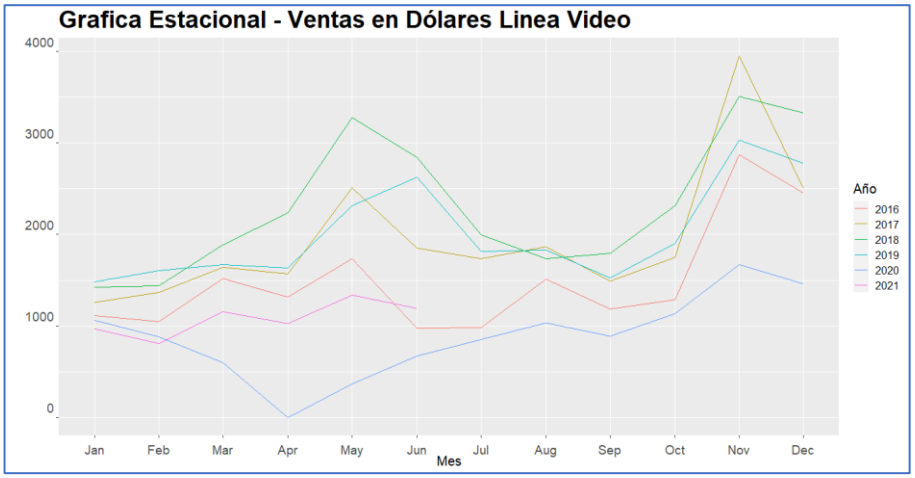

Ilustración 3-4 Serie Estacional de Ventas Categoría Video

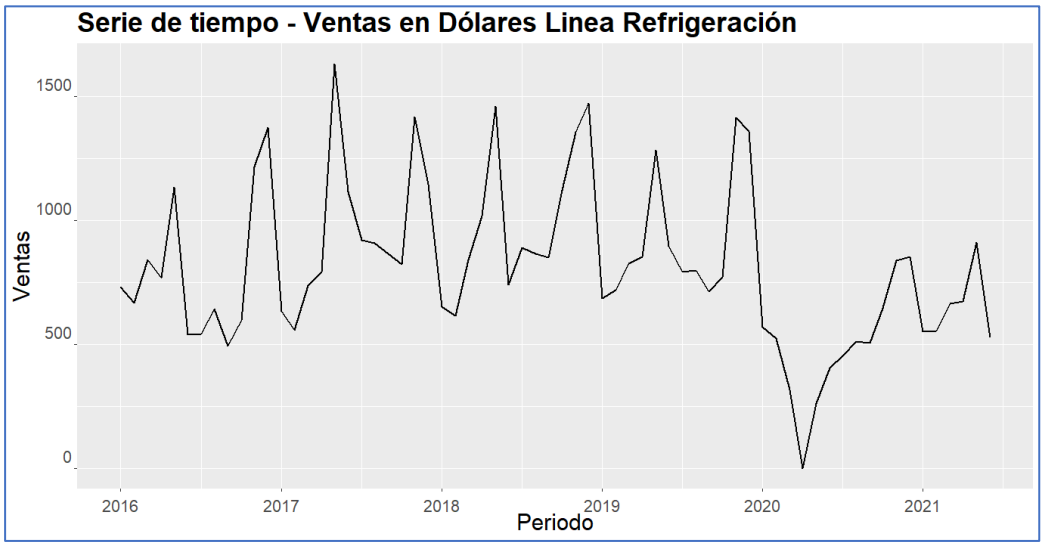

Ilustración 3-5 Serie Ventas Categoría Refrigeración

La serie de la ilustración 3-5 mantiene picos constantes en los meses de mayo y noviembre, evidentemente también se visualiza la caída de las ventas en el año 2020 por la pandemia de COVID19 aumentando la venta a partir de octubre en adelante.

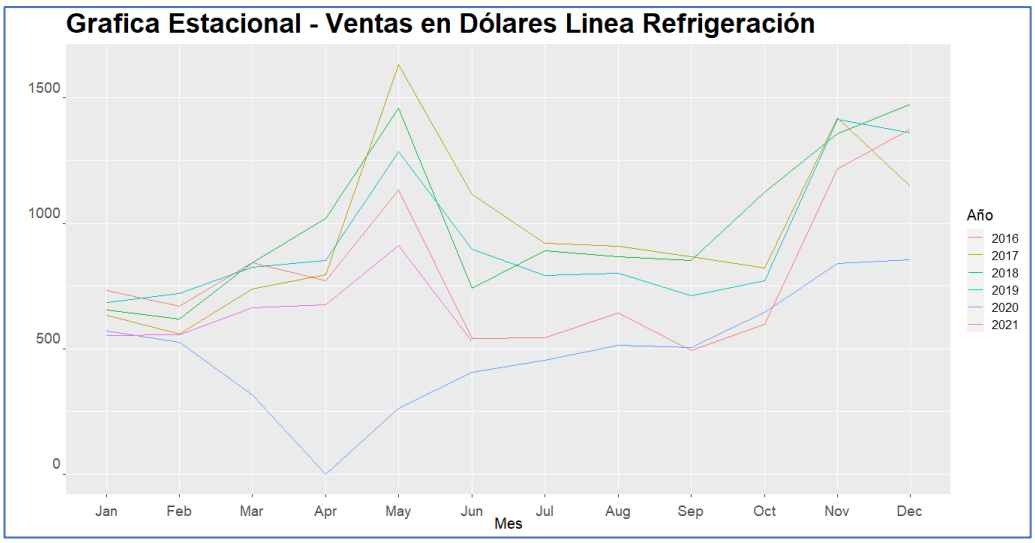

Ilustración 3-6 Serie Ventas Categoría Refrigeración

En la ilustración 3-6 se observa que al igual que la línea video, existen picos de venta en mayo, noviembre y diciembre y en los demás meses se muestra constante la venta.

3.1.2 Exploración de las variables económicas y de empleo

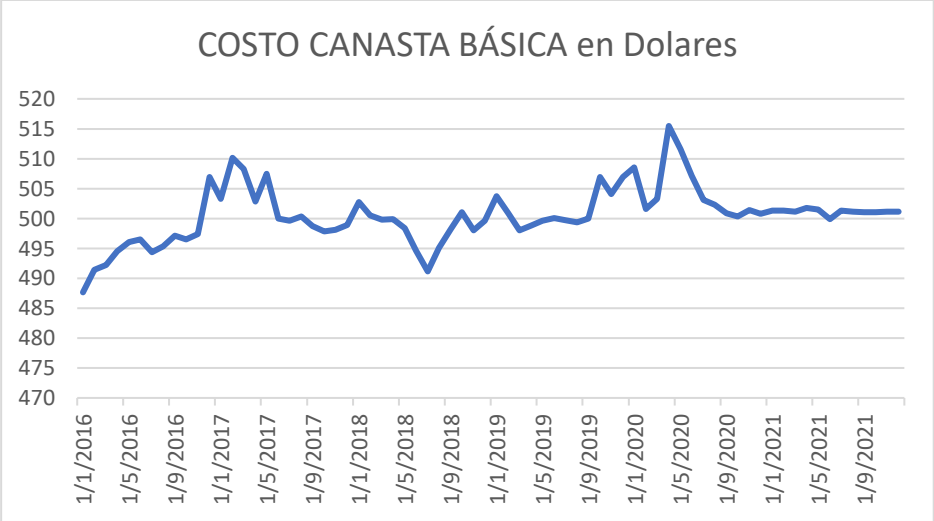

Ilustración 3-7 Serie Costo Canasta Básica

En la ilustración 3-7 de costo de la canasta básica se observa que incrementa en el intervalo de tiempo de estudio, no se observa alguna estacionalidad, pero existen picos al comienzo del 2017 y en enero a mayo del 2020. Hay que recordar que en marzo del 2020 comenzó las restricciones de movilización por la pandemia, incrementando los costos de los productos.

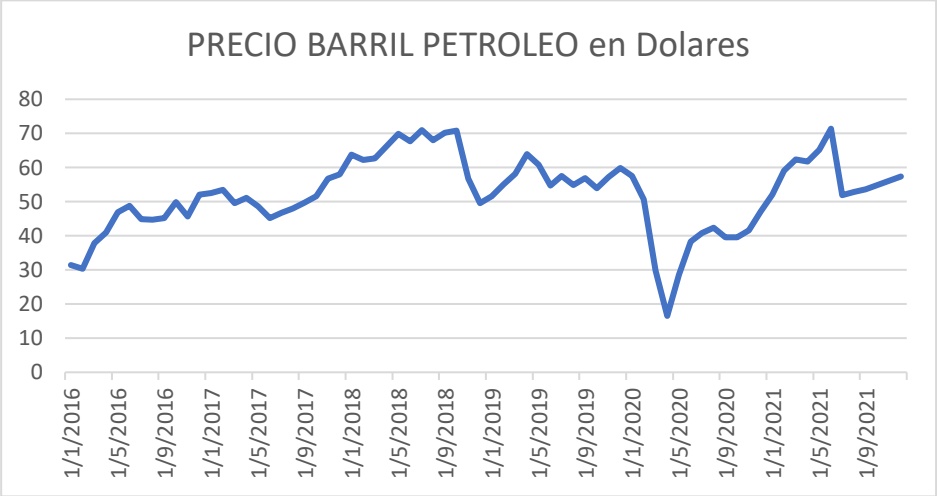

Ilustración 3-8 Serie precio del petróleo

Observamos la serie de los precios del barril del petróleo en la Ilustración 3-8 la cual presenta una tendencia creciente, pero para el año 2020 baja considerablemente por la pandemia regulándose a partir del 2021.

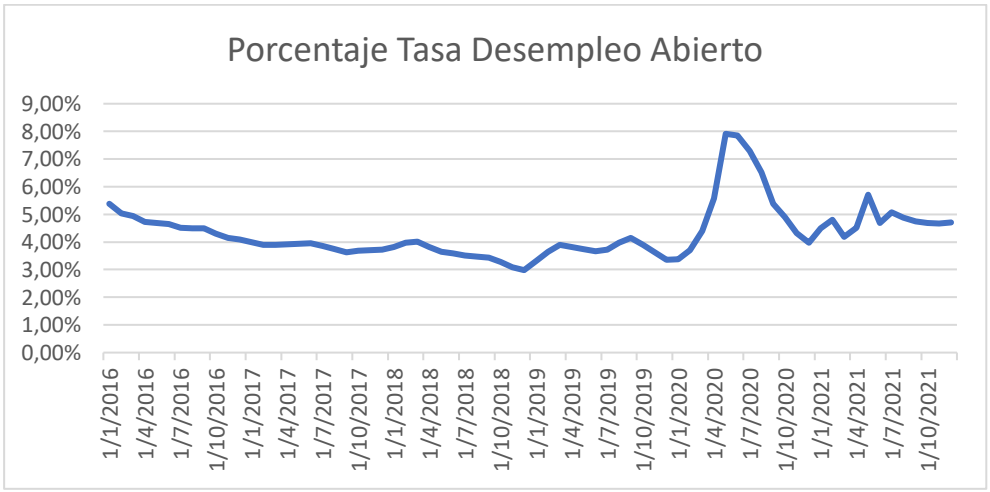

Ilustración 3-9 Serie Tasa de Desempleo Abierto

Para el caso de la tasa de Desempleo presentaba una tendencia decreciente hasta los primeros meses del año 2020, luego existió un incremento significativo en los periodos de marzo a mayo por el motivo de la pandemia reduciendo los valores hasta el presente año 2021 que muestra una tasa estabilizada, ilustración 3-9.

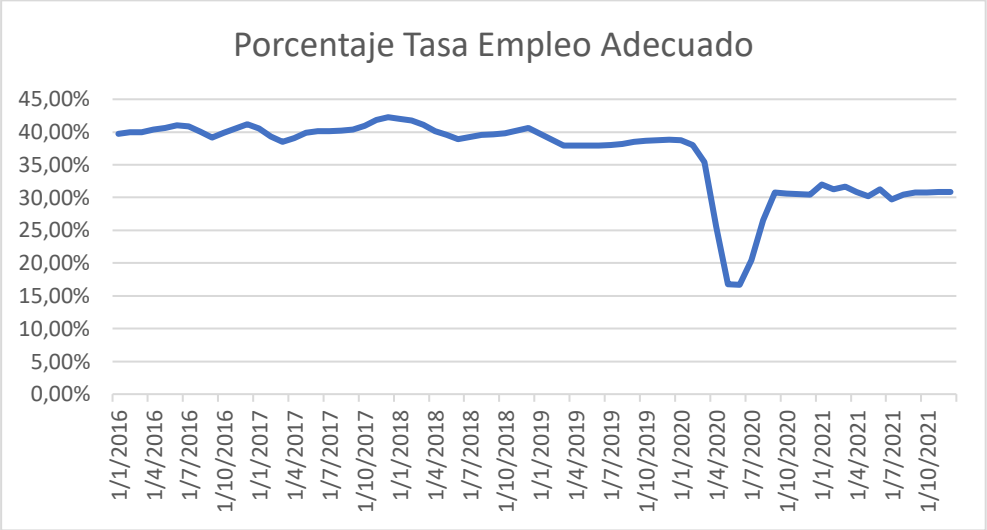

Ilustración 3-10 Serie Tasa Empleo Adecuado

En el caso de la Tasa de Empleo adecuado (ilustración 3-10) no presenta alguna tendencia, pero si un decrecimiento en los meses de marzo a mayo 2020 por la pandemia. Desde octubre 2020 la tasa empieza a oscilar en un valor de 30% muy por debajo comparado con años anteriores al 2020.

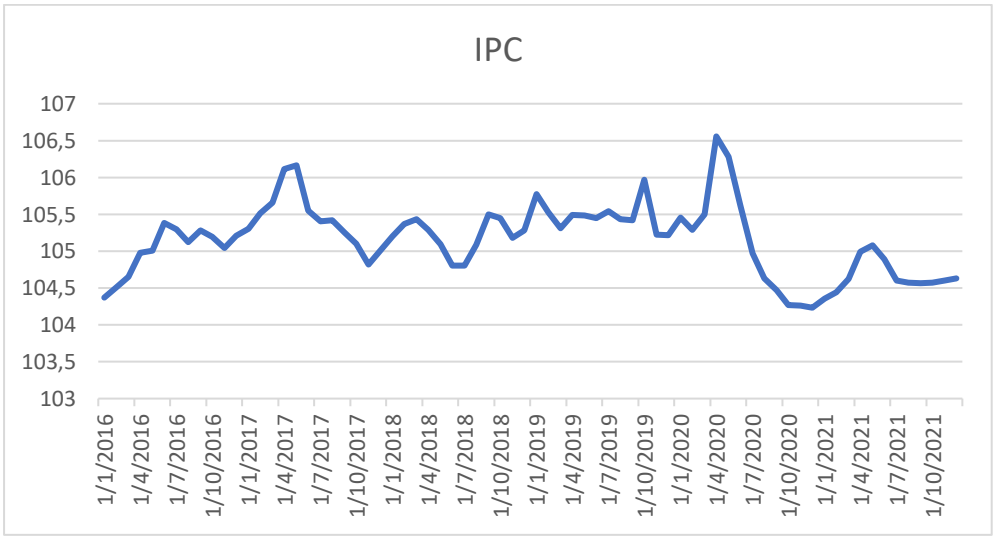

Ilustración 3-11 Serie Índice del precio al consumidor IPC

En la ilustración 3-11 se presenta la serie del IPC, no se observa una tendencia clara, pero si picos en los meses de abril.

3.1.3 Exploración de Variables para el Sistema de Recomendaciones

Para la aplicación del modelo de sistema de recomendaciones exploraremos de manera descriptivas las variables a utilizar:

1) Sexo del cliente
2) Sublinea comprada por el cliente
3) Edad del Cliente
4) Unidades Vendidas al cliente
5) Provincia
6) Id Cliente
7) Ingresos del cliente
8) Perfil del cliente

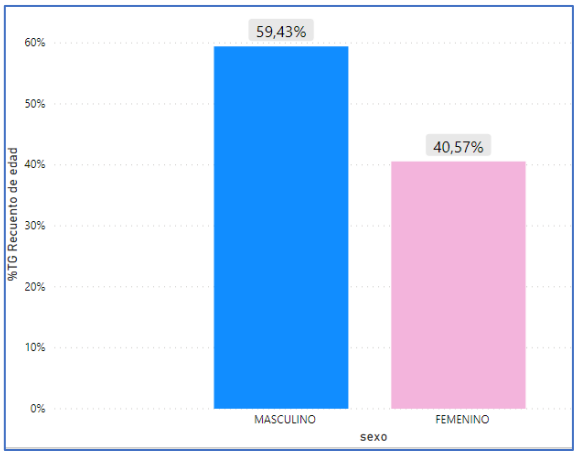

Ilustración 3-12 Porcentaje Género en la muestra

El set de datos contiene 554907 personas del género masculino que equivalen a un 59,43% del total de la data depurada y 378345 personas del género femenino equivalente a un 40,57% de la data (Ilustración 3-12).

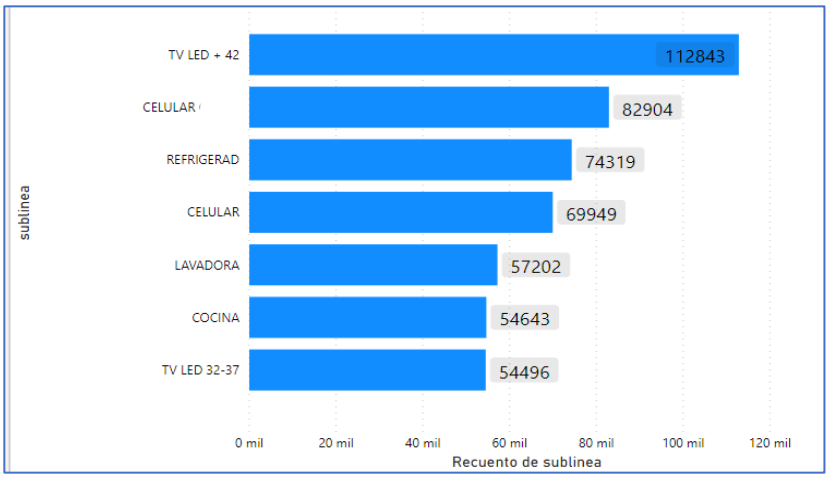

Ilustración 3-13 Recuento Top 5 Sublinea más vendida

En la ilustración 3-13 se observa que la sublinea mayormente demandada en el histórico de ventas para el periodo en estudio es el TV LED-42 con una participación del 12,09% seguido de la sublinea celular compañía “AAA” con una participación del 8,88%. Completando el top 5 con Refrigeradora 7,96%, Celular compañía “BBB” 7,49% y Lavadora con el 6,13%.

