UNIVERSIDAD CENTRAL DE VENEZUELA
FACULTAD DE CIENCIAS
ESCUELA DE MATEMÁTICAS
POSTGRADO MODELOS ALEATORIOS

APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA PREDECIR PATRONES DE CONSUMO DE LA CADENA CENTRAL MADEIRENSE Y CLASIFICAR SUS TIENDAS SEGÚN EL DESEMPEÑO.

Autor: LIC. ILIANA VARGAS

Tutor: DR. RICARDO RIOS

Trabajo de Grado de Maestría presentado ante la ilustre Universidad Central de Venezuela para optar al título de Magister Scientiarium, mención Modelos Aleatorios.

UNIVERSIDAD CENTRAL DE VENEZUELA
FACULTAD DE CIENCIAS
COMISIÓN DE ESTUDIOS DE POSTGRADO

VEREDICTO

Comisión de Estudios
de Postgrado

Quienes suscriben, miembros del jurado designado por el Consejo de la Facultad de Ciencias de la Universidad Central de Venezuela, para examinar el **Trabajo de Grado** presentado por: Iliana Beatriz Vargas Camacho, **Cédula de identidad** 19.165.384, bajo el título "**APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA PREDECIR PATRONES DE CONSUMO DE LA CADENA MINORISTA CENTRAL MADEIRENSE Y CLASIFICAR SUS SUCURSALES SEGÚN SU DESEMPEÑO**", a fin de cumplir con el requisito legal para optar al grado académico de **MAGÍSTER SCIENTIARUM, MENCIÓN MODELOS ALEATORIOS**, dejan constancia de lo siguiente:

1.- Leído como fue dicho trabajo por cada uno de los miembros del jurado, se fijó el día 19 de febrero de 2019 a las 09:15 am., para que la autora lo defendiera en forma pública, lo que ésta hizo en LAPROESA, mediante un resumen oral de su contenido, luego de lo cual respondió satisfactoriamente a las preguntas que le fueron formuladas por el jurado, todo ello conforme con lo dispuesto en el Reglamento de Estudios de Postgrado.

2.- Finalizada la defensa del trabajo, el jurado decidió **aprobarla** por considerar, sin hacerse solidario con la ideas expuestas por el autor, que se ajusta a lo dispuesto y exigido en el Reglamento de Estudios de Postgrado.

Para dar este veredicto, el jurado estimó que el trabajo proporciona información sobre el uso de herramientas de la minería de datos para la clasificación fina de las sucursales de la cadena Central Madeirense mediante la detección de pocas variables, vía modelos de clasificación, el desempeño de cada una, logrando una interesante combinación de herramientas teóricas de la estadística matemáticas con algoritmos de fácil uso para el cliente no especializado.

En fe de lo cual se levanta la presente ACTA, a los 19 días del mes de febrero del año 2019, conforme a lo dispuesto en el Reglamento de Estudios de Postgrado, actuó como Coordinador del jurado Ricardo Ríos.

Mairene Colina / C.I. 12.761.954
Institución UCV

Daniel Barráez / C.I. 6.197.305
Institución UCV

Ricardo Ríos / C.I. 3.949.476
Institución UCV
Tutor

UNIVERSIDAD CENTRAL DE VENEZUELA
FACULTAD DE CIENCIAS
ESCUELA DE MATEMÁTICAS
POSTGRADO MODELOS ALEATORIOS

Maestría en Modelos Aleatorios

APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS PARA PREDECIR PATRONES DE CONSUMO DE LA CADENA CENTRAL MADEIRENSE Y CLASIFICAR SUS TIENDAS SEGÚN EL DESEMPEÑO.

Autor: Lic. Iliana Vargas
Tutor: Dr. Ricardo Ríos

RESUMEN

Este trabajo de grado de maestría tiene como finalidad predecir el desempeño de las sucursales de la cadena minorista Central Madeirense mediante la construcción de un modelo de clasificación con el análisis de series temporales y técnicas de minería de datos sobre el comportamiento de compra de los clientes e indicadores económicos y operativos establecidos por la compañía. Las metodologías estudiadas para este fin fueron: análisis de series temporales por el método Box-Jenkins, regresión logística ordinal y el clasificador de Bayes ingenuo. Mantener la rentabilidad de la compañía y al mismo tiempo ofrecer una buena experiencia de compra a los clientes se ha convertido en un problema complejo debido al decaimiento económico, productivo y desabastecimiento en Venezuela en los últimos años, por lo que esta situación incentiva a la construcción de propuestas que permitan comprender mejor el desempeño del negocio para tomar acciones estratégicas a tiempo. Por medio de este estudio se obtiene un panorama global mes a mes del negocio, permitiendo establecer distintas estrategias, entre ellas comerciales y de mercadeo, para mejorar las compras a proveedores, la rotación de los productos, disminuir los gastos, aumentar las ventas y mantener la rentabilidad de la compañía, además este desempeño se podrá monitorear mediante geolocalización de las sucursales sobre el mapa del país. Se concluye que el modelo que mejor se ajusta a los datos procesados para la clasificación de las sucursales que Central Madeirense dispone a nivel nacional es el clasificador de Bayes ingenuo.

Palabras claves: *Regresión Logística Ordinal, Bayes ingenuo, método Box-Jenkins, Central Madeirense.*

Índice

1. Introducción

Este trabajo de investigación relaciona el análisis de series temporales y técnicas de minería de datos con el comportamiento de compra por parte de los clientes de la cadena minorista Central Madeirense, mediante métodos de aprendizaje estadístico, se construye un modelo de clasificación para las sucursales según su desempeño económico y operativo basándose en los indicadores establecidos por la empresa. El objetivo de esta investigación radica en extraer, analizar y establecer un modelo de pronóstico sobre las ganancias y consumo de los clientes y un modelo de clasificación de las 52 sucursales que la empresa dispone a nivel nacional.

1.1. El problema de los patrones de consumo

Mantener la rentabilidad de la compañía y al mismo tiempo ofrecer una buena experiencia de compra a los clientes se ha convertido en un problema complejo debido al decaimiento económico, productivo y desabasteciemiento en el país en los últimos años, por lo que esta situación incentiva a la contrucción de propuestas que permitan comprender mejor el desempeño del negocio para tomar acciones estratégicas a tiempo.

El objetivo de este trabajo de grado es establecer metodologías de análisis sobre el comportamiento de las ventas en las sucursales, así como también construir un modelo de clasificación de estas, tomando en cuenta el desempeño económico y operacional que permita obtener un panorama global mes a mes del negocio, permitiendo establecer distintas estrategias, entre ellas comerciales y de mercadeo, para así mejorar las compras a proveedores,la rotación de los productos, disminuir los gastos, aumentar las ventas y mantener la rentabilidad de la compañía.

1.2. Objetivos de estudio

Para el desarrollo de este trabajo se definieron los siguientes objetivos, generales y específicas:

1.2.1. Objetivo General

Elaborar metodologías de análisis y modelos de predicción para cada unas de las 52 sucursales por medio de la técnica de Box-Jenkins y construir un modelo que clasifique a las tiendas según su desempeño a partir de la aplicación de la regresión logística ordinal y el clasificador de Bayes ingenuo.

1.2.2. Objetivos Específicos

1. Extraer, procesar y realizar analásis descriptivo de los datos de las series temporales de número de transacciones y ventas regulares.

2. Analásis descriptivo de los datos sobre los indicadores financieros y operacional por cada sucursal.

3. Construir modelos de pronóstico sobre las ganancias y consumo de los clientes en cada una de las 52 sucursales.

4. Construir un modelo de clasificación del desempeño económico y operacional de las 52 sucursales.

1.3. Justificación

El término Minería de Datos se refiere a un campo de la estadística y las ciencias de la computación que consiste en extraer información de un conjunto de datos y transformarla en una estructura comprensible para su uso posterior. Estas técnicas han sido utilizadas en una gran variedad de industrias en los últimos años con el objetivo de incrementar la competitividad, participación de mercado así como también mejorar la gestión de clientes, procesos de negocios, relación con proveedores, entre otros.

La disponibilidad de un gran volumen de datos sobre los consumidores, posibilitado por el avance de herramientas tecnológicas que junto al desarrollo de nuevas técnicas dentro del campo de la Estadística, permiten la conversión de datos en conocimiento de una manera eficiente y a gran escala. Este tipo de análisis han creado oportunidades y desafíos para que las empresas aprovechen esa información y obtengan una ventaja competitiva para tomar mejores decisiones y administrar el negocios con mayor eficacia.

Para esto la empresa Central Madeirense ha colocado a disposición una bases de datos que resume los regitros de ventas regulares en bolívares y número total de transacciones realizadas por cada sucursal durante los años 2012 a abril 2018 y otra base de datos que resume los registros de distintos indicadores financiero y operacionales manejados internamente por la empresa durante el año 2017.

En este trabajo especial de grado nos enfocaremos en analizar el problema desde los datos de manera de obtener una fuente de información veraz que permita aplicar distintas técnicas de minería de datos para lograr construir un modelo de clasificación que pueda pronósticar el desempeño de las sucursales mes a mes, así como establecer modelos de pronósticos de las ventas y transacciones por sucursal adaptables a los cambios económicos que vive actualmente el país.

2. Fundamentos sobre series temporales

En esta sección se plantea los conceptos básicos, descritos en el texto de Peter Brockwell y Richard Davis, para el análisis de series temporales y procesos estocásticos que serán la base para comprender los modelos AR, MA, ARMA, ARIMA y SARIMA, procesos que se discuten en forma introductoria.

2.1. Ideas Básicas

Una serie temporal es un conjunto de observaciones x_t, cada uno registrado en un tiempo específico t. Las series temporales pueden ser: discreta, el cual es aquella donde el conjunto T_0 de los tiempos, en el cual las observaciones son realizadas, es un conjunto discreto; y continuas las cuales son obtenidas cuando las observaciones son registradas sobre algún intervalo de tiempo, por ejemplo, cuando $T_0 = [0, 1]$.[1]

La utilidad de estas series radica en obtener patrones de comportamiento de una variable mediante la observación de sus datos en el transcurso de un periodo de tiempo. En el estudio práctico de las series temporales se mide el tiempo en periodos aproximadamente equidistantes, como por ejemplo minutos, horas, días, meses, años etc.

El estudio de las series temporales permite conocer una variable a lo largo del tiempo y con ello realizar predicciones. Los comportamientos en la variable estudiada pueden obedecer a patrones deterministas o a patrones aleatorios.

Una parte importante del análisis de series de tiempo es la seleción de un modelo probabilístico adecuado (o clases de modelos) para los datos. Para permitir la posible impredecible naturaleza de las observaciones futuras es natural suponer que cada observación x_t es la realización de ciertas variables aleatorias X_t.[1]

Definición 1. *Un modelo de serie temporal para los datos observados x_t es una especificación de la distribución conjunta (o posiblemente sólo las medias y covarianzas) de una serie de variables aleatorias X_t donde x_t representa una realización.*

Un modelo de series temporales completo para la sucesión de variables aleatorias $(X_1, ..., X_n)'$, $n = 1, 2, ...$, o equivalentemente todas las probabilidades

$$P[X_1 \leq x_1, ..., X_n \leq x_n], \quad -\infty < x_1, ..., x_n < \infty, \quad n = 1, 2, ...$$

Tal especificación es raramente usada en análisis de series temporales (a menos que los datos sean generados por un simple y bien entendido mecanismo), donde en general contendrá muchos parámetros a ser estimados a partir de los datos disponibles. En vez de eso, sólo especificamos el primer y segundo momento de la distribución conjunta, esto es, el valor esperado EX_t y el producto esperado $E(X_{t+h}, X_t), \quad t = 1, 2, ..., \quad h = 0, 1, 2, ...$ enfocándose en propiedades de la sucesión X_t que depende sólo de estos. Tales propiedades de X_t son llamadas propiedades de segundo orden. En el caso particular donde todas las ditribuciones conjuntas sean normales multivariadas, la propiedad de segundo orden de X_t determina completamente las distribuaciones conjuntas y así da una completa caracterización probabilistica de la sucesión.[2]

2.2. Procesos Estocásticos

Un proceso estocástico se entiende como una secuencia de datos que evolucionan en el tiempo, siendo así las series temporales un caso particular de los procesos estocásticos. De manera más formal se puede definir un proceso estocástico como una colección de variables aleatorias ordenadas en el tiempo. [3]

Definición 2. *Un proceso estocástico es una familia de variables aleatorias* $X_t : t\epsilon T$ *definidas sobre un espacio de probabilidad* (ω, A, P)*, donde el conjunto paramétrico* T *es un subconjunto de* R*.*

El conjunto T suele ser un intervalo o un conjunto de valores discretos, por consiguiente el proceso estocástico depende de los argumentos, $X(t, w)$ el tiempo $t \epsilon T$ y el suceso elemental $w \epsilon \omega$. Para cada t fijo, $X(t, .)$ es una variable aleatoria y, para cada w fijo, $X(., w)$ es una realización del proceso, por consiguiente una serie temporal es considerada como una realización de un proceso estocástico.

En síntesis en cada instante t existirá una variable aleatoria distinta X_t, por tanto en un proceso estocástico las características de las variables aleatorias varían en el tiempo.[3]

2.2.1. Procesos Estacionarios y Función de Autocorrelación

Una serie temporal $X_t : t = 0, \pm 1, ...$ se dice ser estacionaria si esta tiene propiedades estadísticas similares a aquellas series "desplazadas en el tiempo"$X_{t+h} : t = 0, \pm 1, ...$ para cada entero h. Con especial atención a aquellas propiedades que dependen sólo en el primer y segundo momento de X_t, podemos hacer que esta idea sea precisa con las siguientes definiciones:

Definición 3. *Sea* X_t *una serie temporal con* $E(X_t^2) < \infty$*. La función de la media de* X_t *es*

$$\mu_X(t) = E(X_t)$$

La función covarianza de X_t *es*

$$\gamma_X(r, s) = Cov(X_r, X_s) = E[(X_r - \mu_X(r))(X_s - \mu_X(s))]$$

para todo entero r *y* s*.*

Definición 4. X_t *es estacionaria débil si*
(i) $\mu_X(t)$ *es independiente de* t *y*
(ii) $\gamma_X(t + h, t)$ *es independiente de* t *para cada* h*.*

Una serie temporal X_t, $t = 0, \pm 1, ...$ es *estrictamente estacionaria* si es definido por la condición de que $(X_1, ..., X_n)$ y $(X_{1+h}, ..., X_{n+h}$ tiene la misma función de distribución conjunta para todo los enteros h y $n > 0$. Es sencillo comprobar que si X_t es estrictamente estacionaria y $E(X_t^2) < \infty$ para todo t, entonces X_t es también estacionaria débil. Siempre que usemos el término estacionaria nos refreiremos a débilmente estacionaria como en la Definición 3, a menos que se indique lo contrario.[2]

Definición 5. X_t *una serie temporal estrictamente estacionaria si*

$$(X_1, ..., X_n)' =^d (X_{1+h}, ..., X_{n+h})'$$

para todo entero h *y* $n \geq 1$*. (Aquí* $=^d$ *es usado para indicar que dos vectores aleatorias tienen la misma función de distribución).*

Para referencia, mostraremos una de las propiedades elementales de las series estrictamente estacionarias.

Propiedades de una serie estrictamente estacionaria

1. Las variables aleatorias X_t son identicamente distribuidas.
2. $(X_t, X_{t+h})' =^d (X_1, ..., X_{1+h})'$ para todo entero t y h.
3. $\{X_t\}$ es débilmente estacionaria si $E(X_t^2) < \infty$ para todo t.
4. Una serie débilmente estacionaria no implica una serie estrictamente estacionaria.
5. Una sucesión iid es estrictamente estacionaria.

Definición 6. *Sea X_t una serie temporal estacionaria. La Función de Autocovarianza (ACVF por sus siglas en inglés) de X_t en el retardo h es:*

$$\gamma_X(h) = Cov(X_{t+h}, X_t)$$

La Función de Autocorrelación (ACF) de X_t en el retardo h es

$$\rho_X(h) \equiv \frac{\gamma_X(h)}{\gamma_X(0)} = Cor(X_{t+h}, X_t)$$

2.2.2. Ruido Blanco, Camino Aleatorio

Ruido Blanco

Dentro de los procesos estocásticos se encuentra un caso simple llamado ruido blanco, donde los valores son independientes e idénticamente distribuidos a lo largo del tiempo con media cero e igual varianza, se notan como ε_t, esto es:

$$\varepsilon_t \sim N(0, \sigma^2), cov(\varepsilon_{t_i}, \varepsilon_{t_j}) = 0, \quad para\ todo\ t\ distintodet_j$$

Camino Aleatorio

Se define camino aleatorio como un proceso estocástico X_t, donde $X_t = X_{t-1} + \varepsilon_t$ es decir $\bigtriangledown X_t = \varepsilon_t$, siendo este resultado un ruido blanco.

2.3. Procesos Autorregresivos y de Media Móvil

Existen modelos que tratan de estructuras estocásticas lineales y su asociación con una serie temporal de datos. Usualmente este tipo de procesos se presentan como combinación lineal de variables aleatorias. Si estos procesos siguen una distribución normal con media cero se presenta la serie como combinación lineal de valores anteriores infinitos de la misma serie más un ruido blanco.

2.3.1. Procesos Autorregresivos y de Media Móvil

Para la correcta identificación de estos procesos es necesario conocer el teorema de descomposición de Wold y la siguiente definición de un proceso estocástico lineal:

Teorema 1. ***(Teorema de descomposión de Wold)***. *Cualquier proceso estacionario*

$$X_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j} + k_t$$

en donde $\varepsilon_t \sim RB(0, \sigma^2)$, k_t es una función determinísta y $\sum_{j=0}^{\infty} \psi_j^2 < \infty$

Definición 7. ***(Proceso estocástico lineal)***. *Un proceso estocástico es un proceso lineal si para todo* $t = 0, \pm 1, ...$ *puede ser representado por:*

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j \varepsilon_{j-t}$$

Donde $\varepsilon_t, t \epsilon T$ *es un proceso de ruido blanco y* $\{\psi_j\}_{j=-\infty}^{\infty}$ *una sucesión de constantes reales absolutamente sumable, es decir, verifica* $\sum_{-\infty}^{\infty} \mid \psi_j \mid < \infty$

Para la interpretación de estos procesos primero se define el operador de retardos así:

Definición 8. ***(Operador de retardos)***. *El operador de retardo de una función del tiempo en un instante proporciona la función en el instante anterior.*

$$BX_t = X_{t-1}$$

Y presenta las siguientes propiedades:
1. $Ba = a$*, con a constante.*
2. $B(aX_t) = aBX_t = aX_{t-1}$
3. $B(aX_t + bY_t) = aBX_t + bBY_t$
4. $B^k X_t = X_{t-k}$*, operador de retardos de orden k.*

Proceso Autorregresivo AR(p)

El primer proceso a trabajar son los modelos autorregresivos de orden p conocidos como $AR(p)$, estos modelos parten del supuesto de que el valor presente de la serie X_t se explica en función de p valores previos así $X_{t-1}, X_{t-2}, ..., X_{t-p}$, siendo p el número de retardos necesarios para pronosticar $\hat{X}_t$.[1]

El proceso general de $AR(p)$ se puede modelizar bajo la siguiente ecuación:

$$X_t = c + \phi_1 X_{t-1} + ... + \phi_p X_{t-p} + \varepsilon_t$$

Donde $X_t, X_{t-1}, ...$ representa las variables aleatorias concebidas como realizaciones de un proceso estocástico en los momentos de tiempo $t, t-1$ los cuales se caracterizan por $E(X_t = E(X_{t-1})) = ...c, \phi_1, ..., \phi_p$ y la varianza del proceso σ_t^2 son los parámetros que definen el modelo y que deben ser estimados. Reescribiendo el proceso general de $AR(p)$ en términos del operador de retardos se tiene:

$$(1 - \phi_1 B - ... - \phi_p B^p) X_t = \varepsilon_t$$

$$\phi_p X_t = \varepsilon_t$$

Proceso de medias móviles MA(q)

Una serie temporal X_t de medias móviles de orden $MA(q)$ se representa mediante la ecuación:

$$X_t = \mu + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - ... - \theta_q \varepsilon_{t-q}$$

Donde:

X_t es una variable aleatoria concebida como realización de un modelo estocástico en los momentos de tiempo t, presentando la característica de $E(Y_t) = E(Y_{t-1}) = ... = \infty$.

$\mu, \theta_1, ..., \theta_q$ y σ_t^2, representan los parámetros del modelo a estimar.

ε_t representa la variable aleatoria ruido blanco.

La característica fundamental de estos modelos suponen que el valor presente de la serie X_t vienen determinados por una fuente externo. La representación de este modelo en términos del operador de retardos viene dada así:

$$X_t - \mu = (1 - \theta_1 B - \theta_2 B^2 - ... - \theta_q B^q)\varepsilon_t$$

$$X_t - \mu = \theta_q(B)\varepsilon_t$$

Nota: El problema de raíces unitarias en series temporales ocurre cuando tanto el polinomio autorregresivo o el polinomio de media móvil de un modelo $ARMA$ tiene una raíz en o sobre un círculo unitario. Una raíz unitaria en cualquiera de estos polinomios tiene implicaciones importante para el modelado.

Por ejemplo, una raíz del polinomio autorregresivo cerca de 1 sugiere que los datos deben ser diferenciados antes de ajustar un modelo $ARMA$, mientras que una raíz del polinomio de medias móviles cerca de 1 indica que los datos fueron sobre diferenciados.[1]

2.3.2. Modelos ARMA, ARIMA y SARIMA

Una serie de tiempo X_t, que presente las características AR y MA de manera conjunta, seguirá un proceso $ARMA(p, q)$, con p términos autorregresivos y q términos de media móvil, estos modelos permiten aproximar la estructura de covarianza de un proceso estacionario hasta el nivel que se fije previamente. Los modelos ARIMA son una extensión de los que se utiliza para modelizar algunos procesos no estacionarios y los $SARIMA$ son la generalización de los modelos $ARIMA$ incluyendo una componente estacional.[1]

Definición 9. ***(Modelos ARMA)****. Se dice que la serie estacionaria X_t tiene estructura $ARMA(p, q)$ si admite una representación del tipo:*

$$X_t - \mu = \phi_1(X_{t-1} - \mu) + ... + \phi_p(X_{t-p} - \mu) + \varepsilon_t + \theta_1\varepsilon_{t-1} + ... + \theta_q\varepsilon_{t-q}$$

La definición anterior la podemos escribir por medio del operador de retardos así:

$$BX_t = X_{t-1}$$

Definiendo los polinomios:

$$\Phi(x) = 1 - \phi_1 x - ... - \phi_p x^p$$

$$\Theta(x) = 1 + \theta_1 x - ... - \theta_q x^q$$

El modelo $ARMA(p, q)$ toma la forma:

$$\Phi(B)(X_t - \mu) = \Theta(B)\varepsilon_t$$

Definición 10. ***(Modelos ARIMA)****. Se dice que un proceso X_t tiene estructura $ARIMA(p, q, d)$ si existen dos polinomios $\Phi(x)$ y $\Theta(x)$ de grado p y q, respectivamente, verificando que*

$$\Phi(B)(1 - B)^d(X_t - \mu) = \Theta(B)\varepsilon_t$$

Diferenciar la serie X_t en el retardo s en una manera muy conveniente de eliminar la componente de periodo s. Si ajustamos un modelo $ARMA(p,q)$, $\phi(B)Y_t = \theta(B)\varepsilon_t$, a la serie diferenciada $Y_t = (1-B^s)X_t$, entonces el modelo para la serie original es $\phi(B)(1-B^s)X_t = \theta(B)\varepsilon_t$. Este es un caso particular del modelo general $ARIMA$ estacional ($SARIMA$) definido como sigue:

Definición 11. ***(Modelos SARIMA)***. *Si d y D son enteros no-negativos, entonces $\{X_t\}$ es un proceso $ARIMA(p,d,q)\times(P,Q,D)$ con periodo s si la serie diferenciada $Y_t = (1-B)^d(1-B^s)^D X_t$ es un proceso $ARMA$ definido por*

$$\phi(B)\Phi(B^s)Y_t = \theta(B)\Theta(B^s)\varepsilon_t$$

Donde ε_t sigue un proceso de ruido blanco.