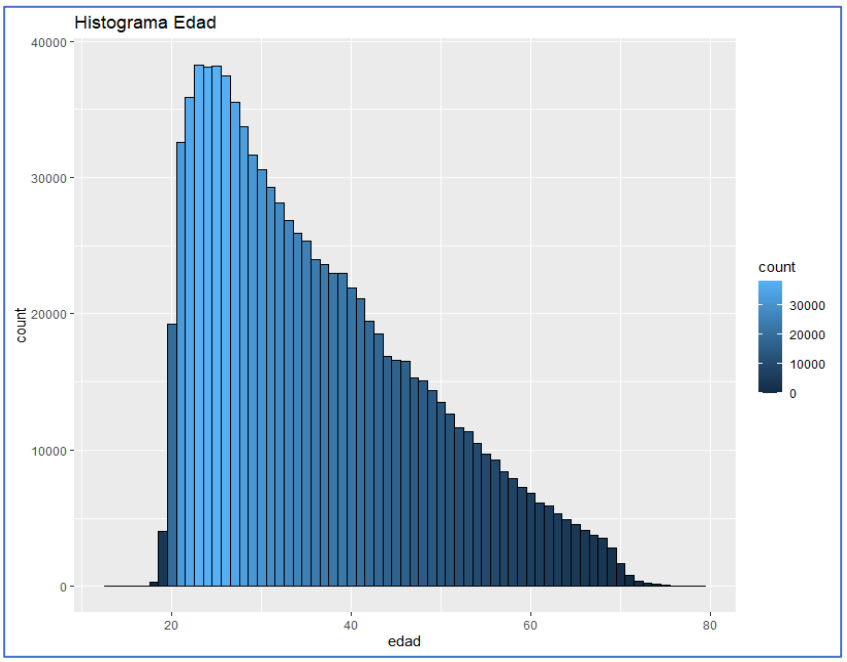

Ilustración 3-14 Histograma Edad Clientes

En la ilustración 3-14 se muestra la incidencia de las edades de los clientes y en que rangos oscilan en mayor proporción, por lo que a priori se visualiza que nuestra mayor clientela se encuentra en edades entre 20 y 40 años.

3.2 Prototipos de algoritmos, modelos, y módulos del sistema

En los siguientes apartados se mostrarán la construcción de los modelos de predicción para las dos categorías de producto y del sistema de recomendaciones. Para la creación de los modelos de predicción se utilizaron el 80% para entrenamiento (53 meses) y el 20% para pruebas (13 meses). Para la comprobación de la estacionariedad de las series las pruebas ADF y KPSS, las hipótesis son las siguientes:

Test ADF

H_o: $Esta\ serie\ tiene\ raizes\ unitarias, no\ es\ estacionaria\ y$

H_a: $Esta\ serie\ es\ estacionaria$

Test KPSS

H_o: $Esta\ serie\ es\ estacionaria\ y$

H_a: $Esta\ serie\ no\ es\ estacionaria$

Se analizó la normalidad de los residuos, para esto se utilizó el test de Shapiro wilk el cual compara las siguientes hipótesis:

Test SHAPIRO - WILKS

H_o: $La\ distribución\ es\ normal\ y$

H_a: $La\ distribución\ no\ es\ normal$

Al elaborar el sistema de recomendaciones contamos inicialmente con 415792 clientes únicos el cual al depurar aquellos que tuvieron una solo compra obtuvimos 243540, estos serán los clientes que analizaremos para el sistema combinando con las categorías de producto que han comprado. Para el

entrenamiento de los modelos se seleccionó el historial de compra (Factura) de los clientes exceptuando su última factura, estas facturas serán tomadas como data de prueba del modelo y así poder comparar las recomendaciones arrojadas con las compras realizadas de los clientes. Contamos con 512498 registros para el entrenamiento y 243540 de prueba.

3.2.1 Elaboración de Modelos Línea Video

3.2.1.1 Modelo Arima Video

Se Muestra la Serie con los datos de Entrenamiento.

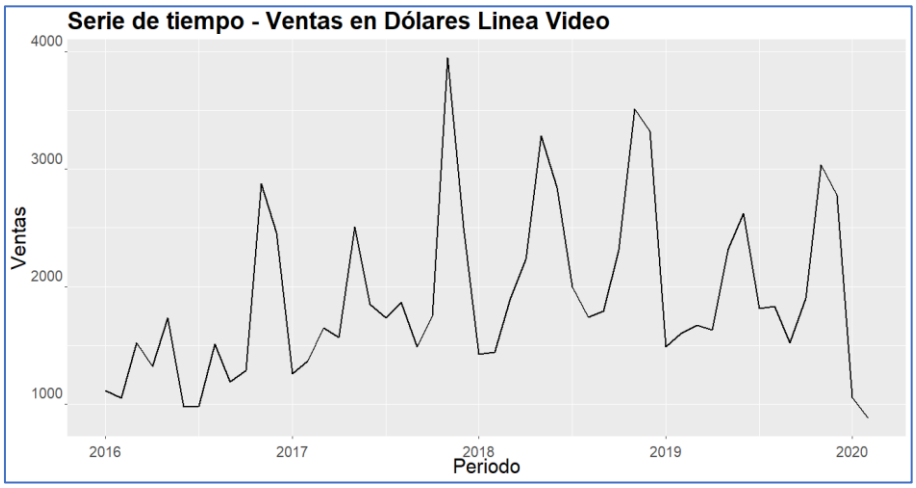

Ilustración 3-15 Serie de Tiempo Ventas VIDEO

En la figura 3-15 se logra visualizar que la serie no presenta estacionalidad, para esto se aplica a la serie la transformada de BoxCox dando un valor de lambda = 0.4177686, se presenta la serie resultante con la cual realizaremos el entrenamiento del modelo.

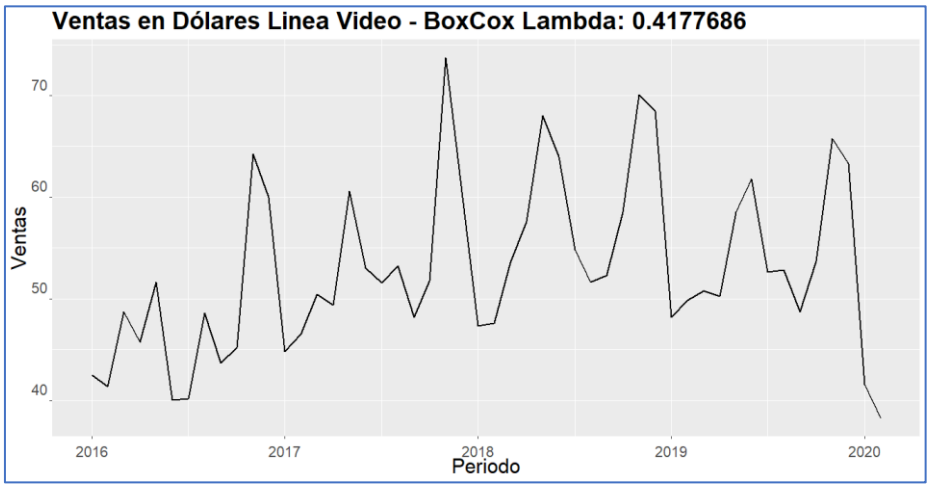

Ilustración 3-16 Serie Ventas VIDEO Transformada

Observamos (Ilustración 3-16) que luego de aplicar la transformada no se logra estabilizar la serie, se aplicara las funciones ndiff() y sndiff que nos indicarán el número de diferenciaciones regulares y estacionales para que la serie sea estacionaria.

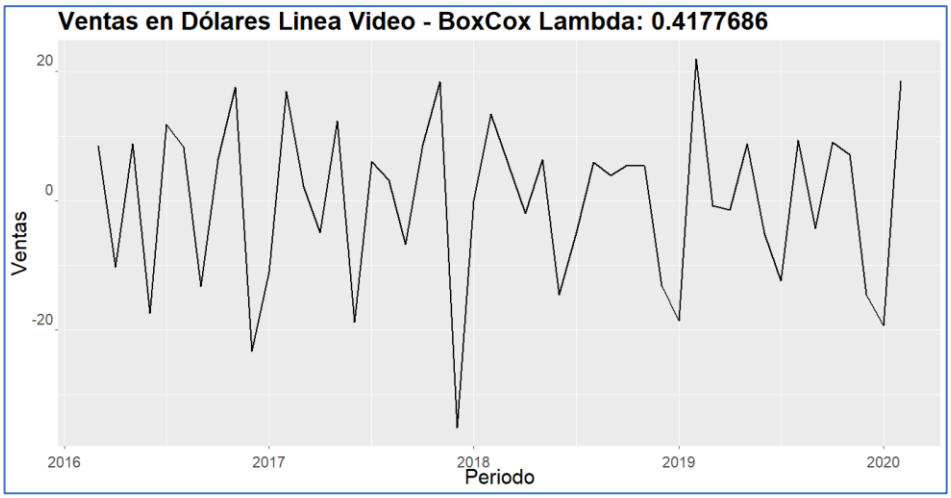

Ilustración 3-17 Serie Ventas Video BoxCox

Al aplicar la transformada de BoxCox, con una diferencia regular y una estacional se obtiene una serie estacionaria con varianza y media constante. Para comprobar la estacionariedad de la serie

se efectúa el test ADF mostrando un valor p de 0.01 rechazando la H_o la cual nos indicaba que la serie no era estacionaria y el test de KPSS dándonos un valor p de 0.1 aceptando la H_o, que nos indica que la serie es estacionaria, con la ayuda de estos test podemos afirmar que nuestra serie es estacionaria.

Se procede a determinar los valores de “p” componente del AR y “q” componente del MA a través de los gráficos de autocorrelaciones.

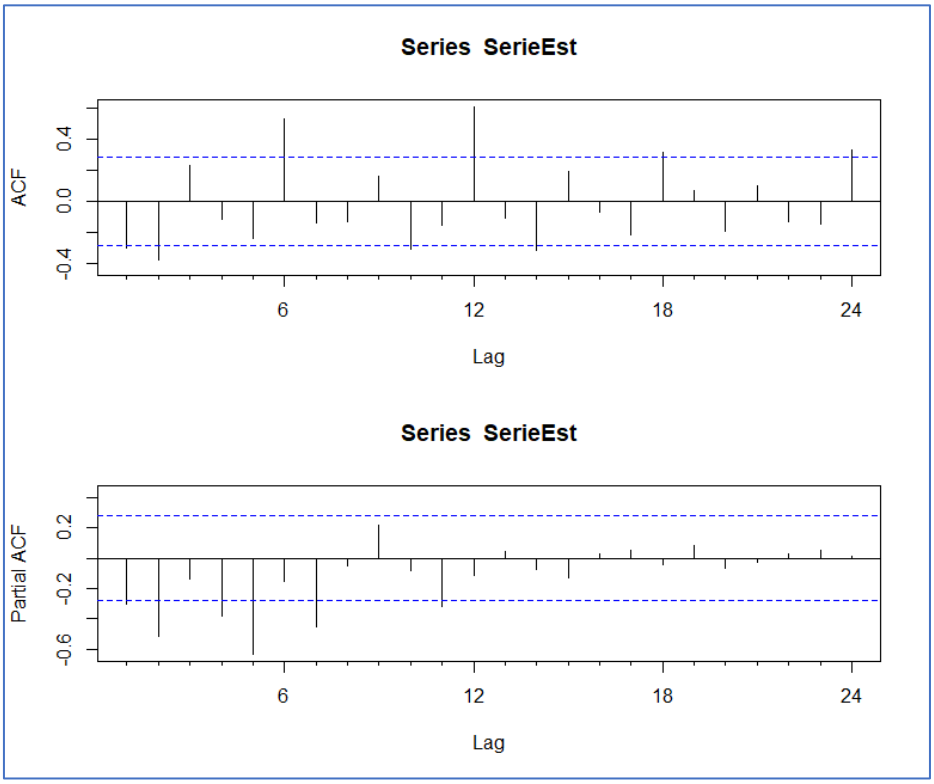

Ilustración 3-18 Autocorrelaciones Serie Video

Analizando los correlogramas se puede determinar de manera visual los componentes p y q. De acuerdo con el PACF (Función de autocorrelaciones parciales) sugiere un AR (0) o AR (1) mientras que en la gráfica de ACF muestra rezagos que indican un posible MA (1) o MA (2) y MA (1) estacional.

Se realizaron varios modelos con las combinaciones de cada uno de los parámetros, para seleccionar el mejor se tomó el criterio del AIC:

Tabla 3-1 Selección Mejor Modelo Criterio AIC Serie Video

MODELO	AIC
ARIMA (0,1,1) (0,1,1)	221.1149
ARIMA (0,1,2) (0,1,1)	222.5791
ARIMA (1,1,1) (0,1,1)	222.3794
ARIMA (0,1,1) (0,1,0)	219.5409
ARIMA (0,1,1) (1,1,1)	223.5499
ARIMA (0,1,2) (1,1,1)	225.1367
ARIMA (1,1,1) (1,1,1)	224.9284
ARIMA (2,1,2) (1,1,1)	230.4789
ARIMA (0,1,2) (0,1,0)	220.8081
ARIMA (1,1,1) (0,1,0)	220.5733

En la tabla se muestran los modelos propuestos en base a las combinaciones que consideramos de los parámetros AR y MA, en donde se escogió los dos modelos que tienen menor AIC para analizarlas en el apartado de Modelo con Errores Arima. A continuación, mostraremos las medidas de los modelos y comprobaremos los supuestos de los residuos.

Para el modelo ARIMA (0,1,1) (0,1,0) obtuvimos un coeficiente para la media móvil de (q) de -0.5237 mostrada en la Tabla 3-2, mientras que en la tabla 3-3 encontramos las medias de error del modelo

Tabla 3-2 Coeficiente Media Móvil Arima (0,1,1)(0,1,0) Video

Coeficiente:	Valor
ma1	-0.5237
s.e.	0.2145

Tabla 3-3 Medías Error Arima (0,1,1)(0,1,0) Video

Medidas de Error – Datos de entrenamiento:					
	ME	**RMSE**	**MAE**	**MPE**	**MAPE**
Datos entrenamiento	-39.52	347.41	243.33	-3.29	11.96

Procedemos a comprobar los supuestos de normalidad y de ruido blanco de los residuos:

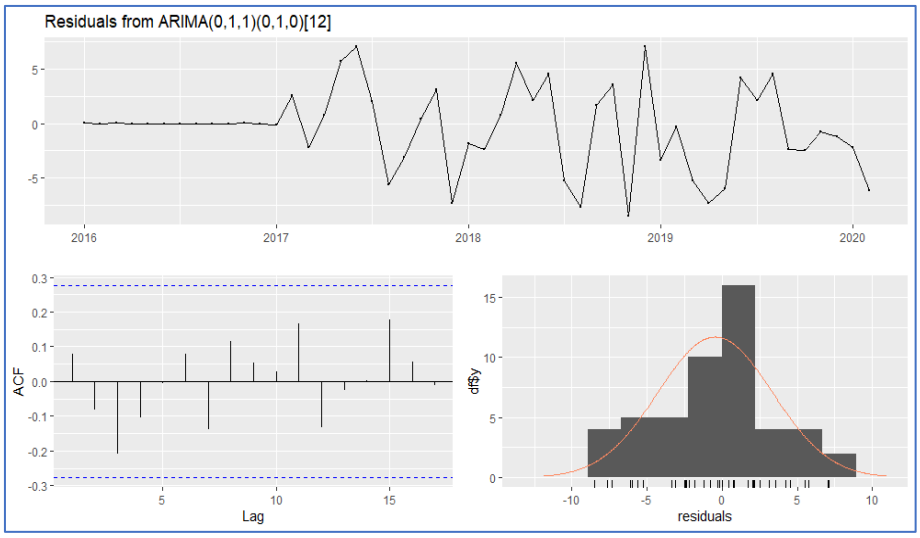

Ilustración 3-19 Supuestos Normalidad y Ruido Blanco Residuos Serie Video Arima (0,1,1) (0,1,0)

En la ilustración 3-20 podemos observar que los residuos poseen una conducta de ruido blanco, lo comprobamos con el test de Ljung-Box al obtener un valor p de 0.7058 indicándonos que los errores cumplen con el supuesto de independencia, en el gráfico de autocorrelación no presentan ningún rezago. Para comprobar el supuesto de normalidad se utilizó el test de Shapiro-Wilk cuyo valor p es de 0.144 comprobando este supuesto, en la gráfica 1-16 se puede distinguir que en el histograma se asemeja a una campana como la distribución normal esto es una forma gráfica de comprobar el supuesto.

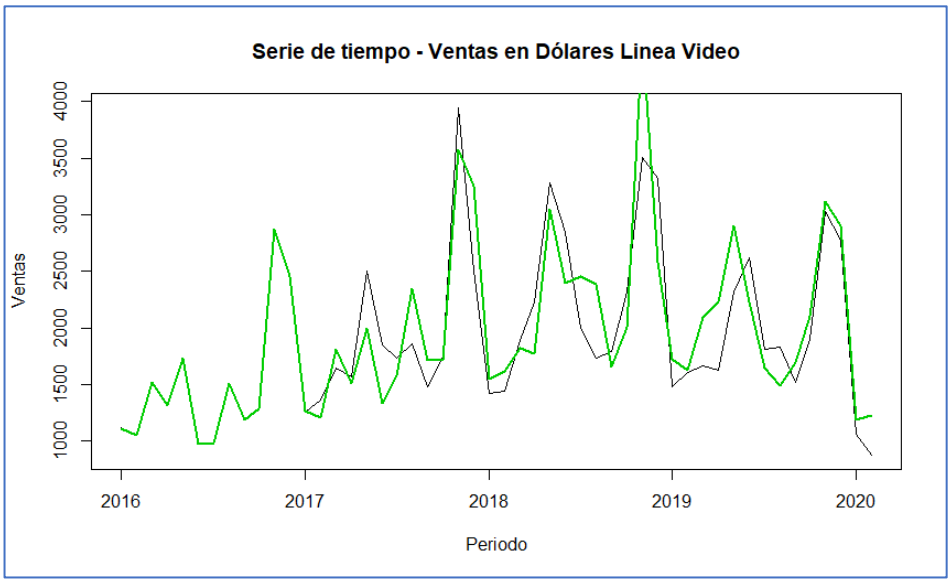

Ilustración 3-20 Comparación Serie Original vs Serie Ajustada Video Arima (0,1,1) (0,1,0)

En la imagen 3-20 se compara la serie normal y la serie ajustada por el modelo (color verde), se puede apreciar que se asemeja a la mayoría de los puntos.

Para el modelo ARIMA (1,1,1) (0,1,0) obtuvimos un coeficiente para la parte autorregresiva (p) de 0.2951 y de media móvil de (q) de -0.7451 mostrada en la Tabla 3-4, mientras que en la Tabla 3-5 encontramos las medias de error del modelo

Tabla 3-4 Coeficientes ARIMA (1,1,1) (0,1,0) Video

Coeficiente	ar1	ma1
Valor	0.2951	-0.7451
s.e.	0.2321	0.1515

Tabla 3-5 Medias Error del Modelo ARIMA (1,1,1) (0,1,0) Video

Medidas de Error – Datos de entrenamiento:					
	ME	RMSE	MAE	MPE	MAPE
Datos entrenamiento	-47.52	345.77	234.38	-3.60	11.52

Procedemos a comprobar los supuestos de normalidad y de ruido blanco de los residuos:

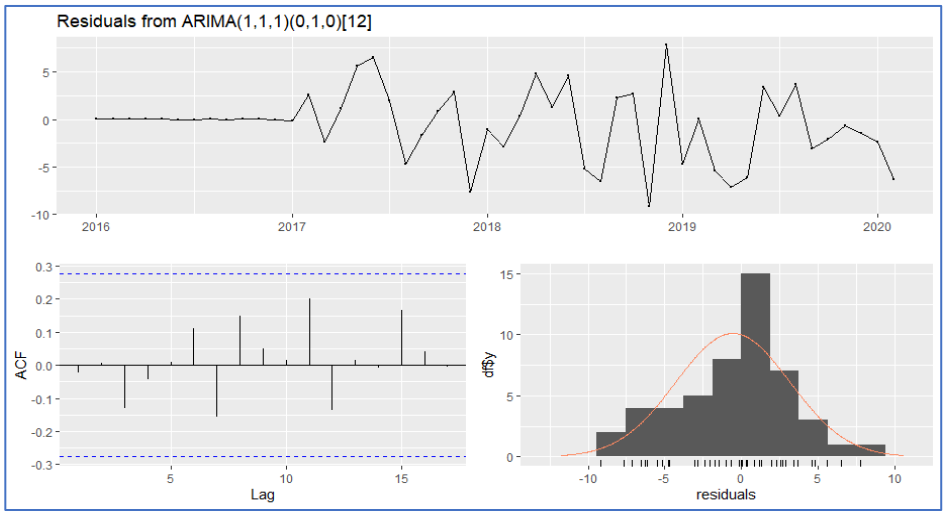

Ilustración 3-21 Supuestos Normalidad y Ruido Blanco ARIMA (1,1,1) (0,1,0) Video

En la ilustración 3-21 podemos observar que los residuos poseen una conducta de ruido blanco, lo comprobamos con el test de Ljung-Box al obtener un valor p de 0.7716 indicándonos que los errores cumplen con el supuesto de independencia, en el gráfico de autocorrelación no presentan ningún rezago. Para comprobar el supuesto de normalidad se utilizó el test de Shapiro-Wilk cuyo valor p es de 0.1771 comprobando este supuesto, de igual manera lo comprobamos con la gráfica del histograma.

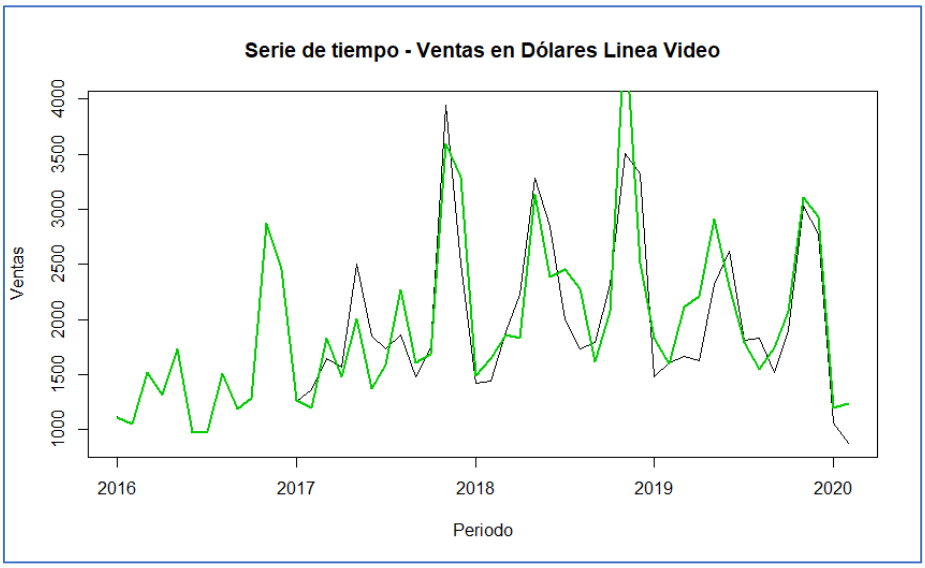

Ilustración 3-22 Comparación Serie Original vs Serie Ajustada Video Arima (1,1,1) (0,1,0)

En la imagen 3-22 se compara la serie normal y la serie ajustada por el modelo (color verde), se puede apreciar que se asemeja a la mayoría de los puntos, este modelo tiene un menor RMSE y MAPE.

3.2.1.2 Modelo de Regresión Múltiple Video

Una vez analizado los modelos ARIMA, se procede a la creación de modelo de Regresión múltiple con la variable de interés que es la venta total y las variables explicativas que se componen en variables económicas y variables de Empleo detalladas en la Tabla 1-2.

Se elabora una matriz de correlación entre las variables para determinar el nivel de relación que tienen entre sí; y si es negativa o positiva esta correlación.

Tabla 3-6 Correlación Variables Externas vs Video

	Venta	Costo Canasta	Precio Barril	Desempleo Abierto	Empleo Adecuado	Sub Empleo	IPC	VarIPC
Venta	1	0.11	0.37	-0.52	0.10	0.23	-0.07	-0.42
Costo Canasta	0.11	1	0.24	-0.48	-0.16	0.17	0.72	0.09
Precio Barril	0.37	0.24	1	-0.71	-0.14	0.20	0.21	-0.09
Desempleo Abierto	-0.52	-0.48	-0.71	1	0.16	-0.25	-0.40	0.19
Empleo Adecuado	0.10	-0.16	-0.14	0.16	1	0.37	-0.30	0.07
Sub Empleo	0.23	0.17	0.20	-0.25	0.37	1	0.07	-0.14
IPC	-0.07	0.72	0.21	-0.40	-0.30	0.07	1	0.20
VarIPC	-0.42	0.09	-0.09	0.19	0.07	-0	0.20	1.00

Analizando la matriz de correlación observamos que el costo de canasta básica, el precio de barril de petróleo, empleo adecuado y el subempleo se correlacionan positivamente con las ventas siendo el precio de barril quien tiene un valor más alto 0.37. Las variables de desempleo abierto, el IPC y la variación del IPC se correlacionan negativamente con las Ventas, es decir mientras aumentan de valor disminuye las ventas el que más se correlación es el desempleo abierto con un valor de -0.52.

Analizando las correlaciones entre las variables explicativas, no observamos una correlación cercana a uno siendo las más alta entre las variables de precio de barril y desempleo abierto con una correlación negativa de -0.71.

Para seleccionar el mejor modelo utilizaremos el método de Forward, el cual consiste en comenzar con un modelo “vacío” solo con la variable de interés e ir seleccionando uno por uno las variables explicativas de acuerdo con cuál de estas disminuye el AIC repitiendo el proceso hasta que ninguna de las variables logre disminuir el indicador.

En la tabla 3-7 mostramos los modelos con cada una de las variables que fueron seleccionadas con su respectivo AIC:

Tabla 3-7 Modelos Regresión Múltiple Video

Modelos	AIC	R^2
Venta = 1	658.25	0
Venta = DesempleoAbierto	644.68	0.252
Venta = DesempleoAbierto + VarIPC	638.59	0.351
Venta = DesempleoAbierto + VarIPC + IPC	636.78	0.385

Observamos que la primera variable elegida fue Desempleo abierto, luego fueron VarIPC e IPC obteniendo tres variables explicativas para el modelo las otras variables aumentan el AIC como se muestra en la tabla 3-8, en donde la mejor opción fue no elegir más variables ya que aumentaba el AIC.

Tabla 3-8 Modelo Regresión Múltiple Seleccionado Video

Venta = DesempleoAbierto + VarIPC + IPC		
	Sum of Sq	**AIC**
+ No Hacer nada		636.78
+ Empleo Adecuado	494083	637.05
+ Sub Empleo	155693	638.24
+ Costo Canasta	93108	638.46
+ Precio Barril	180	638.78

Las variables del modelo seleccionado son todas significativas como se muestra en la tabla 3-9 en donde encontramos los coeficientes del modelo.

Tabla 3-9 Coeficientes Modelo Seleccionado Video

	Parametro	**Std. Error**	**t value**	**Pr(>|t|)**
Intercepto	58234.2	28171.5	2.07	0.044
DesempleoAbierto	-81952.7	18606	-4.41	6.28E-05
VarIPC	-788.4	351.9	-2.24	0.03
IPC	-504.3	264.4	-1.91	0.063
Modelo: Venta = 58234.2 -81952.7 DesempleoAbierto -788.4 VarIPC -504.3 IPC				

Analizando los coeficientes del modelo, observamos que por cada punto porcentual que incrementa el DesempleoAbierto las ventas disminuirían en promedio $819.53, por cada aumento en el IPC disminuye en $504.3 la venta, mientras que por cada punto porcentual que incrementa la variación del IPC disminuye la venta en $78.84.

A continuación, mostramos las medidas de error del modelo:

Tabla 3-10 Mediadas de error del modelo seleccionado Video

Medidas de Error – Datos de entrenamiento:					
	ME	**RMSE**	**MAE**	**MPE**	**MAPE**
Datos entrenamiento	-6.81E-15	537.99	427.57	-7.80	24.64

En la ilustración 3-23 se compara la serie normal y la serie ajustada por el modelo de color rojo.

Ilustración 3-23 Serie Original vs Serie Ajustada Modelo Regresión Múltiple Video

3.2.1.3 Modelo con Errores Arima Video

Para esta sección utilizaremos los modelos de ARIMA propuestos y de Regresión múltiple manteniendo dos términos de error al combinar ambos métodos de predicción.

Primero realizaremos el modelo con errores con el modelo ARIMA (0,1,1) (0,1,0) con las variables del modelo de regresión múltiple, obteniendo los siguientes coeficientes:

Tabla 3-11 Coeficiente Modelo con Errores Arima (0,1,1)(0,1,0) Video

Coeficientes	ma1	DesempleoAbierto	VarIPC	IPC
Valor	-0.5674	-88.7548	-4.2863	-0.0545
s.e.	0.1696	282.8076	2.2763	2.1578

A continuación, se muestra las medidas de error del modelo:

Tabla 3-12 Medidas de Error Modelo con Errores Arima (0,1,1) (0,1,0) Video

Medidas de Error – Datos de entrenamiento:					
	ME	RMSE	MAE	MPE	MAPE
Datos entrenamiento	-3.69E+01	335.39	237.29	-2.92	11.56

En la ilustración 3-24 tenemos la curva ajustada del modelo de color morado.

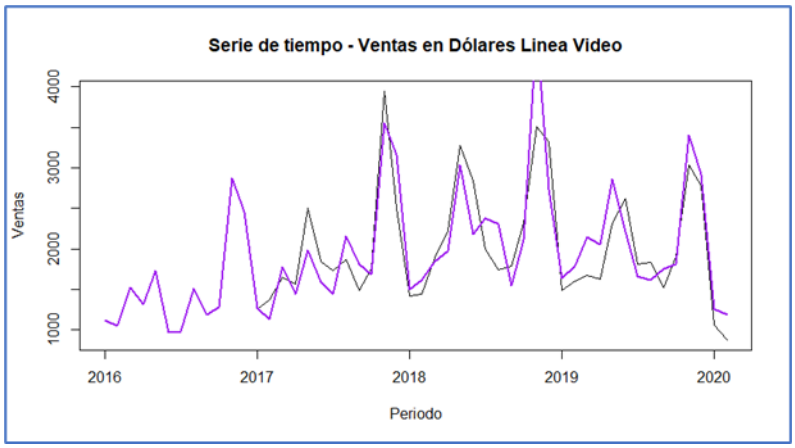

Ilustración 3-24 Serie Original vs Serie Ajustada Modelo con Errores Arima (0,1,1) (0,1,0) Video

Observamos que la línea ajustada difiere en ciertos puntos de la serie original, pero se aprecia buenos pronósticos. Se procederá a analizar el siguiente modelo ARIMA (1,1,1) (0,1,0) con las variables de regresión obteniendo los siguientes coeficientes:

Tabla 3-13 Coeficientes Modelo con Errores Arima (1,1,1) (0,1,0) Video

Coeficientes	ar1	ma1	DesempleoAbierto	VarIPC	IPC
valor	0.1573	-0.6796	-109.8112	-3.748	-0.2333
s.e.	0.3051	0.235	298.83	2.5301	2.2338

Se observa las medidas de error del modelo en la siguiente tabla:

Tabla 3-14 Medidas de error del Modelo con Errores Arima (1,1,1) (0,1,0) Video

Medidas de Error – Datos de entrenamiento:					
	ME	RMSE	MAE	MPE	MAPE
Datos entrenamiento	-3.96E+01	334.61	234.75	-3.03	11.48

En la ilustración 3-25 tenemos la curva ajusta del modelo de color naranja.

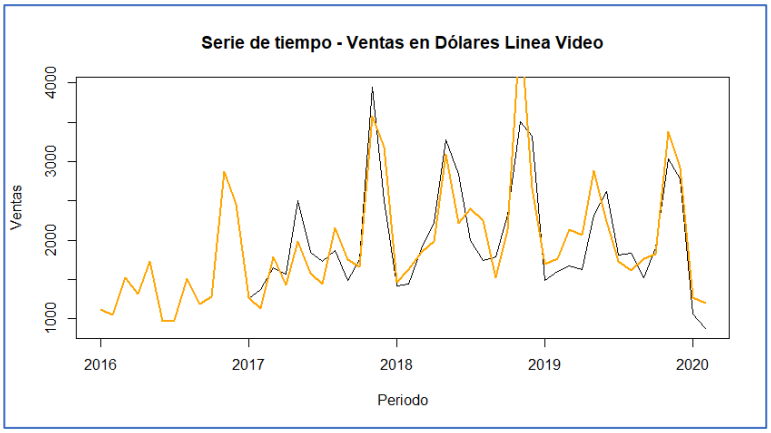

Ilustración 3-25 Serie Original vs Serie Ajustada Modelo con Errores Arima (1,1,1) (0,1,0) Video

Observamos que la línea ajustada difiere en ciertos puntos de la serie original. En ambos modelos seleccionados se presentan medidas de error muy similares, en la sección 3.2.4 del presente capítulo seleccionaremos el mejor modelo con la data de prueba.

3.2.2 Elaboración de Modelos Línea Refrigeración

3.2.2.1 Modelo Arima Refrigeración

Se Muestra la Serie con los datos de Entrenamiento.

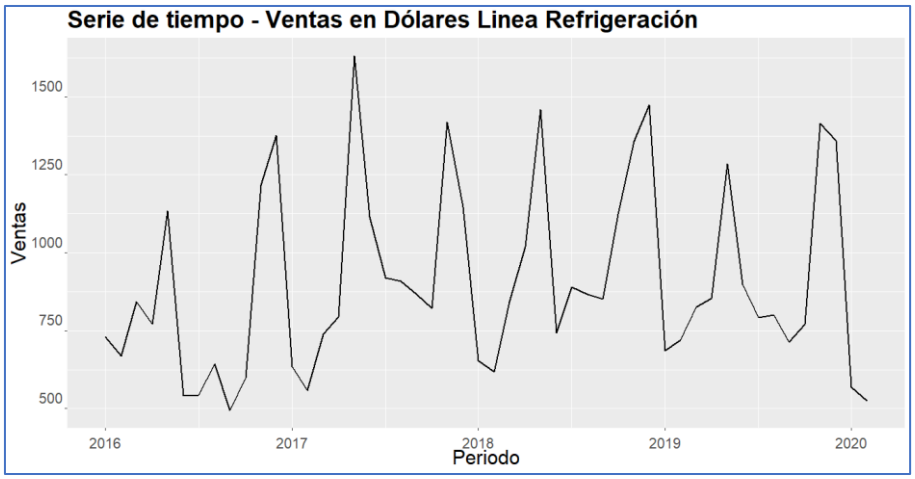

Ilustración 3-26 Serie de Tiempo Ventas REFRIGERACIÓN

En la Ilustración 3-26 se puede observar que la serie no presenta estacionalidad, para esto se aplica a la serie la transformada de BoxCox dando un valor de lambda = 1.080651, se presenta la serie resultante con la cual realizaremos el entrenamiento del modelo.