3. Metodología Box - Jenkins

En este capítulo se estudiará la metodología de Box - Jenkins la cual fue desarrollada en los años 70 por George Box y Gwilym Jenkins, en la actualidad esta metodología ha adquirido gran relevancia por la aplicabilidad que ha tenido gracias al desarrollo de los sistemas de computación. La idea principal del enfoque Box - Jenkins es proponer un conjunto de procedimientos para escoger entre los modelos $ARMA, ARIMA$ y $SARIMA$ o $ARIMA$ estacional, y el que se ajuste a los datos de una serie temporal observada y con ello realizar pronósticos sobre ésta. Este procedimiento plantea cuatro pasos a saber:

1. Identificación: Este primer paso busca establecer los valores apropiados para p, d, q
2. Estimación: Luego se deben estimar los parámetros incluidos en el modelo.
3. Verificación de diagnóstico: Al seleccionar el modelo particular se debe comprobar si el modelo se ajusta a los datos (puede existir otro modelo que presente mejor ajuste). La prueba más simple de ajuste es comprobar si los residuos obtenidos son ruido blanco.
4. Predicción: El último plazo consiste en la elaboración de los pronósticos de la serie temporal en particular.

Como desarrollo de la temática para la presente sección se procede a explicar, de forma más precisa, cada uno de los pasos que comprende Box - Jenkins.

3.1. Identificación

Este paso comprende identificar la estructura no estacionaria, realizar las transformaciones que se requieran para obtener varianza y medias estables y finalmente determinar las órdenes del modelo p, d, q.

Este análisis nos permite identificar en la serie temporal características como la alta frecuencia (característica intrínseca de la serie que no es corregible), el comportamiento no estacionario y la presencia de estacionalidad en los datos.[1]

3.1.1. Estabilización de la no estacionariedad

Para la estabilización de la no estacionariedad es posible realizar transformaciones de Box-Cox, diferenciaciones, entre otras.

Transformaciones de Box-Cox: Para una serie temporal X_t , el proceso que se obtiene luego de realizar una transformación Box-Cox de parámetro λ se encuentra definido por:

$$X_t^{(\lambda)} = \begin{cases} \text{si } \lambda \neq 0, & \frac{(X_t^{(\lambda)}-1)}{\lambda} \\ \text{si } \lambda = 0, & \log(X_t + C) \end{cases} \quad (1)$$

con $t = 0, \pm 1, ...$

Box-Cox además incluye una familia infinita de funciones como logaritmos, raíz cuadrada, etc. También se utiliza esta transformación para solucionar problema de normalidad de los datos.

Diferenciación: este procedimiento implica identificar si la serie temporal tiene un centro de gravedad o si carece de éste, es decir, si se presentan tendencias (para lo cual se usará, sobre todo, el grafico de la serie) y además se busca identificar en la función de autocorrelación muestral un decaimiento lento.

Para estabilizar la media se toman diferenciaciones del tipo:

$$\bigtriangledown_s X_t = (1 - (B)^s)X_t = X_t - X_{t-s},\ t = s+1, ..., T$$

Los instrumentos gráficos son muy útiles, se utiliza el gráfico de la serie, el gráfico de autocorrelación y el de autocorrelación parcial, con ellos se puede realizar una diferenciación regular d o una diferenciación en la parte estacional D.

En la diferenciación en d el gráfico presentará una tendencia clara, la función de autocorrelación muestral decaerá de forma lenta y lineal y la función de autocorrelación parcial muestral presentará un coeficiente de primer retardo cercano a 1. Se debe tener presente que las tendencias lineales se eliminan con $d = 1$ y las tendencias cuadráticas con $d = 2$. Para la diferenciación en D se mostrará un gráfico con pautas repetidas de periodo s y se observara que la función de autocorrelación simple muestral mostrará coeficientes altos que decrecen de manera lenta en los retardos múltiplos de periodo s.[1]

Nota: La diferenciación, a veces, también estabiliza la varianza de la serie X_t.

3.1.2. Identificación de órdenes del proceso

Se proceden a estimar los órdenes p, q, P y Q , para ello se compara las funciones estimadas de autocorrelación simple y parcial con sus respectivas funciones teóricas, se debe seleccionar un conjunto de modelos que se supongan adecuados.

Los coeficientes de autocorrelación muéstrales se estiman mediante la ecuación:

$$\hat{\rho}_k = \frac{\sum_{t=d+sD+1}^{T-k}(\omega_t - \bar{\omega})(\omega_{t+k} - \hat{\omega})}{\sum_{t=d+sD+1}^{T}(\omega_t - \bar{\omega})^2}$$

con $k = 1, 2, ...$ y donde $\omega_t = \bigtriangledown^d \bigtriangledown_s^D X_t$, que representa la serie estacionaria. Para la obtención de esta fórmula se debe primero calcular la covarianza muestral del retardo k, $\hat{\Omega}_k$ y la varianza muestral $\hat{\Omega}_0$, definidas como:

$$\hat{\Omega}_k = \frac{\sum(\omega_t - \omega)(\omega_{t+k} - \bar{\omega})}{n}$$

$$\hat{\Omega}_0 = \frac{\sum(\omega_t - \bar{\omega})^2}{n}$$

Donde $\bar{\omega}$ es la media muestral y n el tamaño de la muestra, por tanto $\hat{\rho}_k = \frac{\hat{\Omega}_k}{\hat{\Omega}_0}$. Para identificar los órdenes del modelo es necesario apoyarse en la función de autocorrelación parcial ya que p_k (coeficiente de correlación parcial de orden k) mide el grado de asociación lineal existente entre variables habiendo

ajustado el efecto lineal de todas las variables intermedias, el cálculo se realiza mediante la regresión lineal entre variables, que representada seria

$$X_t = \mu + p_1 X_{t-1} + ... + p_k X_{t-k} + \varepsilon_t$$

La función de autocorrelación parcial se estima basándose en los datos de la serie y como función de $\hat{\rho_k}$. Es claro que si $\hat{\rho_k} = 0$, la serie X_t es ruido blanco, si por el contrario , la serie X_t no es ruido blanco, por lo anterior se emplea una prueba de significancia conjunta que determina si los coeficientes estimados estadísticamente equivalen a cero, pruebas soportadas por los test de Q de Box y Pierce y por la prueba Ljung- Box.

Para identificar órdenes en un proceso estacional se sigue un proceso similar con la diferencia que se deberán observar los coeficientes en los retardos específicos que muestren estacionalidad ya que indicarán en la función de autocorrelación y la función de autocorrelación simple los órdenes P y Q.

3.2. Estimación

En este paso se realiza la estimación de los parámetros que constituyen el modelo $AR, MA, ARMA$ y $ARIMA$ (si es el caso), esta estimación se obtiene por diferentes métodos a saber:

3.2.1. Método de los mínimos cuadrados condicionales

Con esta técnica se busca minimizar la suma de cuadrados condicionales $S_c(\varphi_p, \theta_q) = \sum_{t=p+l}^{n} \varepsilon_t^2$ suma que se obtiene de la suma de cuadrados no condicionales, donde $\varepsilon_p, \varepsilon_{p-1}, ..., \varepsilon_{p+1-q}$ son la base de los cálculos de ε_t igualados a cero que corresponde a su valor esperado.

Para un modelo $AR(p)$ la estimación viene dada por al siguiente ecuación:

$$S_c(\varphi_p) = \sum_{t=p+1}^{n} (X_t - \phi_1 X_{t-1} - ... - \phi_1 X_{t-p})^2$$

De donde se plantean las ecuaciones normales $Y'Y\hat{\varphi_p} = Y'W$ siendo:

$$\begin{bmatrix} X_p & X_{p-1} & \cdots & X_1 \\ X_{p+1} & X_p & \cdots & X_2 \\ \vdots & \vdots & \ddots & \vdots \\ X_{n-1} & X_{n-2} & \cdots & X_{n-p} \end{bmatrix} \tag{2}$$

$\hat{\gamma_p} = (\hat{\gamma_1}, \hat{...}, \hat{})$ y $W = (X_{p+1}, ..., X_n)'$ y por tanto $\hat{\varphi_p} = (Y'Y)^{-1}Y'W$.

3.2.2. Algoritmo de máxima verosimilitud

Con este método se busca que los estimadores de los parámetros maximicen la función de verosimilitud con respecto a la varianza del error. Por ello el algoritmo se basa en la función de verosimilitud en modelos $ARMA$ planteada por Newbold en 1974:

$$L((\varphi_p, \theta_q, \sigma^2) \mid X) = g_1(\varphi_p, \theta_q, \sigma^2) \exp\{\frac{-1}{2\sigma^2} S(\varphi_p, \theta_q)\}$$

Donde g_1 representa la función dependiente de los parámetros $\varphi_p, \theta_q, \sigma^2$ y

$$S(\varphi_p, \theta_q) = \sum_{t=1-p-q}^{n} E^2[U_t \mid X, \varphi_p, \theta_q, \sigma^2]$$

La utilización de representaciones de los procesos $ARMA$ como modelos en el espacio de estados y del filtro de Kalman permite calcular de manera exacta la función de verosimilitud. Al maximizar la función de verosimilitud se logran obtener las predicciones.

3.3. Diagnóstico del modelo

En este paso se busca verificar que tan adecuado es el modelo, es decir, se debe comprobar que:

- Las condiciones de estacionariedad e invertibilidad de los coeficientes estimados del modelo se cumplan, así como determinar que estos parámetros estimados sean significativos.
- Los residuos se comporten como ruido blanco.

3.3.1. Diagnóstico de los coeficientes estimados

Para los modelos ARMA(p,q) se plantean los siguientes contrastes de hipótesis:

$$H_0 : \delta = 0 \; vs \; H_a : \delta \neq 0$$

$$H_0 : \phi = 0 \; vs \; H_a : \phi \neq 0$$

$$H_0 : \theta = 0 \; vs \; H_a : \theta \neq 0$$

Siendo δ la constante para la media. Para los coeficientes $\beta = (\delta, \phi_1, ..., \phi_p, \theta_1, ..., \theta_q)$ los cuales presentan distribución asintótica $\hat{\beta}_i \sim N(\beta_i, V(\hat{\beta}_i))$ esto para todo i donde la inversa de la matriz de información permite estimar la varianza, el estadístico t de contraste con distribución normal viene dado por:

$$t = \frac{\hat{\beta}_i - 0}{\sqrt{(V)(\hat{\beta}_i)}} \sim N$$

La hipótesis nula con $\alpha = 0{,}05$ se rechaza cuando : $| \frac{\hat{\beta}_i - 0}{\sqrt{(V)(\hat{\beta}_i)}} | > \frac{N_\alpha(0,1)}{2} \approx 2$

Las condiciones de estacionariedad e invertibilidad se comprueban calculando las raíces del polinomio autorregresivo, $\phi(\hat{B}) = 0$ y las raíces del polinomio de medias móviles $\theta(\hat{B}) = 0$, si alguna se encuentra cercana a 1 se puede presumir falta de estacionariedad o invertibilidad. Por otra parte la matriz de covarianzas permite detectar presencia de factores comunes al modelo valiéndose de los niveles de correlación entre los modelos.[2]

3.3.2. Diagnóstico de los residuos el modelo

Para el modelo $ARMA(p, q)$ se debe comprobar que los residuos presentan un comportamiento de ruido blanco, media cero, varianza constante y autocorrelaciones nulas.

Para el contraste de media cero se plantean las hipótesis

$$H_0 : E(\varepsilon_t) = 0 \; vs \; H_0 : E(\varepsilon_t) \neq 0$$

El estadístico t viene dado por:

$$t = \sqrt{(T)} \frac{\bar{\hat{a}}}{\sqrt{(C_0(\hat{a}))}} \sim N(0, 1)$$

Donde $\bar{\hat{a}}$ y $C_0(\hat{a})$ representan la media y la varianza muestral de los residuos estimados. También se puede usar el análisis grafico de los residuos que junto con el de dispersión nos podrá indicar la existencia de varianza constante.

Para determinar la no existencia de correlaciones entre los residuos se utiliza el estadístico de Ljung-Box el cual viene dado por:

$$Q = T(T+2)\sum_{h=1}^{m}\frac{\hat{\rho_h}}{T-r}$$

Siendo ρ_h el coeficiente de autocorrelación de los residuos estimados, T representa el número de valores de la serie X_t y r representa el número de parámetros estimados. El estadístico Q se distribuye como una Chi-cuadrado, donde el número de grados de libertad es igual a $m-r-1$, siendo m el número de coeficientes utilizados en la suma.

Las hipótesis a probar con este test son:

$$H_0 : \rho_1, \rho_2 = ... = \rho_k = 0 \;\; vs \;\; H_i : \;\; algun \;\; \rho_i \neq 0$$

Gráficamente se pueden observar los coeficientes de las funciones de autocorrelación muestrales (simple y parcial) que no deberán ser significativos para considerar la independencia entre residuos (comparados con las bandas de confianza).

Nota: Alguno puede ser significativo debido al azar.

Si los residuos presentan comportamiento de ruido blanco se procede a calcular las predicciones, de no cumplirse esto se debe repetir el proceso y proponer un nuevo modelo en la fase de identificación.

Como conclusión se puede decir que en esta etapa se selecciona la mejor especificación del modelo para realizar el pronóstico, el cual se estudia a continuación.

3.4. Predicción

Una vez que se cuenta con un modelo estimado que cumpla los criterios de validez anteriores, este puede ser utilizado para realizar pronósticos [3] para instantes observados la predicción a realizar es del tipo X_{n+1} y viene dada a partir de la ecuación de diferencias:

$$X_{n+1} = \rho_1 X_{n+l-1} + ... + \rho_{p+q} X_{n+l-p-q} - \theta_1 \varepsilon_{n+l-1} - ... - \theta_q \varepsilon_{n+l-q} + \varepsilon_{n+l}$$

Como suma infinita ponderada de los valores $\varepsilon_r, r \leq t$:

$$X_{n+1} = \sum_{j=0}^{\infty} \psi_j \varepsilon_{n+l-j}, \;\; (\psi_0 = 1)$$

Como suma infinita ponderada de los valores previos más un ruido

$$X_{n+1} = \varepsilon_{n+1} + \sum_{j=1}^{\infty} \pi_j X_{n+l-j}$$

Si la predicción de la observación X_{n+1} se denomina predicción de l pasos hacia el futuro y se representa por y que se expresa como combinaciones lineales de valores pasados y presentes, siendo además función de los valores pasados y presentes del proceso ruido.

$$X_n(l) = \sum_{j=1}^{\infty} \psi_j^* \varepsilon_{n+l-j}$$

Donde ψ_j^* representa los pesos que minimizan el error cuadrático medio de la predicción, luego el error de la predicción se expresa como:

$$e_n(l) = X_{n+1} - X_n(l) = (\varepsilon_{n+l} + \psi_1 \varepsilon_{n+l-1} + ... + \psi_{l-1}\varepsilon_{n+1})$$

La varianza de la predicción se representa por:

$$Var[e_n(l)] = (1 + \psi_1^2 + ... + \psi_{l-1}^2)\sigma^2$$

Asumiendo que el ruido blanco sea Gaussiano se tiene el siguiente límite de confianza para un $\alpha = 0{,}05$:

$$X_n(l) \pm 1{,}96\sqrt{((\sigma^2(1 + \psi_1^2 + ... + \psi_{l-1}^2)))}$$

4. Datos de rango

Asumiendo que los parámetros de t artículos deben ser estimados. La preferencia de un panel de n consumidores son usados como observaciones. A cada consumidor se le pide ordenar los t artículos, empezando por el que más le gusta y terminando con el que menos le gusta.[7]

Se asume que las preferencias de los consumidores tiene la misma escala de preferencia, así cada observación puede ser tratado como salidas independientes de la misma distribución.

Se debe tener claro la diferencia entre el concepto de ordenamiento y jerarquización. Una jerarquización de t artículos, es definido como

$$r_u = (r_{1u}, r_{2u}, ..., r_{iu}), \qquad \forall u = 1, ..., t!$$

Donde, r_{iu} es el valor del rango escogido para el artículo i en el número jerarquizado u, mientras que el *ordenamiento* de los artículos es definido como:

$$h_u = (h_{1u}, h_{2u}, ..., h_{iu}), \qquad \forall u = 1, ..., t!$$

Donde h_{iu} es el número del artículo con valor rango i en el número jerarquizado u.

La relación entre una jerarquización r_u y un ordenamiento h_u es único y por lo tanto los datos pueden ser observados en ambas formas.

Todos las posibles jerarquizaciones de los t artículos pueden ser descritos por todas las posibles permutaciones de los índices de los artículos. Esto es, $t!$ diferentes jerarquizaciones de los t artículos.

Para derivar un modelo matemático que estime los parámetros de un artículo a partir de un conjunto de observaciones jerarquizadas, la variable estocástica Y_{uk} de las observaciones jerarquizadas es definido como:

$$Y_{uk} = \begin{cases} 1, & \text{si el consumidor k jerarquiza los artículos de acuerdo a la jerarquización} r_u \\ 0, & \text{en otro caso.} \end{cases}$$

para todo $u = 1, ..., t!$ y $k = 1, ..., n$.

Escrito en otra forma

$$Y_{uk} \sim bin(p_u, 1) \qquad \forall u = 1, ..., t!,$$

Donde $p_u = P(Y_{uk} = 1)$.

Como un consumidor debe elegir uno y sólo una jerarquización para todos los otros, la probabilidad p_u, para $u = 1, ..., t!$ debe sumar uno,

$$\sum_{u=1}^{t!} P(R_u) = 1,$$

Donde R_u es usado como una notación para el evento que un consumidor jerarquiza los artículos de acuerdo a la jerarquización r_u, $\{R_u\} = \{Y_{uk} = 1\}$.

Hasta ahora, todos los modelos de jerarquización deben coincidir, sin embargo se han desarrolado durante años distintas aproximaciones para describir la probabilidad p_u.[7]

5. Clasificación

5.1. Regresión Logística

Los métodos de Regresión se han convertido en un componente integral para cualquier análisis de datos que involucre la descripción de la relación entre la variable respuesta y una o más variables explicatorias. Es frecuente en el caso en que la variable de salida es discreta, tomando dos o más posibles valores. En las últimas décadas el modelo de regresión logística se ha convertido, en muchos campos, el método estándar para el análisis esta situación.[4]

El objetivo de un análisis usando este método es el mismo que el de cualquier técnica para la construcción de un modelo usado en estadística: encontrar el mejor ajuste, el más parsimonioso y más razonable para describir la relación entre una variable salida (dependiente o respuesta) y un conjunto de variables independientes (predictor o explicatoria). Estas variables indpendientes son llamadas a menudo covariables. El ejemplo más común de modelado es el usual modelo de regresión lineal donde la variable de salida es asumida como continua.

Lo que distingue un modelo de regresión logística de un modelo de regresión lineal es que la variable de salida en la regresión logistica es binaria o dicotómica. Esta diferencia entre la regresión logistica y lineal están ambos reflejados es la escogencia de un modelo paramétrico y sus supuestos. Una vez que esta diferencia es tomada en cuenta, los métodos en un análisis usando regresión logística son los mismos principios generales usados en la regresión lineal. Así, las técnicas usadas en el análisis de regresión lineal motivará al acercamiento a una regresión logística.

Probabilidades

Las probabilidades de un evento son la relación entre la probabilidad de que ocurra un evento y la probabilidad de que no ocurra. Si la probabilidad de que un evento ocurra es p, la probabilidad de que el evento no ocurra es $(1-p)$. Entonces las probabilidades correspondientes son un valor dado por

$$Probabilidad\{Evento\} = \frac{p}{1-p}$$

Como la regresión logística calcula la probabilidad de que un evento ocurra sobre la probabilidad de que un evento no ocurra, el impacto de las variables independientes generalmente se explica en términos de probabilidades. Con la regresión logística, la media de la variable de respuesta p en términos de una variable explicativa x se relaciona con p y x a través de la ecuación $p = \alpha + \beta x$.

Desafortunadamente, esto no es un buen modelo porque los valores extremos de x darán valores de $p = \alpha + \beta x$ que no se encuentran entre 0 y 1. La solución de regresión logística a este problema consiste en transformar las probabilidades usando el logaritmo natural. Con la regresión logística modelamos las probabilidades naturales del registro como una función lineal de la variable explicativa:

$$logit(y) = \ln(odds) = \ln(\frac{p}{1-p}) = \alpha + \beta x \qquad (1)$$

Donde p es la probabilidad del resultado interesado y x es la variable explicativa. Los parámetros de la regresión logística son α y β. Este es el modelo logístico simple. Tomando el antilogaritmo de la ecuación (1) en ambos lados, se puede derivar una ecuación para la predicción de la probabilidad de que ocurra un resultado interesado como

$$p = P(Y = \text{resultado esperado}|X = x, \text{un valor especifico}) = \frac{\epsilon^{\alpha+\beta x}}{1-\epsilon^{\alpha+\beta x}}$$

Extendiendo la lógica de la regresión logística simple a múltiples predictores, uno puede construir una regresión logística compleja como:

$$logit(y) = \ln(odds) = \ln(\frac{p}{1-p}) = \alpha + \beta_1 x_1 + ... + \beta_k x_k$$

Por lo tanto,

$$p = P(Y = \text{resultado esperado} | X_1 = x_1, ...X_k = x_k) = \frac{\epsilon^{\alpha+\beta x}}{1 - \epsilon^{\beta_1 x_1 + ... + \beta_k x_k}}$$

Razón de probabilidad

La razón de probabilidad, Odds Ratio en inglés (OR) es una medida comparativa de dos probabilidades con respecto a diferentes eventos. Para dos eventos A y B, las probabilidades correspondientes de A que ocurren con respecto a B son:

$$\text{Razón de probabilidad[A vs B]} = \frac{\text{razón de probabilidad[A]}}{\text{razón de probabilidad[B]}} = \frac{P_A/(1-P_A)}{P_B/(1-P_B)}$$

La razón de probabilidad es una medida de asociación entre una exposición y un resultado. La razón de probabilidad representa la probabilidad de que se produzca un resultado (por ejemplo, enfermedad o trastorno) dada una exposición particular (por ejemplo, comportamiento de salud, historial médico), en comparación con las probabilidades de que el resultado se produzca en ausencia de esa exposición.