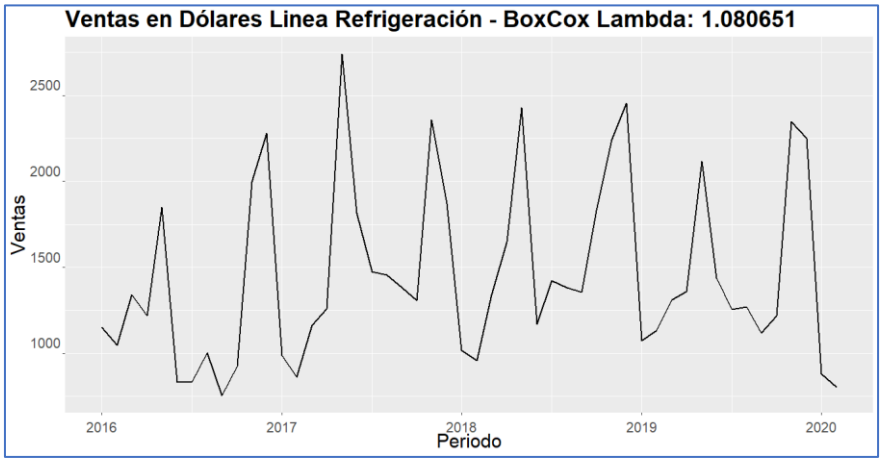

Ilustración 3-27 Serie de Tiempo Ventas Refrigeración Transformada

Observamos en la ilustración 3-27 que luego de aplicar la transformada no se logra estabilizar la serie, se aplicara las funciones ndiff() y sndiff que nos indicaran el número de diferenciaciones regulares y estacionales para que la serie sea estacionaria.

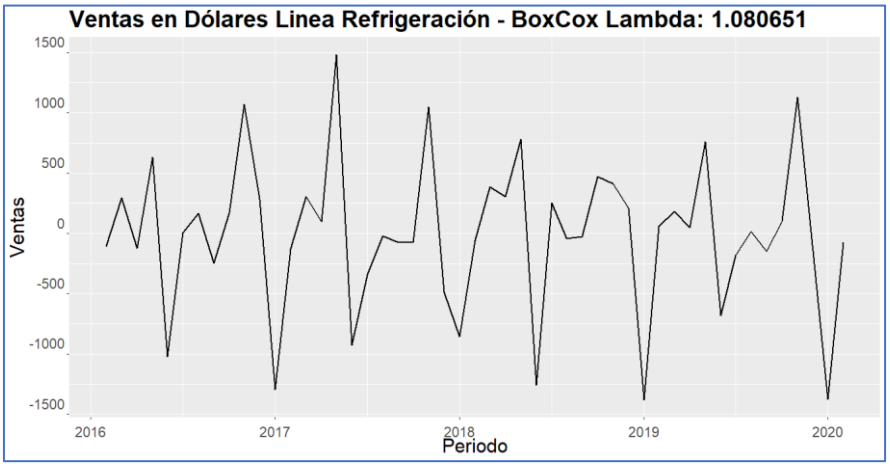

Ilustración 3-28 Serie Ventas Refrigeración Transformada Diferenciada

Al aplicar la transformada de BoxCox, con una diferencia regular y estacional se obtiene una serie estacionaria con varianza y media constante. Para comprobar la estacionariedad de la serie se efectúa el test ADF mostrando un valor p de 0.01 rechazando la H_o la cual nos indicaba que la serie no era estacionaria y el test de KPSS dándonos un valor p de 0.1 aceptando la H_o, que nos indica que la serie es estacionaria, con la ayuda de estos test podemos afirmar que nuestra serie es estacionaria.

Se procede a determinar los valores de “p” componente del AR y “q” componente del MA a través de los gráficos de autocorrelaciones (Ilustración 3-29).

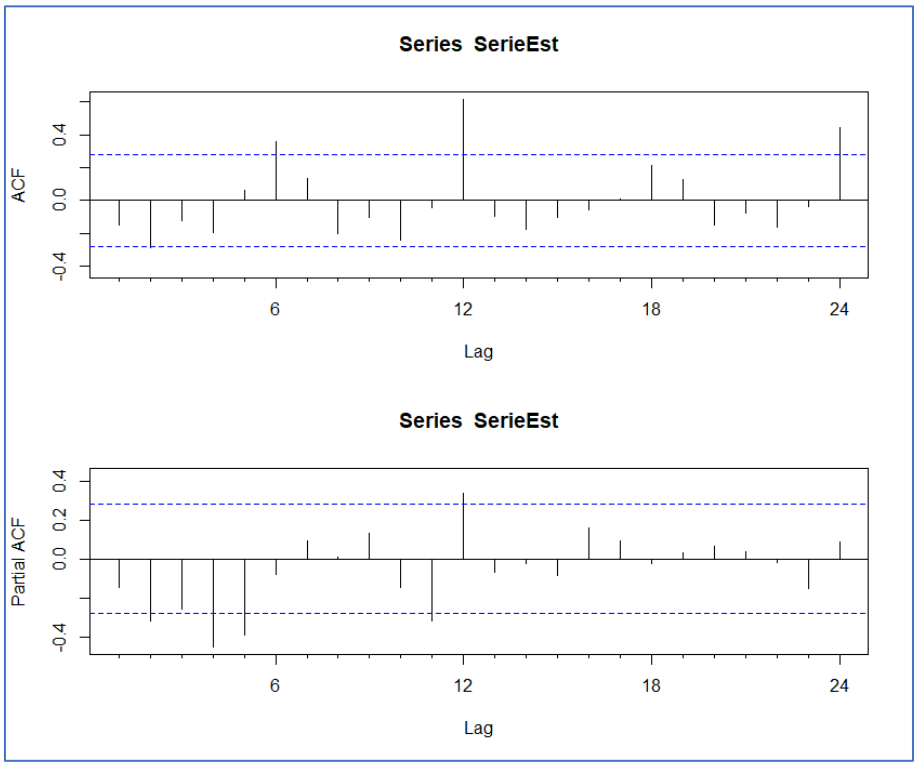

Ilustración 3-29 Autocorrelaciones Serie Refrigeración

Analizando los correlogramas se puede determinar de manera visual los componentes p y q. De acuerdo con el PACF (Función de autocorrelaciones parciales) sugiere un AR (0) o AR (1) y AR (1) estacional mientras que en la gráfica de ACF muestra rezagos que indican un posible MA (0) o MA (1) y MA (1) estacional.

Se realizaron varios modelos con las combinaciones de cada uno de los parámetros, para seleccionar el mejor se tomó el criterio del AIC:

Tabla 3-15 Selección Modelo Serie Refrigeración Criterio AIC

MODELO	AICc
ARIMA (0,1,1) (1,1,1)	526.3822
ARIMA (1,1,0) (1,1,1)	526.9114
ARIMA (0,1,1) (0,1,1)	528.4417
ARIMA (1,1,1) (1,1,1)	528.9679
ARIMA (1,1,0) (0,1,1)	529.1086
ARIMA (1,1,1) (0,1,1)	530.16
ARIMA (1,1,0) (1,1,0)	531.0957
ARIMA (1,1,1) (1,1,0)	531.7161
ARIMA (0,1,1) (0,1,0)	545.2172
ARIMA (1,1,1) (0,1,0)	545.3249

La tabla 3-15 muestra los modelos propuestos que mejor se ajustan a la serie de la categoría refrigeración en base a las combinaciones de los parámetros AR y MA, para lo cual elegiremos los dos modelos que tienen menor AIC para analizarlas en el apartado de Modelo con Errores ARIMA. A continuación, mostraremos las medidas de los modelos y comprobaremos los supuestos de los residuos.

Para el modelo ARIMA (0,1,1) (1,1,1) obtuvimos un coeficiente para la media móvil de (q) de -0.3484 y la parte estacional de -0.9998 y para la parte autorregresiva estacional de -0.41192 mostrada en la Tabla 3-16, mientras que en la tabla 3-17 encontramos las medias de error del modelo

Tabla 3-16 Coeficientes Modelo Arima (0,1,1)(1,1,1) Serie Refrigeración

Coeficientes	ma1	sar1	sma1
valores	-0.3484	-0.4192	-0.9998
s.e.	0.1669	0.1619	0.4743

Tabla 3-17 Medias error del Modelo Arima (0,1,1) (1,1,1) Serie Refrigeración

Medidas de Error – Datos de entrenamiento:					
	ME	**RMSE**	**MAE**	**MPE**	**MAPE**
Datos entrenamiento	2.30	90.36	57.69	-0.12	6.03

A continuación, comprobaremos los supuestos de normalidad y de ruido blanco de los residuos:

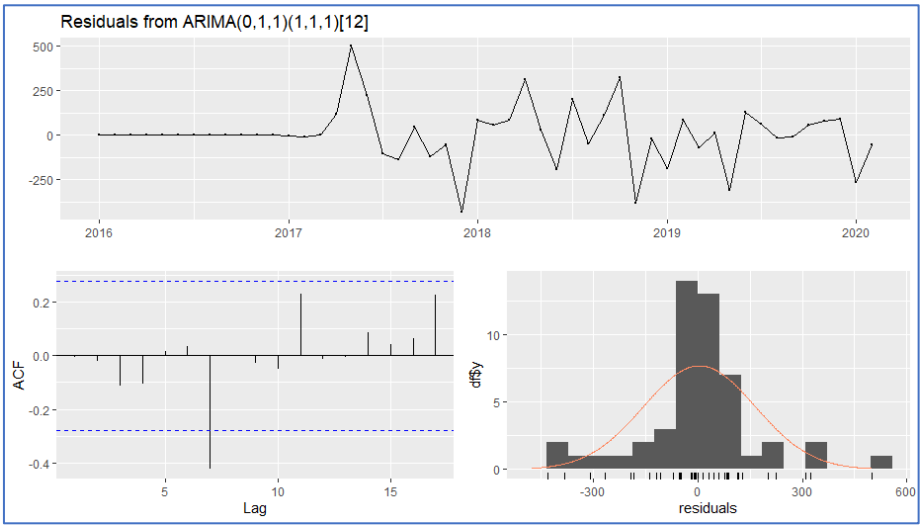

Ilustración 3-30 Supuestos Normalidad y Ruido Blanco Modelo Arima (0,1,1) (1,1,1) Serie Refrigeración

En la ilustración 3-30 podemos observar que los residuos asumen una conducta de ruido blanco, lo comprobamos con el test de Ljung-Box al obtener un valor p de 0.08416 indicándonos que los errores cumplen con el supuesto de independencia, en el gráfico de autocorrelación presenta un rezago. Para comprobar el supuesto de normalidad se utilizó el test de Shapiro-Wilk en la gráfica (arriba) se puede distinguir que en el histograma se asemeja a una campana como la distribución normal esto es una forma gráfica de comprobar el supuesto.

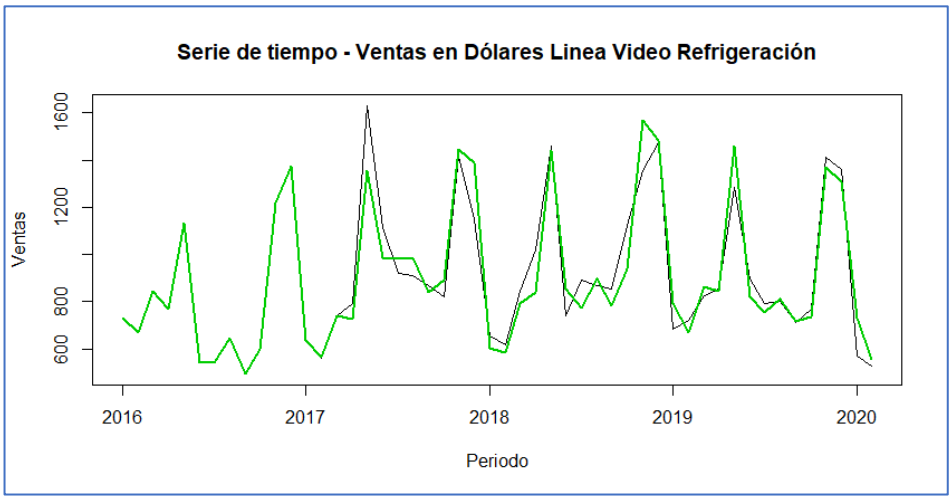

Ilustración 3-31 Serie Original vs Serie Ajustada Modelo Arima (0,1,1) (1,1,1) Serie Refrigeración

En la ilustración 3-31 se compara la serie normal y la serie ajustada por el modelo de color verde, se puede apreciar que se asemeja a la mayoría de los puntos, este modelo tiene un menor RMSE y MAPE.

Para el modelo ARIMA (1,1,0) (1,1,1) obtuvimos un coeficiente para la parte autorregresiva (p) de -0.2929 la parte estacional de -0.4215 y de media móvil de (q) de -1 mostrada en la Tabla 3-18, mientras que en la Tabla 3-19 encontramos las medias de error del modelo

Tabla 3-18 Coeficientes Modelo Arima (1,1,0) (1,1,1) Serie Refrigeración

Coeficientes	ar1	sar1	sma1
valores	-0.2929	-0.4215	-1
s.e.	0.155	0.1605	0.4606

Tabla 3-19 Medias error del Modelo Arima (1,1,0) (1,1,1) Serie Refrigeración

Medidas de Error – Datos de entrenamiento:					
	ME	RMSE	MAE	MPE	MAPE
Datos entrenamiento	1.92	91.02	58.87	-0.05	6.22

Procedemos a comprobar los supuestos de normalidad y de ruido blanco de los residuos:

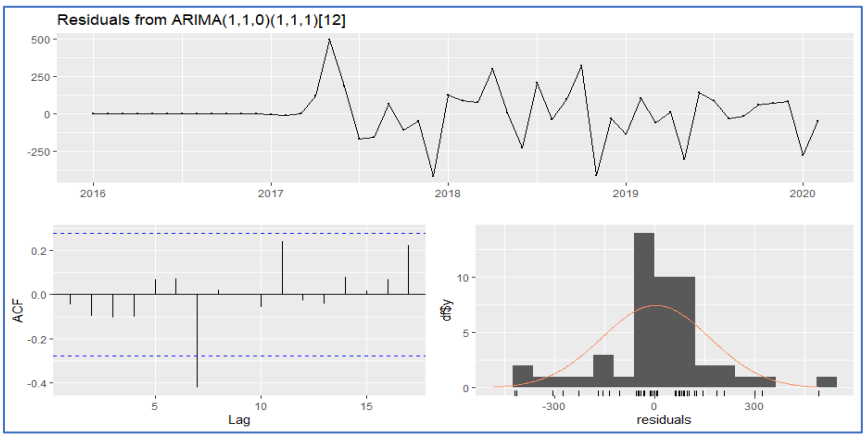

Ilustración 3-32 Supuestos Normalidad y Ruido Blanco Modelo Arima (1,1,0) (1,1,1) Serie Refrigeración

En la imagen 3-32 podemos observar que los residuos poseen una conducta de ruido blanco, lo comprobamos con el test de Ljung-Box al obtener un valor p de 0.06225 indicándonos que los errores cumplen con el supuesto de independencia, en el gráfico de autocorrelación presenta un único rezago. Para comprobar el supuesto de normalidad se utilizó el test de Shapiro-Wilk, de igual manera lo comprobamos con la gráfica del histograma.

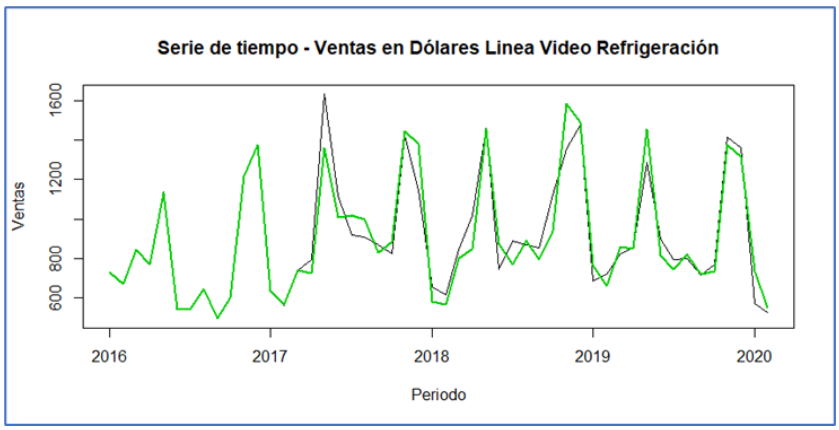

Ilustración 3-33 Serie Original vs Serie Ajustada Modelo Arima (1,1,0) (1,1,1) Serie Refrigeración

En la ilustración 3-33 se compara la serie normal y la serie ajustada por el modelo (color verde), se puede apreciar que se asemeja a la mayoría de los puntos.

3.2.2.2 Modelo de Regresión múltiple Refrigeración

Una vez analizado los modelos ARIMA, se procede a la creación de modelo de Regresión múltiple con la variable de interés que es la venta total y las variables explicativas que se componen en variables económicas y variables de Empleo detalladas en la Tabla 3-20.

Tabla 3-20 Correlación Variables Externas vs Refrigeración

	Venta	Costo Canasta	Precio Barril	Desem Abierto	Empleo Adec	Sub Empleo	IPC	Var IPC
Venta	1	0.16	0.20	-0.37	0.14	0.18	0.00	-0.37
Costo Canasta	0.16	1	0.24	-0.48	-0.16	0.17	0.72	0.09
Precio Barril	0.20	0.24	1	-0.71	-0.14	0.20	0.21	-0.09
Desempleo Abierto	-0.37	-0.48	-0.71	1	0.16	-0.25	-0.40	0.19
Empleo Adecuado	0.14	-0.16	-0.14	0.16	1	0.37	-0.30	0.07
Sub Empleo	0.18	0.17	0.20	-0.25	0.37	1	0.07	-0.14
IPC	0.00	0.72	0.21	-0.40	-0.30	0.07	1	0.20
VarIPC	-0.37	0.09	-0.09	0.19	0.07	-0	0.20	1.00

En la matriz de correlación (Tabla 3-20) observamos que el costo de la canasta básica, el precio de barril de petróleo, empleo adecuado y el subempleo se correlacionan positivamente con las ventas siendo el precio de barril quien tiene un valor más alto 0.20. Las variables de desempleo abierto, el IPC y la variación del IPC se correlacionan negativamente con las Ventas, es decir mientras aumentan este valor disminuye las ventas, el que más se correlación es el desempleo abierto y VarIPC con un valor de -0.37.