Cuando se cálcula una regresión logística, el coeficiente de regresión (b_1) es el aumento estimado de las probabilidades registradas del resultado por aumento unitario en el valor de la variable independiente. En otras palabras, la función exponencial del coeficiente de regresión (ϵ^{b_1}) es la razón de probabilidad asociado con un aumento de unidad única en la variable independiente. La razón de probabilidad también puede usarse para determinar si una exposición en particular es un factor de riesgo para un resultado en particular, y para comparar la magnitud de varios factores de riesgo para ese resultado. $OR = 1$ indica que la exposición no afecta las probabilidades de resultado. $OR \leq 1$ indica la exposición asociada con mayores probabilidades de resultado. $OR \neq 1$ indica la exposición asociada con menores probabilidades de resultado. Por ejemplo, la variable fumar está codificada como 0 (no fumar) y 1 (fumar), y la razón de posibilidades para esta variable es $3,2$. Entonces, las probabilidades de un resultado positivo en los casos de fumar son $3,2$ veces más altas que en los casos de no fumadores. La regresión logística es una forma de generalizar el OR más allá de dos variables binarias. Supongamos que tenemos una variable de respuesta binaria Y y una variable de predicción binaria X, y además tenemos otras variables de predicción $Z_1, ..., Z_k$ que pueden ser o no binarias. Si usamos la regresión logística múltiple para retroceder Y en X, $Z_1, ..., Z_k$, entonces el coeficiente estimado βx para X se relaciona con un OR condicional. Específicamente, a nivel de la población:

$$\epsilon^{\hat{\beta_x}} = \frac{\frac{P(Y=1|X=1,Z_1,...,Z_k)}{P(Y=0|X=1,Z_1,...,Z_k)}}{\frac{P(Y=1|X=0,Z_1,...,Z_k)}{P(Y=0|X=0,Z_1,...,Z_k)}}$$

por lo tanto, $\epsilon^{\hat{\beta_x}}$ es una estimación de este razón de probabilidad condicional. La interpretación de $\epsilon^{\hat{\beta_x}}$ es como una estimación del OR entre Y y X cuando los valores de $Z_1, ..., Z_k$ se mantienen fijos.[7]

Curva Logística

La regresión logística es un método para ajustar una curva de regresión, $y = f(x)$, cuando y consiste en datos codificados en binario (0, 1 – fallo, éxito). Cuando la respuesta es una variable binaria (dicotómica) y x es numérica, la regresión logística ajusta una curva logística a la relación entre x e y. La curva logística es una curva en forma de S o sigmoidea, a menudo utilizada para modelar la población en crecimiento. Una curva logística comienza con un crecimiento lento y lineal, seguido por un crecimiento exponencial, que luego vuelve a disminuir a una tasa estable. Una función logística simple se define por la fórmula:

$$y = \frac{\epsilon^x}{1 + \epsilon^x} = \frac{1}{1 + \epsilon^{-x}}$$

La cual se muestra en la siguiente gráfica:

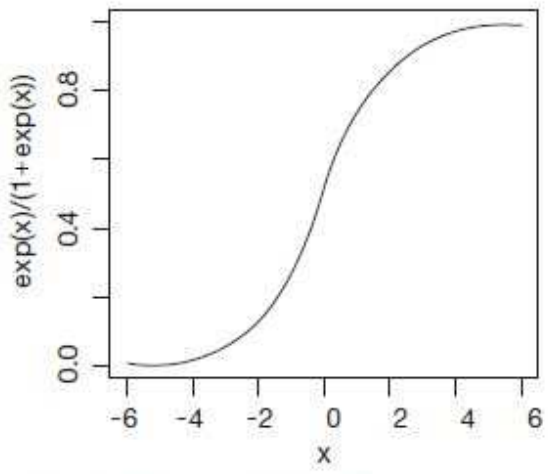


Figure 1. Graph of logistic curve where α = 0 and β = 1.

Para proporcionar flexibilidad, la función logística se puede extender a la forma

$$y = \frac{\epsilon^{\alpha+\beta x}}{1 + \epsilon^{\alpha+\beta x}} = \frac{1}{1 + \epsilon^{-(\alpha+\beta x)}}$$

La regresión logística se ajusta a α y β, los coeficientes de regresión. *Figura* 1 muestra la función logística cuando a α y β son 0 y 1, respectivamente. La función logística o función logit se usa para transformar una curva en forma de "S.[en] una línea recta y para cambiar el rango de la proporción de $0 - 1$ a $-$ínf $-+$ínf como:

$$logit(y) = ln(odds) = ln(\frac{p}{1-p}) = \alpha + \beta x$$

Donde p es la probabilidad del resultado interesado, α es el parámetro de intersección,β es un coeficiente de regresión, y x es un predictor.

Supuestos de Regresión Logística

La regresión logística no requiere muchas de las suposiciones principales de los modelos de regresión lineal que se basan en el método de los mínimos cuadrados ordinarios, particularmente con respecto a la linealidad de la relación entre las variables dependientes e independientes, la normalidad de la distribución de errores, la homocedasticidad de los errores, y el nivel de medición de las variables independientes. La regresión logística puede manejar relaciones no-lineales entre las variables dependientes e independientes, porque aplica una transformación logarítmica no-lineal. Los términos de error(los residuos) no necesitan ser multivariados distribuidos normalmente, aunque la normalidad multivariado produce una solución más estable. La varianza de los errores puede ser heteroscedástica para cada nivel de las variables independientes. La regresión logística puede manejar no solo datos continuos sino también datos discretos como variables independientes. Sin embargo, aún se aplican algunas otras. Primero, la regresión logística requiere que la variable dependiente sea discreta, principalmente dicotómica. Segundo, dado que la regresión logística estima la probabilidad de que ocurra el evento ($P(Y = 1)$), es necesario codificar la variable dependiente en consecuencia. Ese es el resultado deseado que debe codificarse para 1. En tercer lugar, el modelo debe ajustarse correctamente. No se debe sobre ajustar con variables sin sentido incluidas. Además, no debe estar equipado con una variable significativa no incluida. En cuarto lugar, la regresión logística requiere que cada observación sea independiente. Además, el modelo debería tener poca o ninguna multicolinealidad, es decir, las variables independientes no son funciones lineales entre sí. En quinto lugar, aunque la regresión logística no requiere una relación lineal entre las variables dependientes e independientes, requiere que las variables independientes estén linealmente relacionadas con las probabilidades de registro de un evento. Por último, la regresión logística requiere grandes tamaños de muestra porque las estimaciones de máxima verosimilitud son menos poderosas que los mínimos cuadrados ordinarios utilizados para estimar los parámetros desconocidos en un modelo de regresión lineal.[4]

5.1.1. Regresión Logística Ordinal

La Regresión logística ordinal es grandemente utilizados para modelar la relación entre un conjunto de predictores y una respuesta ordinal. Una respuesta ordinal tiene tres o más resultados que tienen un orden, como por ejemplo: bajo, medio y alto. Usted puede incluir términos de interacción y polinómicos, anidar términos en otros términos y ajustar diferentes funciones de enlace.

Cuando es necesario controlar posibles factores de confusión o incluso cuando es necesario tener en cuenta varios factores, la alternativa natural es un análisis multivariado especial para los datos ordinales. Existen varios enfoques, como el uso de modelos mixtos u otra clase de modelos, probit, por ejemplo, pero los modelos de regresión logística ordinal han sido ampliamente publicitados en la literatura estadística.

Consideraciones acerca de los datos para Regresión logística ordinal.

Para asegurar que los resultados arrojados por el modelo sean válidos, se consideran las siguientes pautas al recopilar datos, realizar el análisis e interpretar los resultados.

- Los predictores pueden ser continuos o categóricos.
- La variable de respuesta debe ser ordinal.
- La correlación entre los predictores, también conocida como multicolinealidad, no debe ser significativa.

Considere la variable de respuesta Y (por ejemplo, puntaje de calidad de vida) con k categorías codificadas en $1, 2, ..., k$ y $\vec{x} = (x_1, x_2, ., x_p)$ el vector de variables explicativas (covariables). Las k categorías de Y condicionalmente a los valores de las covariables ocurren con las probabilidades

$p_1, p_2, ..., p_k$, esto es$p_j = Pr(Y = j|\vec{x})j =$para $j = 1, 2, ...k$. El modelado de datos de respuesta ordinal puede usar probabilidades simples (p_j) o probabilidades acumuladas

$$(p_1 + p_2), (p_1 + p_2 + p_3), ..., (p_1 + p_2 + p_3 + ... + p_k)$$

En el primer caso, la probabilidad de cada categoría se compara con la probabilidad de una categoría de referencia, o cada categoría con la categoría anterior, como en el modelo de categorías adyacentes. Este trabajo presentará modelos logísticos con probabilidades acumuladas.

Razón de probabilidad para datos ordinales

Supongamos que la respuesta objetivo Y sobre la calidad de vida tiene k categorías ordenadas (Y_j con $j = 1, 2, ..., k$) y que se deben comparar dos grupos (AyB). Para la categoría j, OR viene dada por:

$$OR_j = \frac{\frac{P(Y \leq Y_j|x^{(A)})}{1-P(Y \leq Y_j|x^{(A)})}}{\frac{P(Y \leq Y_j|x^{(B)})}{1-P(Y \leq Y_j|x^{(B)})}} = \frac{\frac{P(Y \leq Y_j|x^{(A)})}{P(Y > Y_j|x^{(A)})}}{\frac{P(Y \leq Y_j|x^{(B)})}{P(Y > Y_j|x^{(B)})}}$$

De acuerdo con la definición habitual, OR es la relación entre dos probabilidades, pero ahora las probabilidades se definen en términos de probabilidades acumuladas. Para su interpretación, basta recordar que la respuesta se ha dicotomizado y que el evento debe clasificarse hasta la categoría j. Si A y B representan, respectivamente, la exposición y la no exposición a un factor de riesgo, O cuantifica las probabilidades de que un individuo en el grupo expuesto se clasifique en una categoría determinada, en comparación con las probabilidades del grupo no expuesto. En el contexto de datos ordinales, de acuerdo con la suposición de probabilidades proporcionales, OR es el mismo para todas las categorías de la variable de respuesta.[4]

Modelos logísticos acumulativos para variables repuestas ordinales

El modelo logístico acumulativo de probabilidades proporcionales es posiblemente el modelo más popular para los datos ordinales. Este modelo utiliza probabilidades acumulativas hasta un umbral, lo que hace que toda la gama de categorías ordinales sea binaria en ese umbral. Sea la variable respuesta $Y = 1, 2, 3, ..., J$ donde el orden es el natural. Las probabilidades asociadas son $\pi_1, \pi_2, ..., \pi_J$, y la probabilidad acumulada de respuesta menor o igual a j dado un conjunto de variables aleatorias X es:

$$P(Y \leq j|X) = \pi_1 + \pi_2 + ... + \pi_j$$

Con $X = X_1, X_2, ..., X_p$.Así, el modelo acumulativo logístico está definido como:

$$\log(\frac{P(Y \leq j|X)}{P(Y > j|X)}) = \log(\frac{P(Y \leq j|X)}{1 - P(Y \leq j|X)}) = \log(\frac{\pi_1 + \pi_2 + ... + \pi_j}{\pi_{j+1} + \pi_{j+2} + ... + \pi_J})$$

Esto describe la proporción logarítmica de dos probabilidades acumulativas, una menor y la otra más grande. Esto mide que tan probable la respuesta está en la categoría j o por debajo versus que la respuesta esté en una categoría mayor a j.

La serie Logística acumulativa puede ser definida como:

$$L_1 = \log(\frac{\pi_1}{\pi_2 + \pi_3 + ... + \pi_r})$$

$$L_2 = \log(\frac{\pi_1 + \pi_2}{\pi_3 + \pi_4 + ... + \pi_r})$$

$$L_2 = \log(\frac{\pi_1 + \pi_2}{\pi_3 + \pi_4 + ... + \pi_r})$$

$\vdots$

$$L_{r-1} = \log(\frac{\pi_1 + \pi_2 + ... + \pi_2}{\pi_r})$$

En esta notación L_j es la proporción logarítmica de caer sobre o por debajo la categoría j versus caer sobre ella. Ahora denotaremos el modelo en función de las covariables, como esto:

$$L_1 = \beta_{10} + \beta 11 X_1 + ... + \beta_{1p} X_p$$

$$L_2 = \beta_{20} + \beta_{21} X_1 + ... + \beta_{2p} X_p$$

$\vdots$

$$L_{r-1} = \beta_{r-1,0} + \beta r - 1, 1 X_1 + ... + \beta_{r-q,p} X_p$$

Note que (a diferencia del modelo logístico de categoría adyacente), esto no se trata de una reparametrización lineal del modelo de categoría de línea de base. Los registros acumulativos no son simples diferencias entre los registros de la categoría de base. Por lo tanto, el modelo anterior no dará un ajuste equivalente al del modelo de la categoría de base.

Ahora, supongamos que simplificamos el modelo requiriendo que el coeficiente de cada variable X sea idéntico a través de las ecuaciones $r-1$ logística. Luego, cambiando los nombres de las interceptaciones a α, se convierte en el modelo:

$$L_1 = \alpha_1 + \beta_1 X_1 + ... + \beta_p X_p$$

$$L_2 = \alpha_2 + \beta_1 X_1 + ... + \beta_p X_p$$

$\vdots$

$$L_{r-1} = \alpha_{r-1} + \beta_1 X_1 + ... + \beta_p X_p$$

Este modelo, llamado modelo logítico acumulativo de probabilidades proporcionales, tiene $(r-1)$ intercepciones más p pendientes, para un total de parámetros $r + p - 1$ a ser estimados.