Para seleccionar el mejor modelo utilizaremos el método de Forward, el cual consiste en comenzar con un modelo “vacio” solo con la variable de interés e ir seleccionando uno por uno las variables explicativas de acuerdo con cuál de estas disminuye el AIC repitiendo el proceso hasta que ninguna de las variables logre disminuir el indicador.

En la tabla 3-21 mostramos los modelos con cada una de las variables que se fue seleccionando con su respectivo AIC:

Tabla 3-21 Modelos Regresión Múltiple Refrigeración

Modelos	**AIC**	**R^2**
Venta = 1	570.39	0
Venta = DesempleoAbierto	565.22	0.116
Venta = DesempleoAbierto + VarIPC	561.69	0.191
Venta = DesempleoAbierto + VarIPC + EmpleoAdecuado	560.79	0.220

Observamos que la primera variable elegida fue Desempleo abierto, luego fueron VarIPC y Empleo adecuado obteniendo tres variables explicativas para el modelo, al elegir las otras variables se aumenta el AIC como se muestra en la tabla 3-22, en donde la mejor opción fue no elegir más variables.

Tabla 3-22 Regresión Múltiple Seleccionado Refrigeración

Venta = DesempleoAbierto + VarIPC + EmpleoAdecuado		
	Sum of Sq	AIC
+No Hacer nada		560.79
+ Costo Canasta	17662.9	562.51
+ Precio Barril	15314.9	562.55
+ Sub Empleo	3297.8	562.74
+ IPC	260.8	562.79

La variable de Empleo adecuado no es significativa, para el análisis del modelo final se tendrá en cuenta esto para determinar si es necesario incluir o no la variable.

Tabla 3-23 Coeficientes Modelo Seleccionado Refrigeración

	Parametro	Std. Error	t value	Pr(>|t|)
Intercepto	-472.4	1309.6	-0.36	0.720
DesempleoAbierto	-20459.1	7842.3	-2.61	1.22E-02
VarIPC	-386.1	157.1	-2.46	0.0178
EmpleoAdecuado	5504.7	3319.4	1.66	0.104
Modelo: Venta = -472.4 -20459.1 DesempleoAbierto -386.1 VarIPC +5504.7 EmpleoAdecuado				

Analizando los coeficientes del modelo (Tabla 3-23) se observa que por cada punto porcentual que incrementa el Desempleo Abierto las ventas disminuirían en promedio $204.59, por cada punto porcentual que aumente del empleo adecuado la venta incrementa en $55.04, mientras que por cada punto porcentual que incrementa la variación del IPC disminuye la venta en $38.61.

A continuación, mostramos las medidas de error del modelo:

Tabla 3-24 Mediadas de error del modelo seleccionado Refrigeración

Medidas de Error – Datos de entrenamiento:					
	ME	**RMSE**	**MAE**	**MPE**	**MAPE**
Datos entrenamiento	-1.36E-14	251.61	202.24	-7.45	24.04

En la ilustración 3-34 se compara la serie normal y la serie ajustada por el modelo de color rojo

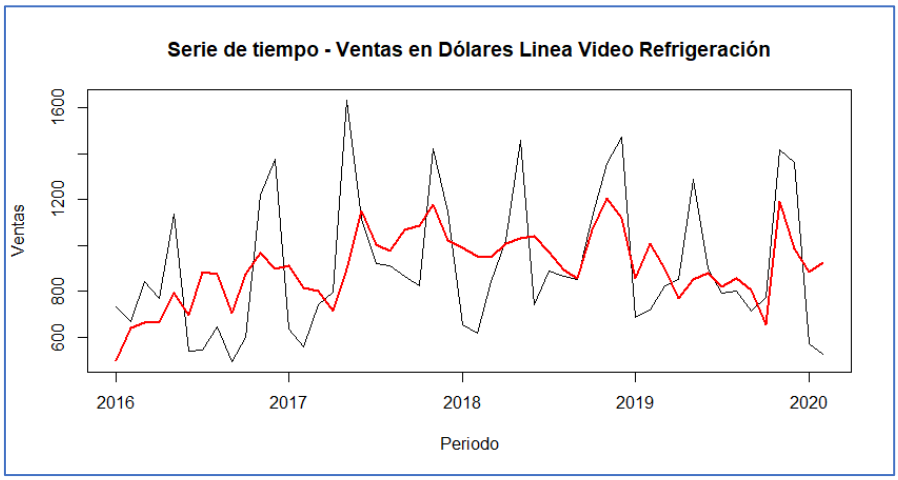

Ilustración 3-34 Serie Original vs Serie Ajustada Modelo Regresión Múltiple Refrigeración

3.2.2.3 Modelo con Errores Arima Refrigeración

Para esta sección utilizaremos los modelos de ARIMA y de Regresión múltiple propuestos manteniendo dos términos de error al combinar ambos métodos de predicción.

Primero realizaremos el modelo de errores ARIMA (0,1,1) (1,1,1) con las variables del modelo de regresión múltiple obteniendo los siguientes coeficientes:

Tabla 3-25 Coeficiente Modelo con Errores Arima (0,1,1)(1,1,1) Refrigeración

Coeficientes	ma1	sar1	sma1	Desempleo Abierto	VarIPC	Empleo Adecuado
valores	-0.3284	-0.2914	-0.9999	-725.9523	-229.6819	2710.224
s.e.	0.157	0.2174	0.4965	17059.2313	139.7056	6344.394

Tabla 3-26 Medidas de Error Modelo con Errores Arima (0,1,1) (1,1,1) Refrigeración

Medidas de Error – Datos de entrenamiento:					
	ME	RMSE	MAE	MPE	MAPE
Datos entrenamiento	3.97E+00	90.56	57.77	0.20	6.15

Visualizamos las medidas de error del modelo en la tabla 3-26

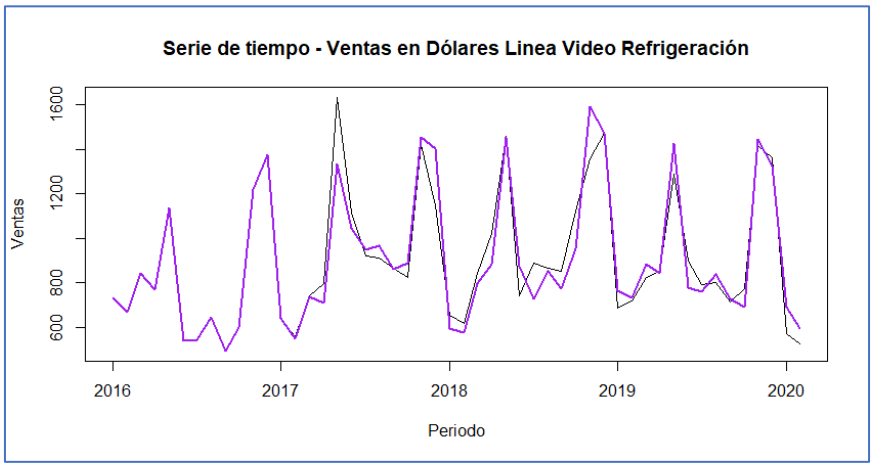

Ilustración 3-35 Serie Original vs Serie Ajustada Modelo con Errores Arima (0,1,1) (1,1,1) Refrigeración

Observamos (Ilustración 3-35) que la línea ajustada difiere en ciertos puntos de la serie original, se procederá a analizar el siguiente modelo ARIMA (1,1,0) (1,1,1), obteniendo la siguiente ecuación (Tabla 3-27):

Tabla 3-27 Coeficiente Modelo con Errores Arima (1,1,0) (1,1,1) Refrigeración

Coeficientes	ar1	sar1	sma1	Desempleo Abierto	VarIPC	Empleo Adecuado
valores	-0.3262	-0.2795	-1	-393.9121	-245.466	4148.434
s.e.	0.1556	0.2221	0.4915	17182.1038	136.8648	6266

Tabla 3-28 Medidas de Error Modelo con Errores Arima (1,1,0) (1,1,1) Refrigeración

Medidas de Error – Datos de entrenamiento:					
	ME	RMSE	MAE	MPE	MAPE
Datos entrenamiento	4.27E+00	90.86	58.94	0.31	6.35

Se visualiza las medidas de error del modelo en la tabla 3-28, en la ilustración 3-36 tenemos la curva ajustada del modelo de color naranja.

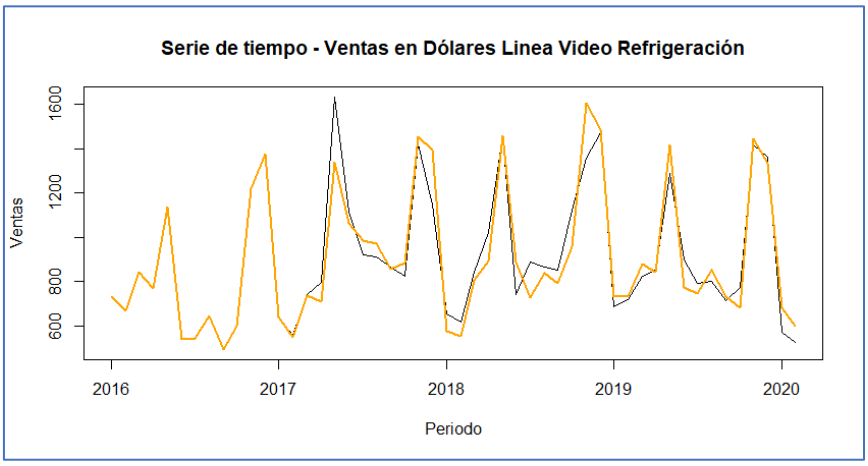

Ilustración 3-36 Serie Original vs Serie Ajustada Modelo con Errores Arima (1,1,0) (1,1,1) Refrigeración

Observamos que la línea ajustada difiere en ciertos puntos de la serie original. En ambos modelos seleccionados presentan medidas de error muy similares, en otra sección seleccionaremos el mejor modelo con la data de prueba.

3.2.3 Elaboración del Sistema de Recomendación

Vamos a realizar 3 enfoques para el sistema de recomendación:

1. Filtrado Colaborativo
2. Basado en Contenido
3. Hibrido

Previamente se depura la data, eliminando aquellos clientes que tienen una sola compra para evitar cualquier tipo de sesgos. Se deja 3 tablas en las cuales en una está detallada alguna característica de la categoría, en la segunda se detalla las características del usuario y la tercera se encuentra los ID del usuario con cada ID de la categoría que ha comprado y su respectivo ranking.

Para el Ranking consideramos la compra de alguna de las categorías por parte de los usuarios como estos pueden ser comprados más de una vez. Por lo tanto, para modelar el interés del usuario agregamos todas las compras que ha realizado de cada categoría realizando una suma ponderada de la intensidad del tipo de interacción y aplicamos una transformación logarítmica para suavizar la distribución.

Construimos la matriz de interacción del ID de cada categoría con el ID del usuario, en la tabla 3-29 se muestra una pequeña parte de la matriz original 199 x 243542 los valores de 0 se consideran a los que nunca han comprado esa sublinea y así de manera exponencial:

Tabla 3-29 Muestra Matriz Interacción Categoría - Usuario

personId	100011543	100107440	100324581	100340843	993005258001	9930200
sublinea						
A/A SPLIT 12000BTU	0.0	0.0	0.0	0.0	0.0	0.0
A/A SPLIT 12000BTU NVERTER	0.0	0.0	0.0	1.0	2.0	3.5
A/A SPLIT 18000BTU	0.0	0.0	0.0	0.0	0.0	0.0
A/A SPLIT 18000BTU NVERTER	0.0	0.0	0.0	0.0	2.0	0.0
A/A SPLIT 24000BTU	0.0	0.0	0.0	0.0	0.0	0.0

Para el primer enfoque basado en contenido utilizaremos la librería Surprise y compararemos los modelos propuestos con medidas de error y cross-validation.

Ilustración 3-37 Medidas de desempeño de cada modelo

En la ilustración 3-37 se muestran las medidas de desempeño de cada uno de los modelos, se aplicó la técnica de cross validation y se sacó el promedio de estos indicadores. El modelo que tiene mejor desempeño es el NMF (Matrices de factorización no negativa) que tiene mejor RMSE y MAE.

Se puede estandarizar la cantidad de recomendaciones que nos arroje el sistema ordenando de acuerdo con la probabilidad de similitud que tenga los ítems con el usuario, en la siguiente ilustración se muestra las recomendaciones de un usuario:

```
1: COCINA,
2: LAVADORA
3: LICUADOR
4: TV LED +
5: TV LED 3
```

Ilustración 3-38 Ejemplo Recomendación Modelo Basado en Contenido

Procedemos a utilizar el paquete LightFM para realizar un modelo basado en contenido de acuerdo con las características de las categorías de producto. Se utiliza diferentes variables que describen a las categorías de

producto como marca, tipo, status para ver similitudes entre ellos. En la imagen observamos similitudes con la categoría Celular “AAA”:

```
Item of interest :CELU  AAA  \RO
Item similar to the above item:
1- TV LCD 40-42
2- TV LCD 32-37
3- SERVICIO INTERNET
4- TABLET HIBRIDA
5- MICROCOMPONENTES
```

Ilustración 3-39 Similitudes Categoría Celular AAA

Luego de tener el modelo basado en contenido de las categorías de productos procedemos a realizar el modelo hibrido con la misma matriz de interacción entre usuario e ítems que se usó para el filtrado colaborativo obteniendo un test (Tabla 3-30) un poco más bajo que el del filtrado colaborativo mostrada en la siguiente tabla.

Tabla 3-30 Test Filtrado Colaborativo

test_time
0.61

Aplicando el modelo para un usuario obtenemos las siguientes recomendaciones:

```
1: COCINA,
2: HOR-MICR,
3: PLANCHA,
4: LAVADORA,
5: SANDWICHERA,
```

Ilustración 3-40 Ejemplo Modelo basado en filtros colaborativos

Posteriormente analizaremos los test con la data de prueba para comprobar los modelos.

3.2.4 Selección del Modelo

En secciones previas analizamos algunos modelos de ARIMA, regresión múltiple y la combinación de ambos denominada “modelo con errores ARIMA”, en esta ocasión vamos a seleccionar el mejor modelo en base al mejor desempeño con la data de prueba ya que nos interesa que el modelo pueda predecir la venta de cada línea con el menor error posible. Las métricas para elegir el mejor modelo son:

- RMSE (Raíz del Error Cuadrático Medio)
- MAE (Error Absoluto Medio)

Medidas de Error que nos ayudaran a determinar el mejor modelo para nuestras series.

3.2.4.1 Modelo Línea Video

Para la línea video analizamos 5 modelos como se puede apreciar en la tabla 3-31. Los dos primeros son los de ARIMA, luego tenemos el de regresión múltiple terminando con los dos modelos de errores ARIMA.

Tabla 3-31 Modelos Línea de Video

		TRAIN		TEST	
MODELO	**AICc**	**RMSE**	**MAE**	**RMSE**	**MAE**
ARIMA (0,1,1) (0,1,0)	219.54	347.41	243.33	606.00	485.96
ARIMA (1,1,1) (0,1,0)	220.57	345.77	234.38	651.66	522.70
Venta ~ DesempleoAbierto + VarIPC + IPC	780.68	537.99	427.57	935.46	827.81
ARIMA (0,1,1) (0,1,0) + DesempleoAbierto + VarIPC + IPC	222.63	335.39	237.29	542.74	451.61
ARIMA (1,1,1) (0,1,0) + DesempleoAbierto + VarIPC + IPC	225.25	334.61	234.75	536.86	454.40

Analizando los modelos se tienen menor error con la data de entrenamiento, esto es debido a que se ajustaron bastante bien a los datos históricos y que en la data de prueba se encuentra meses donde no se realizaron ventas debido a la pandemia aumentando el error.

Los modelos con errores ARIMA tienen las métricas más bajas, siendo el modelo ARIMA (0,1,1) (0,1,0) + DesempleoAbierto + VarIPC + IPC quien tiene mejor MAE con 451.61 y el otro modelo ARIMA (1,1,1) (0,1,0) + DesempleoAbierto + VarIPC + IPC tienen menor RMSE con 536.86. Elegimos el modelo que tiene menor RMSE ya que existe una mayor brecha siendo una raíz.

Procederemos a evaluar los supuestos de los residuos del modelo el cual no debe presentar Autocorrelación y debe seguir una distribución normal.

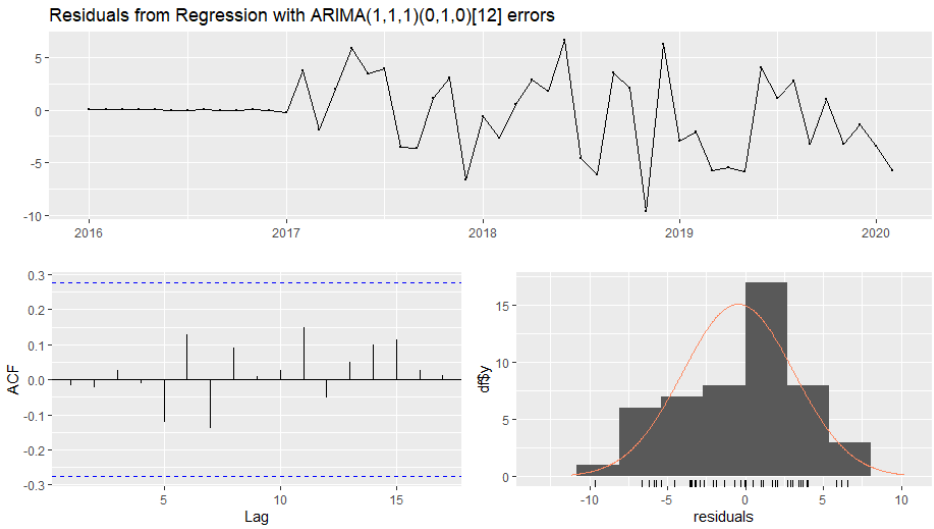

Ilustración 3-41 Supuestos Normalidad y Ruido Blanco Modelo con Errores Arima Video

Observamos que en la ilustración 3-41 la autocorrelación no presenta rezagos, se utilizó el test de Ljung-Box con un valor p de 0.6083 indicándonos que la serie cumple con el supuesto de independencia y mediante la gráfica del histograma podemos determinar que sigue una distribución normal, de igual manera se comprobó con el test de Shapiro con un valor p de 0.3756.