Observa que las intercepciones pueden ser diferentes, pero que la pendiente de cada variable es la misma en diferentes ecuaciones. Uno puede pensar en esto como un conjunto de líneas paralelas (o hiperplanos) con diferentes intercepciones. La condición de probabilidades proporcionales obliga a que las líneas correspondientes a cada logaritmo acumulativo sean paralelas.

Interpretación

- En este modelo, el intercepto α_j es la proporción logarítmica de caer sobre o por debajo de la categoría j cuando $X_1 = X_2 = ... = 0$.
- Un sólo parámetro β_k describe el efecto de x_k en Y tal que β_k es el incremento en la proporción logarítmica cayendo sobre o por debajo de cualquier categoría asociado con un único incremento en X_k, manteniendo a las otras variables constantes.

- Pendiente constante β_k: el efecto de X_k es el mismo para todos los $J-1$ formas de colapsar Y en resultados dicotómicos.

5.2. Razonamiento Bayesiano

El razonamiento bayesiano proporciona un enfoque probabilístico para la inferencia. Se basa en la suposición de que las cantidades de interés se rigen por distribuciones de probabilidad y que las decisiones óptimas se pueden tomar razonando sobre estas probabilidades junto con datos observados. Es importante porque proporciona un enfoque cuantitativo para ponderar la evidencia que respalda hipótesis alternativas. El razonamiento bayesiano proporciona la base para algoritmos de aprendizaje que manipulan directamente las probabilidades, así como un marco para analizar el funcionamiento de otros algoritmos que no manipulan explícitamente las probabilidades.[5]

Los métodos de aprendizaje Bayesianos son relevantes por dos razones diferentes. En primer lugar, los algoritmos de aprendizaje Bayesiano que calculan probabilidades explícitas para hipótesis, como el clasificador ingenuo de Bayes, se encuentran entre los enfoques más prácticos para ciertos tipos de problemas de aprendizaje. Existen estudios que comparan el clasificador ingenuo de Bayes con otros algoritmos de aprendizaje, incluidos los algoritmos de árbol de decisión y de red neuronal. Estos muestran que el clasificador ingenuo de Bayes es competitivo con estos otros algoritmos de aprendizaje en muchos casos y que en algunos casos supera a estos otros métodos. Para tales tareas de aprendizaje, el clasificador ingenuo de Bayes se encuentra entre los algoritmos más efectivos conocidos.

La segunda razón por la que los métodos Bayesianos son importantes es que brindan una perspectiva útil para comprender muchos algoritmos de aprendizaje que no manipulan explícitamente las probabilidades.También usamos un análisis Bayesiano para justificar una elección de diseño clave en algoritmos de aprendizaje de redes neuronales eligiendo minimizar la suma de los errores al cuadrado al buscar espacio de posibles redes neuronales. También derivamos una función de error alternativa, la entropía cruzada, que es más apropiada que la suma de errores al cuadrado cuando se aprenden funciones objetivo que predicen probabilidades. Utilizamos una perspectiva Bayesiana para analizar el sesgo inductivo de los algoritmos de aprendizaje del árbol de decisiones que favorecen los árboles de decisión cortos y examinamos el principio de Longitud de Descripción Mínima estrechamente relacionado. Una familiaridad básica con los métodos bayesianos es importante para comprender y caracterizar el funcionamiento de muchos algoritmos en el aprendizaje automático.[6]

Las características de los métodos de aprendizaje Bayesianos incluyen:

- Cada ejemplo de entrenamiento observado puede disminuir o aumentar gradualmente la probabilidad estimada de que una hipótesis sea correcta. Esto proporciona un enfoque de aprendizaje más flexible que los algoritmos que eliminan por completo una hipótesis si se considera que es inconsistente con un solo ejemplo.

- El conocimiento previo puede combinarse con datos observados para determinar la probabilidad final de una hipótesis. En el aprendizaje Bayesiano, el conocimiento previo se proporciona afirmando (1) una probabilidad previa para cada hipótesis candidata, y (2) una distribución de probabilidad sobre los datos observados para cada hipótesis posible.

- Los métodos bayesianos pueden acomodar hipótesis que hacen predicciones probabilísticas.

- Las nuevas instancias se pueden clasificar combinando las predicciones de múltiples hipótesis, ponderadas por sus probabilidades.

- Incluso en los casos en que los métodos Bayesianos demuestran ser computacionalmente intratables, pueden proporcionar un estándar de toma de decisiones óptimo contra el cual se pueden medir otros métodos prácticos.

Una dificultad práctica en la aplicación de métodos Bayesianos es que típicamente requieren un conocimiento inicial de muchas probabilidades. Cuando estas probabilidades no se conocen de antemano, a menudo se estiman basándose en el conocimiento previo, los datos disponibles previamente

y las suposiciones sobre la forma de las distribuciones subyacentes. Una segunda dificultad práctica es el costo computacional significativo requerido para determinar la hipótesis óptima de Bayes en el caso general (lineal en el número de hipótesis del candidato). En ciertas situaciones especializadas, este costo computacional se puede reducir significativamente.[5]

Teorema de Bayes

A menudo nos interesa determinar la mejor hipótesis desde algún espacio H, dados los datos de entrenamiento observados D. Una forma de especificar lo que queremos decir con la mejor hipótesis, es decir, que exigimos la hipótesis más probable dados los datos D además de cualquier conocimiento inicial sobre las probabilidades previas de las diversas hipótesis en H. Bayes proporciona un método directo para calcular tales probabilidades. Más precisamente, el teorema de Bayes proporciona una forma de calcular la probabilidad de una hipótesis basada en su probabilidad previa, las probabilidades de observar varios datos dada la hipótesis y los datos observados.

Para definir precisamente el teorema de Bayes, primero introduzcamos una pequeña notación. Escribiremos $P(h)$ para denotar la probabilidad inicial que tiene la hipótesis h, antes de haber observado los datos de entrenamiento. $P(h)$ a menudo se denomina probabilidad previa de h y puede reflejar cualquier conocimiento previo que tengamos sobre la posibilidad de que h sea una hipótesis correcta. Si no tenemos ese conocimiento previo, entonces podríamos simplemente asignar la misma probabilidad previa a cada hipótesis candidata. De manera similar, escribiremos $P(D)$ para indicar la probabilidad previa de que se observen los datos de entrenamiento D (es decir, la probabilidad de que D no tenga conocimiento sobre qué hipótesis se cumple). A continuación, escribiremos $P(D|h)$ para denotar la probabilidad de observar los datos D dado un mundo en el que se cumple la hipótesis h. De manera más general, escribimos $P(x|y)$ para denotar la probabilidad de x dada y. Estamos interesados en la probabilidad $P(h|D)$ que h tiene dada la información de entrenamiento observada D. $P(h|D)$ se llama la probabilidad posterior de h, porque refleja nuestra confianza en que h tiene después de haber visto el datos de entrenamiento D. Observe que la probabilidad posterior $P(h|D)$ refleja la influencia de los datos de entrenamiento D, en contraste con la probabilidad previa $P(h)$, que es independiente de D.

El teorema de Bayes es la piedra angular de los métodos de aprendizaje bayesianos porque proporciona una forma de calcular la probabilidad posterior $P(h|D)$, a partir de la probabilidad previa $P(h)$, junto con $P(D)$ y $P(D|h)$.[5]

Expresión Teorema de Bayes

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

Como se podría esperar intuitivamente, $P(h|D)$ aumenta con $P(h)$ y con $P(D|h)$ según el teorema de Bayes. También es razonable ver que $P(h|D)$ disminuye a medida que $P(D)$ aumenta, porque cuanto más probable es que D se observe independientemente de h, menos evidencia D proporciona en apoyo de h. En muchos escenarios de aprendizaje, el alumno considera un conjunto de hipótesis candidatas H y está interesado en encontrar la hipótesis más probable $h \epsilon H$ dados los datos observados D (o al menos uno de los máximos probables si hay varios). Cualquiera de estas hipótesis de máxima probabilidad se denomina hipótesis de máximo a posteriori (MAP). Podemos determinar las hipótesis MAP usando el teorema de Bayes para calcular la probabilidad posterior de cada hipótesis candidata.[5]

Más precisamente, diremos que MAP es una hipótesis proporcionada

$$h_{MAP} \equiv argmax_{h \epsilon H} P(h|D) \equiv argmax_{h \epsilon H} \frac{P(D|h)P(h)}{P(D)} \equiv argmax_{h \epsilon H} P(D|h)P(h)$$

Observe que en el paso anterior hemos descartado el término $P(D)$ porque es una constante independiente de h. En algunos casos, supondremos que cada hipótesis en H es igualmente probable a priori ($P(h_i) = P(h_j)$ para todos los h_i y h_j en H). En este caso, podemos simplificar aún más la ecuación y sólo debemos considerar el término $P(D|h)$ para encontrar la hipótesis más probable. $P(D|h)$ a menudo se denomina la probabilidad de que los datos D dado h, y cualquier hipótesis que maximice $P(D|h)$ se denomina hipótesis de máxima verosimilitud (ML), h_{ML}.

$$h_{ML} \equiv argmax_{h \epsilon H} P(D|h)$$

5.2.1. Bayes Ingenuo

Un método de aprendizaje Bayesiano altamente práctico es el aprendiz ingenuo de Bayes, a menudo llamado Clasificador ingenuo de Bayes .En algunos dominios, se ha demostrado que su rendimiento es comparable al de la red neuronal y el aprendizaje del árbol de decisiones. Esta sección presenta el clasificador ingenuo Bayes.

El clasificador de Bayes ingenuo se aplica a tareas de aprendizaje donde cada instancia x se describe mediante una conjunción de valores de atributo y donde la función objetivo $f(x)$ puede tomar cualquier valor de un conjunto finito V. Un conjunto de ejemplos de entrenamiento de la función objetivo es proporcionado, y se presenta una nueva instancia, descrita por la tupla de valores de atributo $(a_1, a_2, ..., a_n)$. Al aprendiz se le pide que prediga el valor objetivo, o clasificación, para esta nueva instancia.

El enfoque bayesiano para clasificar la nueva instancia es asignar el valor objetivo más probable, v_{MAP}, dado los valores de atributos $(a_1, a_2, ..., a_n)$ que describen la instancia.

$$v_{MAP} = argmax_{v_j \epsilon V} P(v_j|a_1, a_2, ..., a_n)$$

Podemos usar el teorema de Bayes para reescribir esta expresión como:

$$v_{MAP} = argmax_{v_j \epsilon V} \frac{P(a_1, a_2, ..., a_n|v_j)P(v_j)}{P(a_2, ..., a_n)} = argmax_{v_j \epsilon V} P(a_1, a_2, ..., a_n|v_j)P(v_j)$$

Ahora podríamos intentar estimar los dos términos en la ecuación anterior en base a los datos de entrenamiento. Es fácil estimar cada uno de los $P(v_j)$ simplemente contando la frecuencia con la que cada valor objetivo v_j ocurre en los datos de entrenamiento. Sin embargo, estimar los diferentes términos $P(a_1, a_2, ..., a_n|v_j)$ de esta manera no es factible a menos que tengamos un conjunto muy grande de datos de entrenamiento. El problema es que el número de estos términos es igual al número de instancias posibles multiplicado por el número de valores objetivo posibles. Por lo tanto, necesitamos ver cada instancia en el espacio de la instancia muchas veces para obtener estimaciones confiables.

El clasificador ingenuo de Bayes se basa en la suposición simplificadora de que los valores de los atributos son condicionalmente independientes dado el valor objetivo. En otras palabras, la suposición es que dado el valor objetivo de la instancia, la probabilidad de observar la conjunción $a_1, a_2, ..., a_n$ es sólo el producto de las probabilidades para los atributos individuales: $P(a_1, a_2, ..., a_n|v_j) = \prod_i P(a_i|v_j)$. Sustituyendo esto en la ecuación anterior, tenemos el enfoque utilizado por el clasificador ingenuo de Bayes.[4]

Clasificador Ingenuo de Bayes

$$v_{NB} = argmax_{v_j \epsilon V} P(v_j) \prod_i P(a_i|v_j)$$

Donde v_{NB} denota el valor objetivo generado por el clasificador ingenuo de Bayes. Observe que en un clasificador de Bayes ingenuo, el número de términos $P(a_i|v_j)$ distintos que deben estimarse a partir de los datos de entrenamiento es simplemente el número de valores de atributos distintos multiplicado por el número de valores objetivo distintos, un número mucho más pequeño que si lo fuéramos a estimar los términos $P(a_1, a_2, ..., a_n|v_j)$ según lo contemplado en primer lugar.