Se muestra en la tabla 3-32 los valores del forecast de venta con la data real.

Tabla 3-32 Forecast Modelo Elegido Video

Fecha	Forecast	Real	Lo 80	Hi 80
mar-20	1006	601	715	1355
abr-20	836	1	553	1190
may-20	1421	369	1004	1924
jun-20	1758	670	1260	2354
jul-20	1210	851	800	1722
ago-20	1183	1035	764	1712
sep-20	1012	889	620	1521
oct-20	1487	1137	970	2137
nov-20	2025	1674	1378	2821
dic-20	2054	1464	1383	2883
ene-21	655	967	326	1124
feb-21	470	811	201	879
mar-21	661	1155	246	1329
abr-21	673	1028	230	1409
may-21	1044	1339	418	2020
jun-21	1365	1194	586	2546

En los meses de pandemia no existe una buena predicción, en el resto de los meses se encuentra dentro de los límites y en los últimos dos meses ya se acerca bastante a la real, para esta línea hay que aclarar que se hizo ciertas promociones a principio del año 2021 para liquidar ciertos productos por tal motivo se observa una gran cantidad de venta en aquellos meses que no se venden mucho.

3.2.4.2 Modelo Línea Refrigeración

Para la línea Refrigeración analizamos 7 modelos como se muestra en la tabla 3-33. Los dos primeros son los de ARIMA, luego tenemos el de regresión múltiple terminando con los 4 modelos de errores ARIMA.

Tabla 3-33 Análisis de los modelos y Selección del mejor modelo

		TRAIN		TEST	
MODELO	AICc	RMSE	MAE	RMSE	MAE
ARIMA (0,1,1)(1,1,1)	526.38	90.36	57.69	444.99	348.25
ARIMA (1,1,0)(1,1,1)	526.91	91.02	58.87	431.63	330.38
Venta = DesempleoAbierto + VarIPC + EmpleoAdecuado	560.79	537.99	427.57	692.68	584.64
ARIMA (0,1,1)(1,1,1)) + DesempleoAbierto + VarIPC	529.22	90.22	56.60	407.53	308.03
ARIMA (1,1,0)(1,1,1) + DesempleoAbierto + VarIPC	529.39	90.35	57.68	407.66	305.35
ARIMA (0,1,1)(1,1,1) + DesempleoAbierto + VarIPC + EmpleoAdecuado	532.09	90.56	57.77	267.37	184.09
ARIMA (1,1,0)(1,1,1) + DesempleoAbierto + VarIPC + EmpleoAdecuado	532.01	90.86	58.94	222.09	177.23

Analizando los modelos se tienen menor error con la data de entrenamiento, esto es debido a que se ajustaron bastante bien a los datos históricos y que en la data de prueba se encuentra meses que no hubo mucha venta por la pandemia aumentando el error.

Los modelos con errores ARIMA tienen las métricas más bajas, siendo el modelo ARIMA (1,1,0) (1,1,1) + DesempleoAbierto + VarIPC + EmpleoAdecuado quien tiene mejor MAE con 177.23 y menor RMSE con 222.09 elegimos este modelo en base a la métrica planteada. La variable Empleo adecuado no fue significativa en el modelo de regresión, pero al analizar el RMSE determinamos que si influye en la predicción de la venta.

Procederemos a evaluar los supuestos de los residuos del modelo el cual no debe presentar Autocorrelacion y debe seguir una distribución normal.

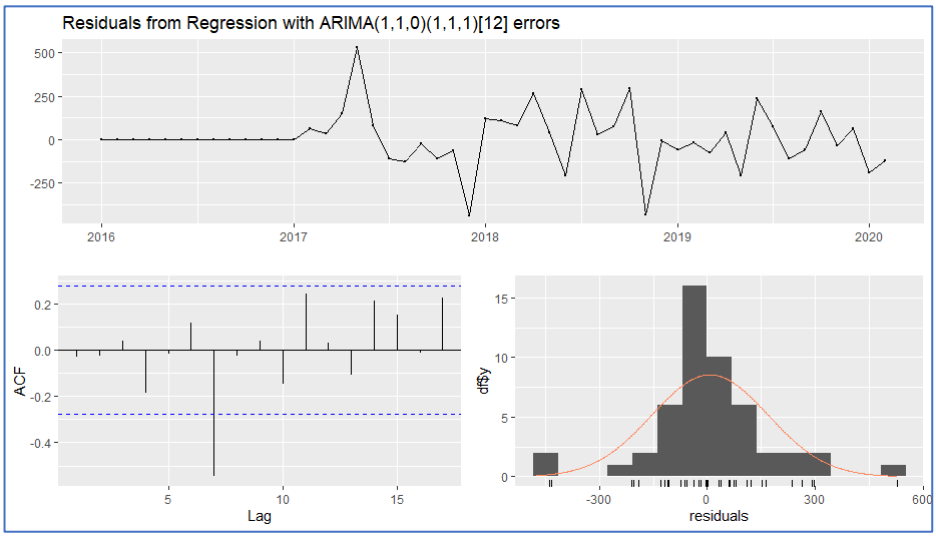

Ilustración 3-42 Supuestos Normalidad y Ruido Blanco Modelo con Errores Arima (1,1,0) (1,1,1) Refrigeración

Observamos que en la ilustración 3-42 de autocorrelación presenta un rezago, se utilizó el test de Ljung-Box con un valor p de 0.0001528 indicándonos que la serie no cumple con el supuesto de independencia y mediante la gráfica del histograma podemos determinar que sigue una distribución normal, de igual manera se comprobó con el test de Shapiro con un valor p de 0.001762 que no sigue una distribución normal.

Ya que el modelo no cumple con los supuestos, se procede a analizar el siguiente modelo con errores ARIMA ARIMA (0,1,1)(1,1,1) + DesempleoAbierto + VarIPC + EmpleoAdecuado.

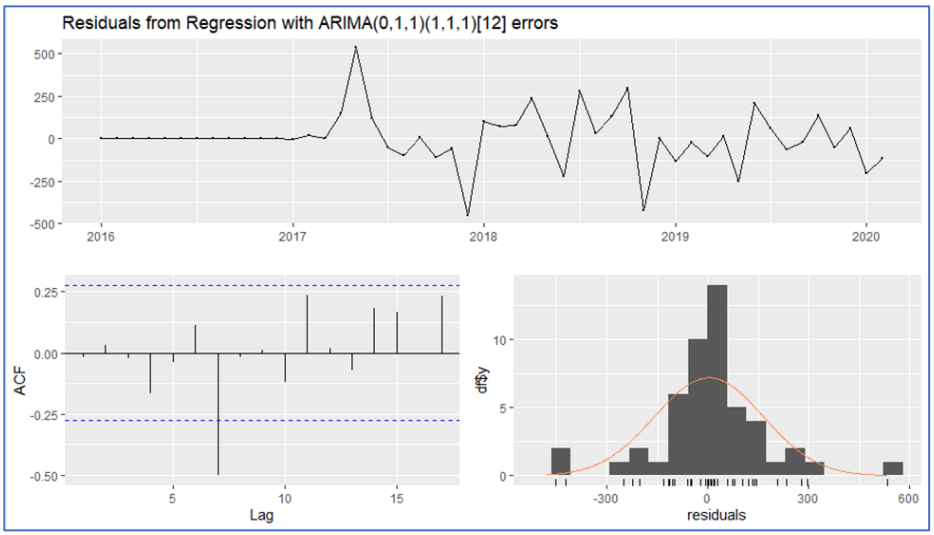

Ilustración 3-43 Supuestos Normalidad y Ruido Blanco Modelo con Errores Arima (0,1,1)(1,1,1) Refrigeración

Utilizamos el test de Ljung-Box con un valor p de 0.002179 indicándonos que la serie no cumple con el supuesto de independencia.

Los modelos con errores ARIMA no cumplen con los supuestos de los residuos, por ende, se selecciona el modelo ARIMA (1,1,0) (1,1,1) cumpliendo las condiciones, a continuación, se muestra el forecast y la data real del modelo seleccionado:

Tabla 3-34 Forecast Modelo Elegido Refrigeración

Fecha	Forecast	Real	Lo 80	Hi 80
mar-20	726	317	558	891
abr-20	801	1	597	1002
may-20	1389	263	1157	1618
jun-20	748	406	469	1018
jul-20	741	454	432	1039
ago-20	754	514	418	1077
sep-20	698	506	334	1047
oct-20	817	647	436	1184
nov-20	1264	840	878	1640
dic-20	1252	854	846	1648
ene-21	600	553	135	1033
feb-21	569	557	79	1022
mar-21	736	664	256	1190
abr-21	790	675	302	1253
may-21	1314	911	843	1771
jun-21	778	528	268	1261

De igual manera en los meses de pandemia no hay un buen pronóstico (Tabla 3-34), en el resto de los meses el valor real está dentro de los límites y en su mayoría cercano al valor predicho.

3.2.4.3 Sistema de Recomendación

Para estos modelos utilizamos test de precisión para corroborar la validez, en la siguiente tabla se presentan los resultados de los test:

Tabla 3-35 Resultado de los test de precisión

MODELO	test_time
Filtrado colaborativo	0.61
Basado en Contenido	0.84
Hibrido	0.88

Como observamos el modelo Hibrido de recomendaciones es el que mejor desempeño tiene (Tabla 3-35).

Recalcamos que al aplicar el modelo de filtrado colaborativo tiene buenas predicciones. Por tanto, si una empresa no tiene característica del usuario o categoría, este modelo es buena opción para aplicar en el negocio teniendo buenas recomendaciones de categoría.

3.2.5 Módulos de sistema

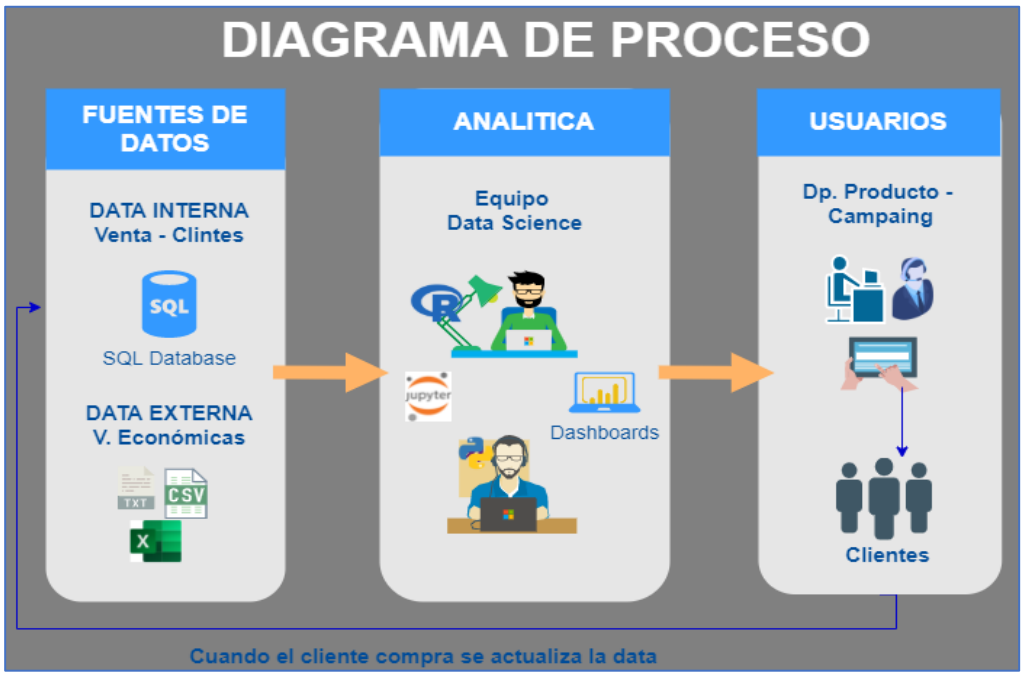

Ilustración 3-44 Diagrama de Procesos

La ilustración 3-44 muestra los pasos que sigue la analítica a implementar, en primera instancia obtenemos datos de ventas con características de clientes cuya fuente es las tablas transaccionales de la empresa y tenemos variables económicas en archivos “csv” o “txt” que son consumidas por el equipo de Data Science para elaborar la analítica, en nuestro proyecto son las predicciones y un sistema de recomendaciones de categoría de producto con su respectiva visualización de datos; el mismo equipo es encargado de darle mantenimiento continuo a los modelos predictivos y al sistema de recomendación.

Los jefes de productos consumen la herramienta de predicciones para la toma de decisiones acorde al comportamiento de cada una de las categorías, por otro lado, los encargados del campaing utilizaran las recomendaciones de las categorías de productos para contactar al cliente con oferta dirigida con alta probabilidad de compra. Al comprar el cliente algún producto, la data se actualiza periódicamente optimizando el sistema de recomendación y los pronósticos de ventas.

3.3 Plataformas y prototipos de visualización

Como Prototipo realizamos visualización en Power BI la cual interactúa con los comandos de Rstudio mostrando estadísticas descriptivas de las variables de interés, las series de tiempo de las categorías y de los modelos elegidos anteriormente definidos, por ultimo las recomendaciones de categorías y productos para un usuario seleccionado con su número de cedula (Para el desarrollo del sistema de recomendaciones utilizamos la herramienta Python).

A continuación, se mostrará dos imágenes preliminares de las visualizaciones de series de tiempo:

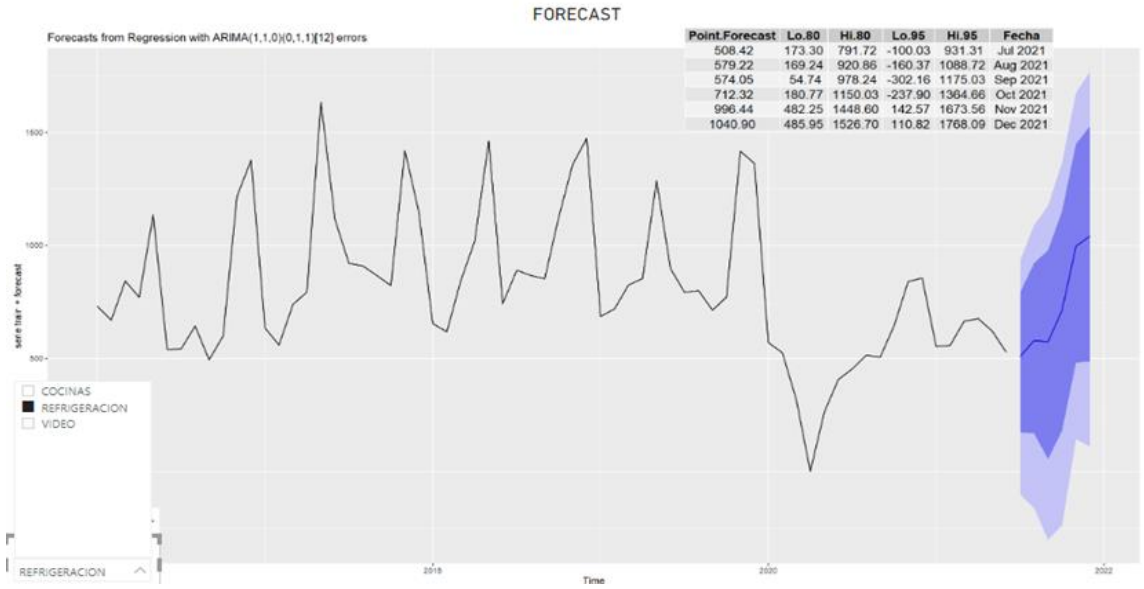

Ilustración 3-45 Prototipo Visualización Imagen 1

En la Ilustración 3-45 se encuentra la serie de la Línea Refrigeración con sus respectivas predicciones de 6 meses e intervalos de confianza de acuerdo con el modelo; de igual manera al seleccionar otra línea de producto cambia el gráfico de la serie con sus respectivas predicciones como se muestra en la Ilustración 3-46 el gráfico de la Línea Video.

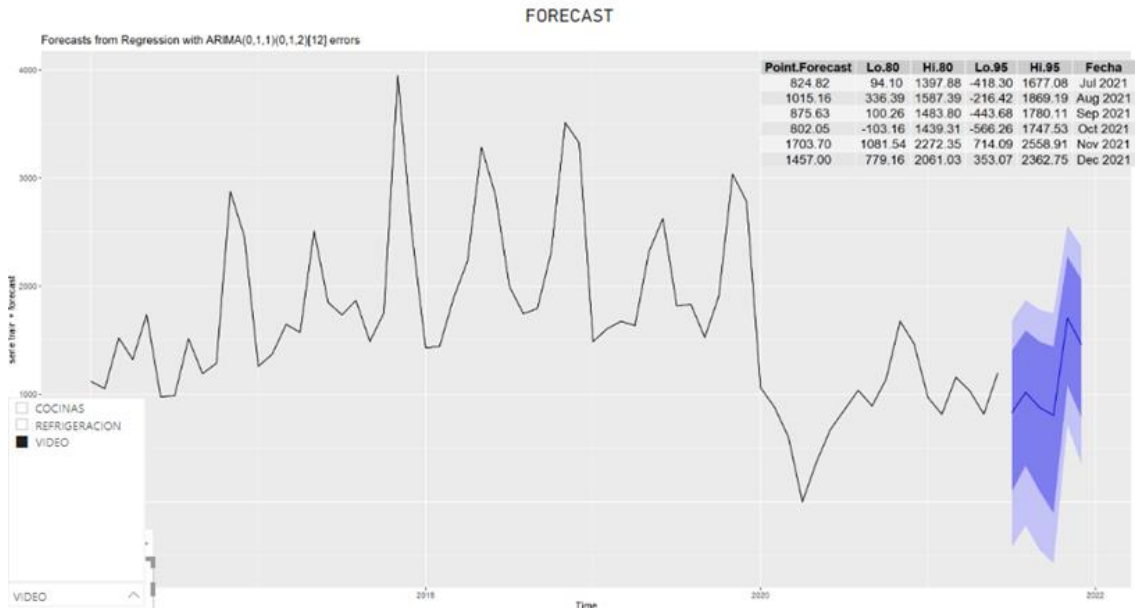

Ilustración 3-46 Prototipo Visualización Imagen 2

El usuario podrá seleccionar la categoría de su interés y el tipo de venta (Crédito – Contado - Todas). La dificultad de este prototipo es que actualmente los códigos de R no son compatibles con la versión web de Power BI, imposibilitándonos compartir de forma rápida a los usuarios finales.

Como solución a este inconveniente se aprovechará que la empresa cuenta con licencias de Microsoft para el personal, cada uno cuenta con un correo personal siendo este utilizado para compartir el dashboard mediante OneDrive a los usuarios finales y solamente dichos usuarios podrán visualizar el reporte restringiendo a que pueda ser compartido a otras personas.

3.4 Métricas y comunicación de resultados

La tabla 3-36 compara las predicciones de los modelos seleccionados vs el forecast de la empresa acorde a la venta real.

Tabla 3-36 Línea Video

Tabla Fecha	Forecast	Real	Pronostico Interno
jul-21	825	982	817
ago-21	1015	845	1021
sep-21	876	762	882
oct-21	802	964	1164
nov-21	1704	2213	1720
dic-21	1457	1539	1381

Se evidencia un mejor pronóstico del modelo propuesto para la línea Video en los distintos meses exceptuando noviembre ya que en este mes el pronóstico interno se acerca más al dato real de venta (mes de Black Friday). Concluimos que el modelo propuesto da buenos pronósticos mejorando al que actualmente tiene la empresa.

Tabla 3-37 Línea Refrigeración

Fecha	Forecast	Real	Pronostico Interno
jul-21	508	514	462
ago-21	579	526	449
sep-21	574	469	451
oct-21	712	591	685
nov-21	996	1284	918
dic-21	1041	949	731

Para la línea Refrigeración se muestran las predicciones en la tabla 3-37 al comparar con la venta real y los pronósticos internos se ve un mejor pronóstico exceptuando el mes de septiembre y octubre en donde pronostica un valor más alto que lo real.

Los pronósticos de los modelos propuestos se asemejan más al dato real de venta de cada una de las líneas de productos analizadas, algunos de los valores de forecast de la empresa se localizan en los intervalos que maneja el modelo.

Con la implementación del sistema de recomendaciones para consumo local, se espera incrementar las ventas y la concreción de las campañas realizadas a través de sms (aplicaciones que reciben mensajes como whatsapp, telegram) y mediante tele vendedores que realizaran llamadas ofreciendo estas recomendaciones a los clientes con el imput de contar con un producto diferenciado para cada uno.

Tabla 3-38 Métricas y Resultados

Acciones	Análisis	%Concreción Actual	%Concreción Implementado
Implementar el sistema de recomendación para consumo local (campaing, marketing)	Envió de sms y correo, el cliente compra el producto recomendado.	**1.9%**	**3.5% - 5.5%**
Implementar el sistema de recomendación para consumo local (tele vendedores)	Llama al cliente a ofrecer el producto.	**2.5%**	**4.5% - 6.5%**

Para analizar el incremento de esta métrica se lanzó una campaña de sms a ciertos clientes con una promoción de categoría arrojada por nuestro modelo de recomendación, la cual en 2 semanas resulto con una concreción de 2.9% mejor que anteriores campañas realizadas por esta vía contando con menos tiempo para que el cliente se acerque a la tienda (Tabla 3-38).

Evidenciamos una mejoría tanto de pronóstico como de concreción de venta en las primeras pruebas que se realizaron con los modelos seleccionados, dejándonos buenas señales para la implementación final en la empresa.

CAPÍTULO 4

4. ANÁLISIS DE RESULTADOS

4.1 Recolección de datos y estrategias para validación del proyecto

4.1.1 Recolección de datos

En la recopilación de los datos utilizamos varias fuentes categorizándolas fuente de datos internos y fuente de datos externos; la data interna fue seleccionada de las tablas transaccionales de ventas alojadas en un almacén de datos dentro de los servidores de la institución para un periodo desde ene-2016 hasta junio-2021. Por otra parte, la data externa fue elegida de fuentes de instituciones públicas básicamente del INEC y el Banco central que mensualmente estiman indicadores económicos y de empleo para el mismo periodo establecido en los datos internos.

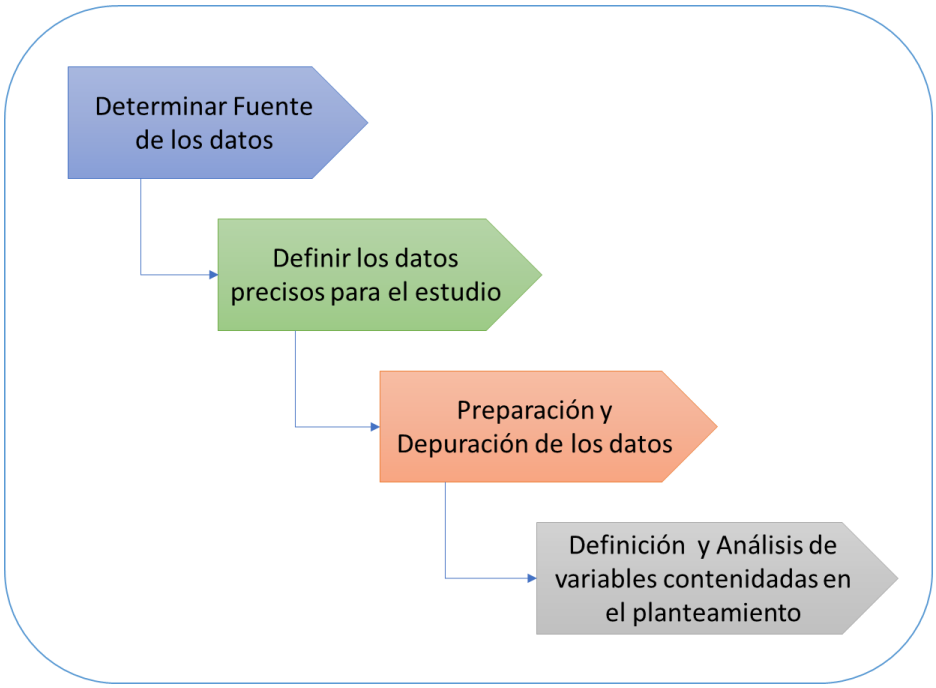

Ilustración 4-1 Diagrama Procesos Recolección de datos

En la ilustración 4-1 se observa un diagrama de procesos de la recolección de los datos en primera instancia se ***determinó la fuente de los datos*** las cuales ya hicimos mención en anteriores apartados:

- Base de datos Transaccional Ventas (Interno de la Empresa)
- INEC (Indicadores de empleo y subempleo)
- Banco Central (Indicadores Económicos)

En el siguiente punto del proceso se ***definieron los datos y variables*** para el estudio que pudieran responder a la problemática planteada, obteniendo un total de 19 variables en la data interna y 8 variables para los datos externos.

La preparación y depuración de los datos es el proceso más importante en este apartado ya que implica una manipulación y estructuración de la data para el caso puntual del dataset de las ventas transaccionales que se encuentran distribuidas a lo largo de todo el periodo establecido. Por lo consiguiente a priori se realizó una depuración de los valores erróneos, así como de los vacíos encontrados en la data, posteriormente se agruparon los datos de tal manera que se pueda obtener un valor de las ventas mensuales en dólares y unidades por cada categoría de la empresa conteniendo una medida comparativa al dataset externo que se utilizaron para los modelos predictivos de ventas por categorías.

A su vez para el sistema de recomendaciones se utilizó la data transaccional de ventas depurada con variables demográficas y que caractericen la compra del cliente.

Como último procedimiento realizado para la preparación de la data, ***analizamos las variables*** de forma descriptiva para observar las categorías más vendidas, tendencias y proporciones que respondan al planteamiento de la problemática.

4.1.2 Estrategias para validación del proyecto

Basado al planteamiento del presente proyecto y a la implementación de el mismo se realizará una propuesta de verificación, comprobación o validación de todas las metodologías aplicadas durante el desarrollo de este.

Por lo tanto, la propuesta de validación es la de experimentación, se plantea que los usuarios finales puedan interactuar con los sistemas generados (“Dashboard con las Predicciones”, “Sistema Recomendaciones”) esta iteración se evaluara bajo rangos y parámetros que estableceremos en conjunto con los altos mandos de la empresa. Esta evaluación nos dará resultados que evidencien y justifique la implementación del proyecto pretendiendo que la propuesta sea aplicable.

A pesar de que la propuesta es mejorar las predicciones de ventas con metodologías eficientes y realizar un sistema de recomendaciones de las principales categorías de la empresa debido a no contar con herramientas estadísticas actualizadas que posibiliten tener una mejor planificación y toma de decisiones a nivel gerencial, la validación de las nuevas implementaciones se basará también en contrastar las anteriores metodologías versus las que se proponen en esta tesis y así encontrar diferencias significativas que conlleven a establecer que metodología es la más acertada y semejante a la realidad.

Para poder cuantificar la experimentación de los sistemas planteados analizaremos las siguientes características:

- Alcance (Si los objetivos planteados fueron solventados durante el desarrollo del proyecto)
- Comparación (Comparar metodologías vigentes versus metodologías propuestas)

- Funcionamiento (Si la propuesta implementada funciona correctamente)
- Aplicabilidad (Si el proyecto es aplicable en los diferentes contextos propuestos)

Para este apartado utilizaremos métricas medibles para la evaluación de cada una de las características mencionadas; Por tanto, planteamos elaborar un cuestionario que plasme o cuantifique cada propiedad ponderando con un peso mayoritario la comparación de metodologías, que tiene como finalidad establecer mejores predicciones y recomendaciones que las actuales.

4.2 Puesta en marcha y funcionamiento

Se procedió a realizar el dashbord en Power BI de manera que sea interactivo y amigable para los usuarios con el fin de que tomen las mejores decisiones en base a los pronósticos y recomendaciones brindadas.

Las principales cualidades en la que se decidió al uso de esta herramienta:

- Fácil acceso a los datos.
- El dashboard se puede revisar en tiempo real.
- Fácil manejo y totalmente personalizable para los usuarios.
- Toma de decisiones basado en datos de la empresa
- Interactividad con herramientas estadísticas Rstudio y Python.

La empresa cuenta con licencias y facilitó la implementación del dashboard para esta primera implementación, uno de los problemas que se evidencia es que los complementos de Rstudio y Python no son compatibles con la versión web de Power BI, para esto se tuvo que compartir el archivo en una carpeta de onedrive y luego asignarles permisos a los usuarios de la empresa que manejaran las opciones.

Procederemos a explicar el funcionamiento del aplicativo implementado en Power BI; como primer paso el usuario deberá tener un usuario y contraseña para poder ingresar y visualizar la información contenida en el aplicativo, estableciendo este punto procederemos a mostrar los diferentes lienzos contenidos en el Dash:

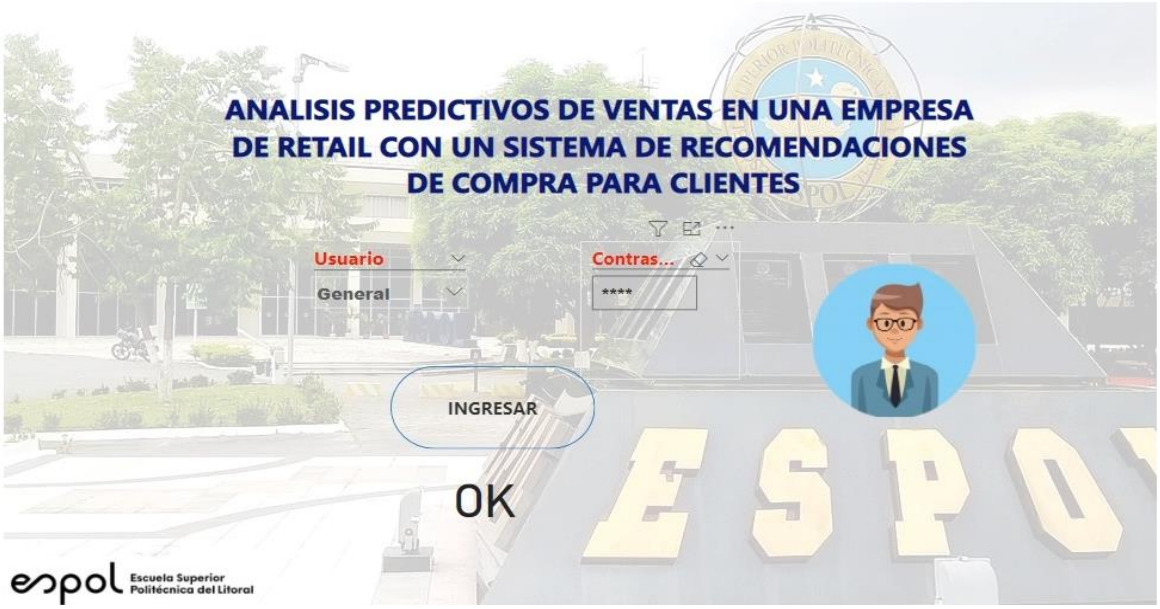

Ilustración 4-2 Dashboard Visualización Login

El usuario ingresara al Dashboard con el usuario y contraseña proporcionado por el administrador del aplicativo (Ilustración 4-2).

Ilustración 4-3 Dashboard Visualización Menú Principal

Posterior de haberse logeado se mostrará un menú con las alternativas de análisis que deseé visualizar (Ilustración 4-3), el usuario podrá seleccionar cualquier alternativa dando click en cualquier recuadro lo cual lo llevará al análisis seleccionado.

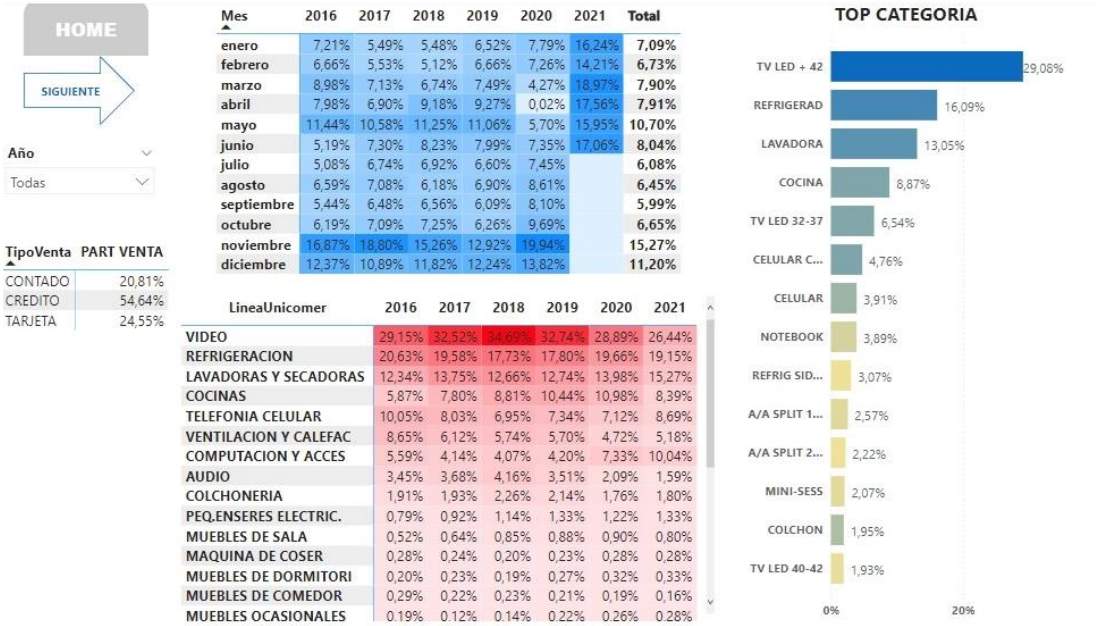

Ilustración 4-4 Dashboard Visualización Descriptivas 1

Si el usuario seleccionó la opción Descriptivo se muestra en pantalla un análisis de la proporción de ventas totales por cada línea de producto en cada año de análisis, así como también los meses en los que más se vendió porcentualmente y por último en la parte derecha el top de los productos más vendidos. Adicional el usuario podrá interactuar con el dashboard seleccionando los años que desee visualizar en los descriptivos (Ilustración 4-4).

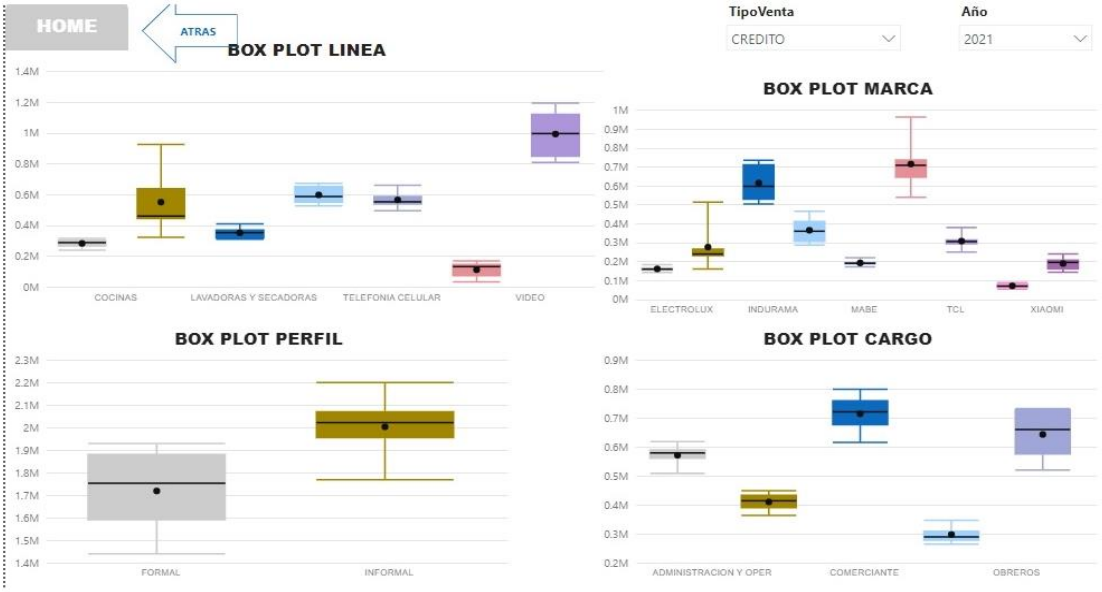

Ilustración 4-5 Dashboard Visualización Descriptivas 2

La ilustración 4-5 muestra la segunda parte del análisis descriptivo, básicamente esta muestra los diagramas de cajas de ciertas variables categóricas respecto al volumen de venta, estos boxplot pueden ser filtrados por el tipo de venta y año que se desee.