En resumen, el método ingenuo de aprendizaje de Bayes implica un paso de aprendizaje en el que se estiman los diversos términos $P(v_j)$ y $P(a_i|v_j)$, en función de sus frecuencias sobre los datos de entrenamiento. El conjunto de estas estimaciones corresponde a la hipótesis aprendida. Esta hipótesis se usa luego para clasificar cada nueva instancia aplicando la regla de v_{NB}. Siempre que se cumpla la suposición ingenua de Bayes de independencia condicional, esta ingenua clasificación v_{NB} de Bayes es idéntica a la clasificación MAP. Una diferencia interesante entre el ingenuo método de aprendizaje de Bayes y otros métodos de aprendizaje es que no hay una búsqueda explícita en el espacio de posibles hipótesis (en este caso, el espacio de posibles hipótesis es el espacio de valores posibles que se pueden asignar a los diversos términos $P(v_j)$ y $P(a_i|v_j$. En cambio, la hipótesis se forma sin buscar, simplemente contando la frecuencia de varias combinaciones de datos dentro de los ejemplos de entrenamiento.[4]

6. Extracción y análisis general de los datos

El desarrollo de esta sección se enmarca en la etapa de extracción, preprocesamiento y análisis descriptivo de los datos para luego construir modelos de predicción de los mismos. La extracción y preparación de los datos para su posterior uso no es un proceso sencillo, y es una tarea de vital importancia ya que del resultado obtenido en esta etapa dependerá en gran medida la capacidad de clasificación y predicción de los modelos que se propongan.

6.1. Datos para la construcción de los modelos de predicción para las series temporales de ventas y transacciones

6.1.1. Extracción de los datos

Los datos de la cadena Central Madeirense están almacenados en bases de datos relacionales, bajo un sistema elaborado por la empresa Oracle llamado Oracle Retail Merchandising System. Entre las bases de datos que conforman este sistema está aquella que almacena información financiera llamada POSDBP, desde allí se extrajo las ventas en bolívares efectuadas entre enero 2012 hasta abril 2018 para cada unas de las 52 sucursales que conforman la empresa. Se complementaron dichos datos con los registros históricos de las transacciones obtenidas por cada sucursal en el mismo periodo de tiempo, estas extraídas desde la Data Warehouse (base de datos que consolida la información hitórica de la empresa).

La migración de los datos desde Oracle se hizo a través de un código en R en el cual se usaron las siguientes librerías : sqldf y ROracle. Una vez los datos estaban en un objeto de R estos fueron exportados a un archivo Excel donde se hicieron modificaciones a los datos, entre ellos:

- Recolocación de la información de las ventas y transacciones del mes de enero debido a que el año fiscal de la empresa termina en ese mes por lo que, por ejemplo, el mes de enero del 2016 lo consideran como último mes del año 2015 y no como el mes en que inicia el año 2016, por lo que se debía recolocar la secuencia de ese mes en todos los años para nuestros propósito.
- Consolidación de las ventas y el número de transacciones del mes de febrero de 2016 para cada una de las sucursales, debido que para ese mes hubo problemas con el almacenamiento de la información en la base de datos consolidada (Data Warehouse), por lo que esos valores se dividieron en dos filas, teniedo dos valores de ese mes en un mismo año.
- Se excluyó la información sobre la sucursal 12-Isabelica- Estado Carabobo debido que está clausurada por el saqueo ocurrido durante las manifestaciones de mayo del año 2017, por lo que esa sucursal no se tomará en cuenta para los modelos de predicción.

Luego de los cambios antes mencionados, los datos se volvieron a importar como un objeto de formato tabla a R para sus análisis posteriores.

6.1.2. Descripción general de los datos

Para nuestro estudio contamos con una base de datos compuesta por 5 variables de 66 registros cada una, las variables resgistran lo siguiente:

- UBICACIÓN : variable tipo factor que almacena la identificación de cada una de las 52 sucursales.
- AÑO: variable tipo factor que almacena la información del rango de tiempo en el que fueron extraidos los datos (enero 2012- abril 2018).

- MES: variable tipo factor que guarda los meses del año.
- VENTAS_REGULAR: variable del tipo numérico que almacena los registros de las ventas en bolívares por cada sucursal durante el periodo estudiado.
- NRO_TRANSACCIONES: variable de tipo numérico que almacena el número de transacciones realizadas por cada sucursal durante el periodo estudiado.

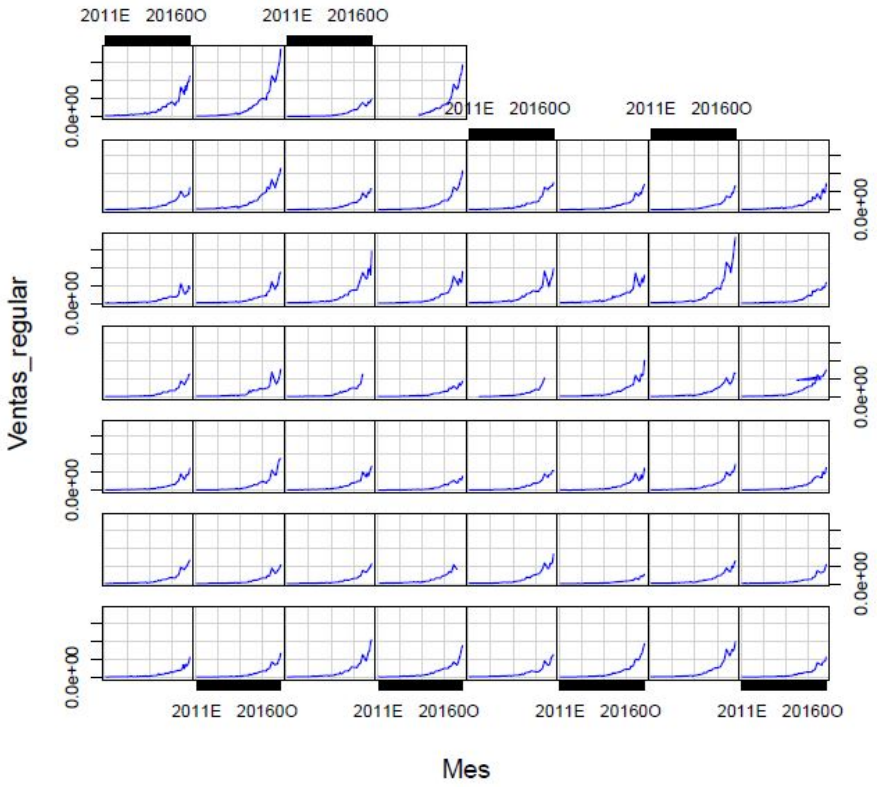
6.1.3. Análisis descriptivo de los datos

```
'data.frame':        3856 obs. of  5 variables:
 $ Ubicacion        : Factor w/ 52 levels "001 La Urbina 1",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Ano              : Factor w/ 7 levels "2012","2013",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Mes              : Factor w/ 86 levels "201112 ENERO",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ Ventas_regular   : num  11970989 9676116 11605383 10973011 12677781 ...
 $ Nro_transacciones: num  106696 112325 124134 135690 151457 ...$
```

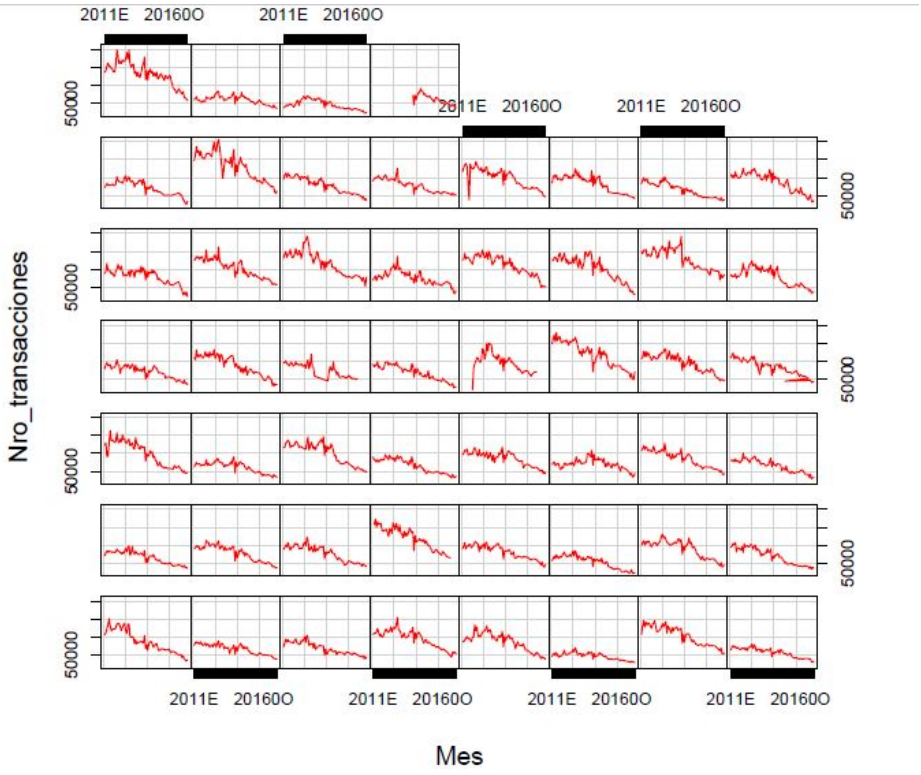
```
        Ubicacion  Ano            Mes Ventas_regular Nro_transacciones
1 001 La Urbina 1 2012   201112 ENERO       11970989            106696
2 001 La Urbina 1 2012 201201 FEBRERO        9676116            112325
3 001 La Urbina 1 2012   201202 MARZO       11605383            124134
4 001 La Urbina 1 2012   201203 ABRIL       10973011            135690
5 001 La Urbina 1 2012    201204 MAYO       12677781            151457
6 001 La Urbina 1 2012   201205 JUNIO       12549587            130723
```

Cada sucursal posee observaciones de 66 meses (enero 2012- abril 2018) tanto para las ventas en bolívares como para el número de transacciones, excepto las sucursales: 27 - Margarita, 29 - El Recreo Barquisimeto y 52 - Manzanares, las dos primeras tienes datos faltantes en sus registros y la última sólo tomaremos 40 meses de registros debido a que se inauguró al final del año 2014. Las sucursales 27 y 29 no se tomarán en cuenta por ahora para el presente estudio debido a los datos faltantes.

A continuación, se presenta las series temporales para las ventas en bolívares por cada una de las sucursales a estudiar.



Mediante esta gráfica podemos tener un panorama general del comportamiento de las ventas en la cadena Central Madeirense durante enero 2012 hasta abril de 2018. La primera impresión que se tiene al ver el gráfico es que a partir de finales de 2015 y principio de 2016 es que las ventas han incrementado considerablemente después de el periodo mencionado, sin embargo debido a los grandes problemas económicos que ha vivido nuestro país a partir del 2014, lo que realmente está permitiendo este comportamiento es la alto incremento de la inflación en los últimos años entre otros factores económicos; este hecho lo podemos corroborar con el siguiente gráfico que muestra el comportamiento de las transacciones por sucursal durante el mismo periodo de tiempo de estudio.



Las series temporales en este gráfico coinciden en una caida abrupta en las transacciones durante el primer trimestre del año 2015, el cual coincide además con la puesta en vigencia de la regulación de los productos de primera necesidad, a partir de ese mes el comportamiento en el consumo de los clientes ha tenido una tendencia bajista.

En el siguiente capítulo revisaremos en detalle el comportamiento de las series de las ventas y transacciones, estableceremos un modelo para cada una por sucusal que nos permita conseguir la mejor predicción.

6.2. Datos para la construcción del modelo de clasificación según el desempeño

6.2.1. Extracción de los datos

Los datos para la construcción del modelo de clasificación fueron facilitados por la empresa, los cuales corresponden a un consolidado de una serie de indicadores financieros y operativos, algunos extraídos desde la base de datos corporativa y otros contruidos a partir del cierre contable mes a mes para cada una de las 52 sucursales a nivel nacional.