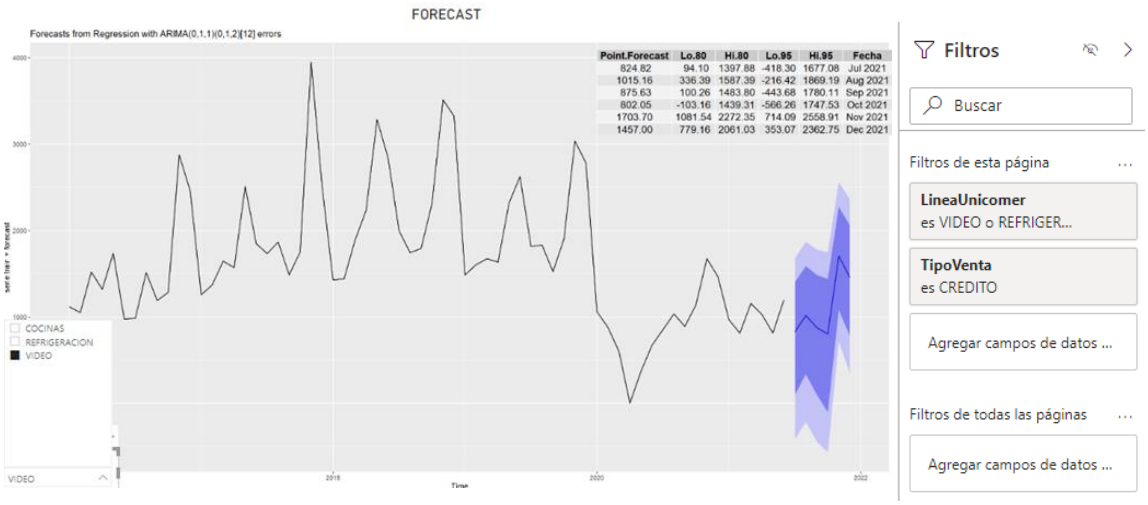

Ilustración 4-6 Dashboard Visualización Pronóstico Video

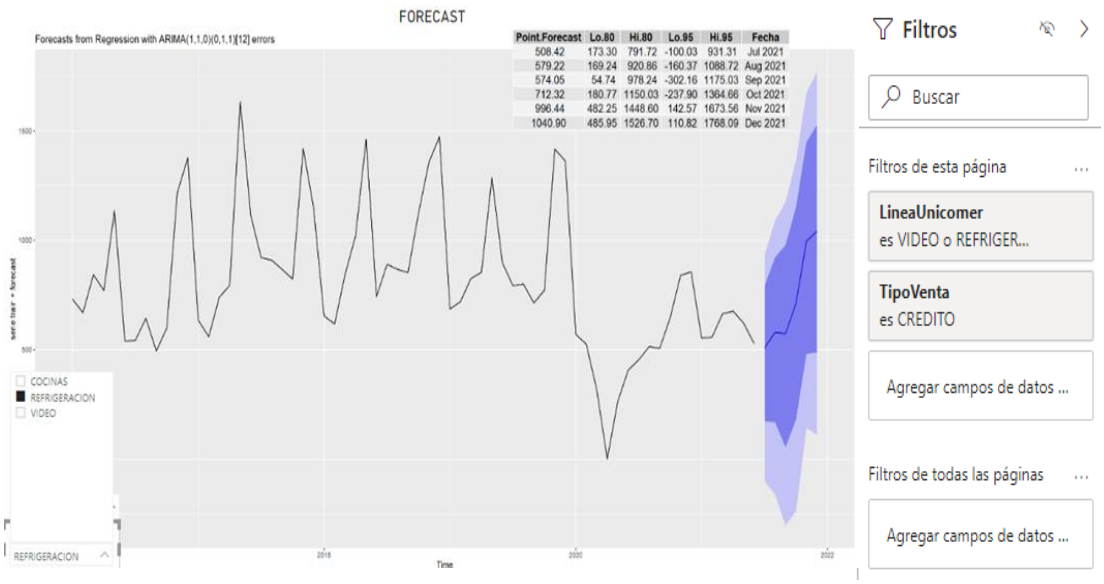

Ilustración 4-7 Dashboard Visualización Pronóstico Refrigeración

Presentamos las predicciones de los modelos en donde el usuario puede filtrar por las categorías de producto y por el tipo de venta que se tiene de acuerdo con las necesidades que tenga. En la ilustración 4-6 está el modelo ARIMA y en la ilustración 4-7 el modelo con ERRORES ARIMA para que se compare los resultados.

Ilustración 4-8 Dashboard Visualización Recomendaciones

En la ilustración 4-8 analizaremos las recomendaciones asociadas a la compra de una categoría como a los clientes. En la parte superior el usuario seleccionara una categoría mostrando como resultado alternativas que puede asociar al producto (esto está enfocado netamente a las categorías). En la parte inferior de la pantalla nos enfocamos en el cliente, seleccionando un id el sistema arroja la recomendación personalizada por cada individuo (Top 3).

4.3 Pruebas de funcionalidad

Como indicamos en los capítulos previos se decidió utilizar el programa de Power BI con los complementos de Rstudios y Python para realizar la implementación.

Los comentarios de los usuarios:

Tabla 4-1 Comentarios – Pruebas de Funcionalidad

USUARIO	MODELO	COMENTARIOS
Usuario 1	Predicción	Se compara con la venta real, siendo muy eficiente en los pronósticos.
Usuario 2	Predicción	Las predicciones son más cercanas que el presupuesto establecido.
Usuario 3	Recomendaciones	Se lanzó una campaña con las recomendaciones la cual nos sirvió de input para los clientes.

Se realizó una campaña de sms con las recomendaciones arrojadas por el modelo logrando aumentar la concreción de venta en 0.6%, se prevé que aumentaran la concreción en 1.2% en el próximo mes ya que se lanzara campaña dirigidas.

Los pronósticos de venta han servido para calibrar el dato del presupuesto en el transcurso de los años, teniendo datos que se asemejan a la realidad. Como las métricas las imponen un ente externo, para esto tipos de pronósticos trabajaremos con indicadores internos logrando una estabilidad en las metas de la empresa.

4.4 Análisis costo/beneficio

Se procede a realizar el análisis de las diferentes acciones posibles al implementar los modelos propuestos, planteamos inicialmente tres posibles acciones de implementación las cuales son:

1. Sistema de recomendación para consumo local (campaing, markting) y modelos predictivos.
2. Sistema de recomendación en los puntos de venta y modelos predictivos.
3. Solo los modelos predictivos

A continuación, haremos un análisis de las acciones posibles bajo diferentes enfoques de costo/beneficio.

Tabla 4-2 Resultado del Análisis

Acciones posibles	Resultados del Análisis
Implementar el sistema de recomendación para consumo local (campaing, marketing) y modelos predictivos.	• Incremento de concreción de ventas • Mejorar las campañas de crédito • Aumento de rentabilidad de la empresa
Implementar el sistema de recomendaciones en los puntos de venta y modelos predictivos.	• Incremento de concreción de ventas. • Mejorar las campañas de crédito • Aumento de rentabilidad de la empresa • Ofertas optimas en los canales de venta
Implementar modelos predictivos.	• Mejorará la gestión de los jefes de producto.

En la Tabla 4-2 compara las posibles implementaciones con los beneficios de cada uno, al realizarlo en los puntos de ventas existen un mejor beneficio al ser más dirigidas las recomendaciones en el momento que el cliente este en el local.

Tabla 4-3 Restricciones / Limitaciones

Acciones posibles	Restricciones
Implementar el sistema de recomendación para consumo local (campaing, marketing) y modelos predictivos.	• Números de teléfonos y correos erróneos de los clientes para envío de promociones, se debe mejorar la contactabilidad. • Se debe gestionar los productos con valor agregado.
Implementar el sistema de recomendaciones en los puntos de venta y modelos predictivos.	• Se gestionará solo en el punto de venta. • Resistencia al cambio por parte de los vendedores • Mal uso de las promociones por parte de punto de venta.
Implementar modelos predictivos.	• Mal enfoque de los jefes de producto.

Se analizaron las limitaciones que tendrían cada opción de implementación, lo cual al realizar mediante campaig la principal limitación es la contactibilidad al tener números de teléfono desactualizados, al realizarlo directamente en puntos de venta existe el riesgo de manipulación por parte de los vendedores realizando mal uso de estas recomendaciones como se detalla en la Tabla 4-3.

Tabla 4-4 Supuestos Estratégicos

Acciones posibles	Supuestos
Implementar el sistema de recomendación para consumo local (campaing, marketing) y modelos predictivos.	• El proyecto cuenta con el aval del jefe inmediato • El proyecto está alineado con la estrategia • Los interesados les van a dar el tiempo que se necesita
Implementar el sistema de recomendaciones en los puntos de venta y modelos predictivos.	• El proyecto está alineado con la estrategia
Implementar modelos predictivos.	• El proyecto cuenta con el aval del jefe inmediato • El proyecto está alineado con la estrategia • Los interesados les van a dar el tiempo que se necesita

La Tabla 4-4 detalla los supuestos estratégicos al implementar cada una de las opciones, todos los casos están alineados con la estrategia de la empresa, pero por el momento solo el implementar el sistema para consumo local y los modelos predictivos cuentan con el aval del jefe inmediato.

Tabla 4-5 Riesgos

Acciones posibles	Riesgos
Implementar el sistema de recomendación para consumo local (campaing, marketing) y modelos predictivos.	• Que no se cumpla con los tiempos • Que no se incrementen las ventas • Resistencia al cambio
Implementar el sistema de recomendaciones en los puntos de venta y modelos predictivos.	• Que no se cumpla con los tiempos • Que no se incrementen las ventas • Resistencia al cambio • Molestia en punto de venta
Implementar modelos predictivos.	• Que no se cumpla con los tiempos • Resistencia al cambio

La Tabla 4-5 nos muestra los Riesgo de implementación de cada uno de los casos, el riesgo surge al no poder incrementar la concreción de venta y no cumplir con los tiempos de implementación. Se deberá capacitar a los usuarios

explicándoles los beneficios del proyecto así evitar resistencias e inconformidades.

Tabla 4-6 Análisis Financiero

Acciones posibles	Financiero
Implementar el sistema de recomendación para consumo local (campaing, marketing) y modelos predictivos.	Actualmente se tiene contrato con empresa de envió de sms y correo
Implementar el sistema de recomendaciones en los puntos de venta y modelos predictivos.	Inversión de sistema para que aparezca en los puntos de venta al momento de realizar un pago o cotización. $5000
Implementar modelos predictivos.	Ninguno

Al analizar las propuestas en el ámbito financiero se determinó que la opción de implementación en los puntos de venta se debería invertir un valor aprox de $5000 mientras que en las otras no es necesario alguna inversión esto se detalla en la Tabla 4-6.

Se comparó las mejoras que lograrían los indicadores al implementar los modelos en diferentes ámbitos y se decide realizar la implementación para consumo local enviando promociones por sms y correos al estar acorde a la necesidad del negocio y no tener algún gasto de implementación. Como una mejora al proyecto se lo realizará en los puntos de ventas y ecommerce cuando se encuentre con los recursos necesarios.

5. CONCLUSIONES Y RECOMENDACIONES

5.1 Conclusiones

- Fue indispensable obtener variables económicas externas y adherirlas a los métodos predictivos para que generen ruido; lo que mejoró las predicciones de los modelos.
- Para los modelos de línea Video las variables económicas de Desempleo y variación del IPC influyen en la predicción de venta, mientras que en Refrigeración no fue significativa ninguna variable externa.
- Las predicciones se ajustan a la venta real de la empresa obteniendo mejores predicciones que el modelo actual.
- Al desarrollar el sistema de recomendación, obtuvimos buenos resultados al desarrollar el modelo hibrido, que consiste en combinar el modelo simple de filtrado colaborativo y el modelo basado en contenido.
- Es indispensable tener características de la categoría del producto para poder recomendar al cliente.
- Se obtuvo un incremento en la concreción de las campañas de crédito al mandar recomendaciones de acuerdo con el modelo.
- Power BI nos permite visualizar de manera efectiva los datos, con conexiones de distintas fuentes.
- Las visualizaciones son interactivas permitiéndonos tomar las decisiones apropiadamente en base a los datos.
- La interacción por parte de los usuarios fue satisfactoria al interactuar entre los lienzos.

5.2 Recomendaciones

- La empresa debe ser más flexible en proporcionar los datos a los encargados de su manipulación y convertirlos en información para la correcta toma de decisiones.
- Dar prioridad a proyectos de ciencia de datos que mejoran el desempeño del negocio incrementando la rentabilidad.

- Automatizar el modelo de recomendaciones en los puntos de venta y tienda online (ECOMMERCE).
- Invertir en proyectos que involucren las mejoras de recopilación de información e inyecta de datos.

BIBLIOGRAFÍA

Carrillo, D. P. (2019). *Repositorio de la Universidad Santiago de Compostela.* Retrieved from Universidad Santiago de Compostela: http://eio.usc.es/pub/mte/descargas/ProyectosFinMaster/Proyecto_1680.pdf

Central, B. (2021). *Banco Central*. Retrieved from Banco Central - Indicadores Económicos: https://www.bce.fin.ec/index.php/informacioneconomica/

Céspedes Urrutia, A. I. (2017, 11). *REPOSITORIO USM.* Retrieved from REPOSITORIO UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA: http://hdl.handle.net/11673/41250

Cruz Trejos, E. U. (2011). Evaluación de la robustez de un modelo de regresión múltiple para predecir las ventas diarias de un hipermercado en Pereira. *Scientia Et Technica,*, 95100.

d'Arc, T. (n.d.). *SmartHint Venta de comercio electrónico*. Retrieved from SmartHint: https://www.smarthint.co/es/o-que-e-um-sistema-de-recomendacao/?utm_source=casedesucesso&utm_medium=blog&utm_campaign=bigboy%2F%2F%2F#01

Ghobar, E. W. (2017). *Universidad Politécnica de Valencia.* Retrieved from Repositorio Universidad Politécnica de Valencia: https://espolec-my.sharepoint.com/personal/jvillao_espol_edu_ec/Documents/Tesis/TUTOR/Paper/WALID%20-%20Un%20sistema%20de%20recomendaci%C3%B3n%20basado%20en%20perfiles%20generados%20por%20agrupamiento%20y%20asociaciones.pdf?CT=1630740010907&OR=ItemsView

INEC. (2021). *INSTITUTO NACIONAL DE ESTADÍSTICAS Y CENSOS*. Retrieved from INEC - Estadísticas Laborales: https://www.ecuadorencifras.gob.ec/estadisticas-laborales-enero-2022/

J. Villavicencio. (2010). *Notas de clases Series de Tiempo.* Retrieved from Notas de clases Series de Tiempo: https://d1wqtxts1xzle7.cloudfront.net/38458362/manual_intro_series_tiempo-with-cover-page-v2.pdf?Expires=1630792825&Signature=KR9a-

Ri7zavkJvZzH8VC9iPSkknnWX3zxCSbw7C63XvM0-lu5ffaWQyP7rza8jLerJpwKRQsBG5DV0EGfdwZ0LAhxd4vFErJdE77xfbAYR8YM~Skpu0RlxMl36PXCegIz

Jinkun Wang, Y. L. (2016). Diversified Recommendation Incorporating Item Content Information Based on MOEA/D . *Hawaii International Conference on System Sciences*, 688 - 696.

Keunho Choi, D. Y. (2011). A hybrid online-product recommendation system: Combining implicit rating-based collaborative filtering and sequential pattern analysis. *Electronic Commerce Research and Applications*, 309-317.

Loepp B., H. T. (2014). Choice-based preference elicitation for collaborative filtering recommender systems. *In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, (pp. 3085-3094).

Matías Bavera, B. B. (2018, 10). *REPOSITORIO UCA.* Retrieved from UNIVERSIDAD CÁTOLICA DE ASUNCIÓN: http://clei2019.utp.ac.pa/storage/app/uploads/public/5d8/ce0/6f2/5d8ce06f2457d533287433.pdf

Paul Paucar, L. Y. (2020, 01 20). *Universidad Técnica Salesiana Sede Cuenca Repositorio.* Retrieved from Universidad Técnica Salesiana Sede Cuenca: https://dspace.ups.edu.ec/handle/123456789/18288

Pramodh, C. (2020, 6 28). *Towards Data Science*. Retrieved from Towards Data Science: https://towardsdatascience.com/netflix-recommender-system-a-big-data-case-study-19cfa6d56ff5

River, G. (2021). *Green River*. Retrieved from Green River: https://www.greenriver.com/portfolio/recommendation-system-case-study

Sampín, S. D. (2021). *Repositorio Espol.* Retrieved from Repositorio Espol.

Tulasi K. Paradarami, N. D. (2017). A Hybrid Recommender System Using Artificial Neural Networks. *Expert Systems With Applications*, 9-22.

Uriel, E. (2013, 09). *Repositorio Universidad de Valencia.* Retrieved from Universidad de Valencia: https://d1wqtxts1xzle7.cloudfront.net/55082644/3_Regresion_lineal_multiple_estimacion_y_propiedades-with-cover-page-v2.pdf?Expires=1630798993&Signature=cNgfT-

A75vmfTNiApzFLObokIn5zmSDY4FZaLuOUJKMzMhTmw~TaaVS8KLjVYXRQd
ncDWRDeKdVPZEolZwUumBKVDyxKiObbdnIt0L9

ANEXOS

A continuación, se comparte el enlace donde se encuentra el archivo que contiene el código utilizado tanto para los modelos de predicción como el de sistema de recomendación:

Codigo.txt