6.2.2. Descripción general de los datos

Esta base de datos cuenta con 13 variables de 51 registros por mes, correspondiente a información de las 51 sucursales, excepto la sucursal número 12 - La Isabelica que no se encuentra operativa en este momento, son datos que van desde febrero 2017 hasta noviembre 2017.

```
'data.frame':        510 obs. of  13 variables:
 $ No                       : num  1 2 3 4 5 6 7 8 9 10 ...
 $ Sucursales               : Factor w/ 51 levels "ACARIGUA","AV. ARAGUA",..: 31 4 43 36 19 51 11
 $ Mes                      : Factor w/ 1 level "Septiembre": 1 1 1 1 1 1 1 1 1 1 ...
 $ RotacionInv              : num  1.67 1.38 1.71 2.17 1.51 1.77 2.91 1.65 2.05 1.94 ...
 $ DiasReposicionInventario: num  18 21.8 17.5 13.8 19.9 ...
 $ TicketsProm              : num  24803 27184 47902 28420 22983 ...
 $ NroTransacciones         : num  37815 41412 41273 53580 32007 ...
 $ Ventas                   : num  9.38e+08 1.13e+09 1.98e+09 1.52e+09 7.36e+08 ...
 $ m2SaladeVenta            : num  1973 812 1502 1901 2062 ...
 $ NroPersonal              : num  99 90 87 123 134 66 125 78 97 78 ...
 $ NroRegistradoras         : num  15 10 12 19 16 9 12 11 11 12 ...
 $ GastoTotal               : num  8.59e+08 9.50e+08 1.57e+09 1.39e+09 8.23e+08 ...
 $ Ganancia                 : num  78952577 175659512 407847196 133776690$ -87579207 ...
```

Las variables registran lo siguiente:

- No : variable tipo numérico que almacena la enumeración de las observaciones.
- Sucursales: variable tipo factor que almacena el nombre de cada una de las sucursales.
- Mes: variable tipo factor que guarda los meses del año.
- RotacionInv: variable tipo numérico que guarda el coeficiente de rotación del inventario de cada sucursal por mes.
- DiasReposicionInventario: variable tipo numérico que almacena los días que transcurre para que se deba reponer el inventario por sucursal por cada mes.
- TicketsProm: variable del tipo numérico que almacena el valor promedio de cuánto gasta un cliente por ticket de compra.
- NroTransacciones: variable del tipo numérico que almacena el total de transacciones registradas por cada tienda en un mes.
- Ventas : variable numérica que guarda el total de ventas en bolívares de cada sucursal por mes.
- m2SaladeVenta: variable del tipo numérico que resgitra los metros cuadrádos del piso de venta de cada sucursal, esta variable es fija mes a mes.
- NroPersonal: variable del tipo numérico que almacena el número de personal activo por cada sucursal mes a mes.
- NroRegistradoras: variable del tipo numérico la cuál registra la cantidad de cajas registradoras por sucursal.
- GastoTotal: variable del tipo numérico que almacena todos los gastos que tiene cada sucursal por el mes.
- Ganancia: variable del tipo numérico calculada a partir de las variables ventas y gastos ($gastos - ventas$) que representa la ganancia neta de la empresa por cada sucursal en un mes.

Veamos las correlaciones existentes entre las variables

	Rotacion de Inventario	Dias de Reposicion Inventario	Tickets Promedio	Nro de transacciones	Ventas	m2 Sala de Venta	Nro Personal	Nro Registradoras	Gasto total	Ganancia
Rotacion de Inventario	1	-0,92	0,10	0,35	0,33	-0,27	-0,16	-0,25	0,38	0,02
Dias de Reposicion Inventario	-0,92	1	-0,03	-0,31	-0,22	0,38	0,19	0,33	-0,27	0,05
Tickets Promedio	0,10	-0,03	1	-0,12	0,72	0,10	-0,13	0,03	0,64	0,70
Nro de transacciones	0,35	-0,31	-0,12	1	0,58	0,37	0,54	0,53	0,64	0,16
Ventas	0,33	-0,22	0,72	0,58	1	0,36	0,27	0,42	0,97	0,67
m2 Sala de Venta	-0,27	0,38	0,10	0,37	0,36	1	0,82	0,89	0,44	-0,05
Nro Personal	-0,16	0,19	-0,13	0,54	0,27	0,82	1	0,84	0,35	-0,11
Nro Registradoras	-0,25	0,33	0,03	0,53	0,42	0,89	0,84	1	0,49	0,00
Gasto total	0,38	-0,27	0,64	0,64	0,97	0,44	0,35	0,49	1	0,49
Ganancia	0,02	0,05	0,70	0,16	0,67	-0,05	-0,11	0,00	0,49	1

Las correlaciones significativas (mayores a $0,5$ o menores a $-0,5$) están señaladas en color amarillo, vemos que las variables Rotacion de Inventario y Días de Reposicion Inventario tiene una alta correlación de -0.92, esto se debe a que ambas variables se calcula a partir del movimiento de inventario que se tuvo durante un determinado periodo de tiempo, en este caso un mes, por lo que sólo tomaremos para el estudio a la variable Rotacion de Inventario; en el caso de las otras variables con altas correlaciones haremos el mismo análisis, las variables Ventas y Gastos no serán incluidas en el análisis debido a que la variable Ganancia se construye a partir de estas dos, y esta variable Ganancia es una de las principales referencias de la rentabilidad y desempeño de una sucursal, por lo tanto, la tomaremos para el estudio. La variable Ticket Promedio tiene alta correlación con Ventas, por lo que no la incluiremos en el análisis; las variables Nro personal, m^2 sala de venta y Nro de registradoras tienen alta correlación entre ellas debido a que dependiendo del tipo de formato de cada sucursal (mini, súper, hiper) dependerá los m^2 de la sala de venta, el número dispuesto de cajas registradoras y el personal activo de la sucursal, por lo que estas variables son generalmente fijas y no las tomaremos en cuenta para el estudio pero si para conclusiones finales; mantendremos también para el estudio a la variable Nro transacciones. En resumen, las variables a considerar serán: Rotación de Inventario, Nro transacciones, Ganancia o Ganancia neta, durante el periodo febrero 2017 hasta noviembre 2017. Estas variables recogen información sobre el desempeño de cada sucursal tanto a nivel operativo (RotInv, Nro transacciones) como financiero (Ganancia).

Ahora bien, para nuestro modelo de clasificación construiremos un clasificador que permita modelar el desempeño de las sucursales a partir de las variables seleccionadas. Esta construcción se realizará de la siguiente manera.

Una sucursal con buen desempeño tomando en cuenta las variables escogidas, es aquella que tenga el valor más grande por cada una de las variables mencionadas, es decir, una sucursal con un buen desempeño sería aquella que tenga la más alta rotación de productos, un alto ticket promedio de compra, que tenga el más alto número de transacciones y la ganancia neta más alta.

Por otro lado, el grueso de sucursales de la cadena Central Madeirense son del formato tipo Super, la cual representa el $77\,\%$ (40 de 51 sucursales) del tipo de tiendas de la cadena, por lo que para la construcción del clasificador sólo usaremos los datos de las sucursales con dicho formato para encontrar patrones de comportamiento que nos permita ser de referencia para la clasificación del desempeño.

	No	Sucursales	Mes	RotacionInv	NroTransacciones	Ganancia
1	1	LA URBINA	FEBRERO	2.035742	48667	309077456
2	2	AV. VICTORIA	FEBRERO	1.551303	44136	313996838
3	3	PLAZA LAS AMERICAS	FEBRERO	1.448619	45217	499767676
4	4	LOS RUICES	FEBRERO	1.933129	44966	343166142
5	5	GUARENAS	FEBRERO	1.311432	40953	272055142
6	7	CHACAITO	FEBRERO	3.128428	63116	605378324

Siguiendo con el mismo orden de ideas, cómo una sucursal tiene buen desempeño cuando las variables de estudio son altas, ordenaremos los valores de las variables en orden descendente, donde el valor más alto tendrá 40 (debido a que para la construcción del clasificador sólo usaremos 40 de las 51 sucursales) y el menor 1.

No	Sucursales	RotacionInv	Rank1	NroTransacciones	Rank2	Ganancia	Rank3
1	LA URBINA	1,82270584	28	49704	23	2626195550	12
2	AV. VICTORIA	1,46655084	10	47205	19	2765447089	14
3	PLAZA LAS AMERICAS	1,60223863	15	40919	4	4418588961	34
4	LOS RUICES	2,17153644	37	61953	35	3843705629	30
5	GUARENAS	1,6731655	18	43715	14	1799126803	2
7	CHACAITO	2,84136134	40	70012	39	4711004301	36
10	PRADO DE MARIA	1,85248935	29	46948	18	2303860214	7
13	LOS LEONES	1,99479562	34	48560	21	3287628256	22

Luego calcularemos una variable que llamaremos *Ranking*, esta variable resumirá la posición final que tuvo cada sucursal por el valor de las variables una vez ordenadas, esta se calculará de la siguiente manera:

$$Ranking = \frac{Rank1 + Rank2 + Rank3}{3}$$

Luego de calcular esta variable *Ranking* por cada sucursal, el resultado se ordena de mayor a menor y las sucursales con mejor desempeño estarán de primero y los de más bajo desempeño de últimos.

No	Sucursales	RotacionInv	Rank1	NroTransacciones	Rank2	Ganancia	Rank3	Ranking
7	CHACAITO	2,84136134	40	70012	39	4711004301	36	38
39	IPSFA LOS PROCERES	2,11151957	35	81290	40	7215799933	39	38
4	LOS RUICES	2,17153644	37	61953	35	3843705629	30	34
45	EL PARAISO	1,94057965	32	61811	34	3613329070	27	31
50	LA ALAMEDA	2,3337984	38	43171	11	9918534370	40	30
15	AV. BOLIVAR	1,69909864	24	61215	33	3503321668	26	28
19	CATIA	2,15148838	36	60245	32	3046563689	17	28
23	SAN ANTONIO	1,86261205	30	51667	27	3408035548	23	27
13	LOS LEONES	1,99479562	34	48560	21	3287628256	22	26
30	NAGUANAGUA	1,52893269	11	68185	38	3816386925	28	26

Luego para definir a la variable clasificadora, se establece un intervalo de confianza al indicador *Ranking* para después definir la variable que llamaremos *Desempeño*, la cual será una variable categórica que almacenará las categorías ALTO, MEDIO, BAJO. Cada categoría se definirá de la siguiente manera:

Si $Ranking > \mu_{Ranking} + \sigma_{Ranking} \rightarrow ALTO$

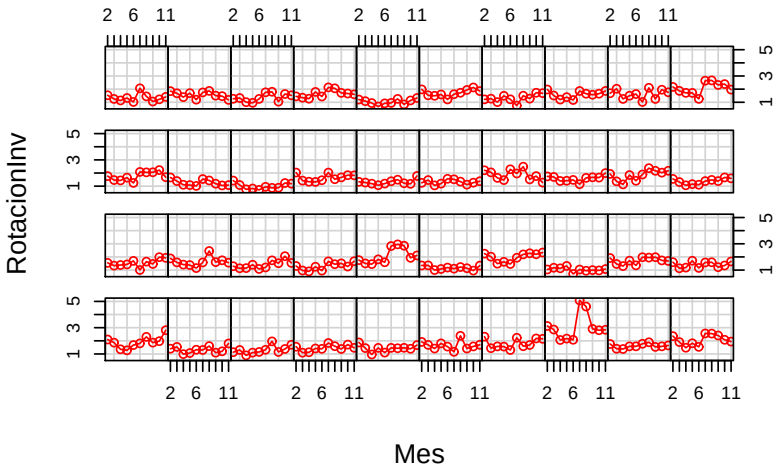
Si $\mu_{Ranking} + \sigma_{Ranking} \leq Ranking \leq \mu_{Ranking} - \sigma_{Ranking} \rightarrow MEDIO$

Si $Ranking < \mu_{Ranking} - \sigma_{Ranking} \rightarrow BAJO$

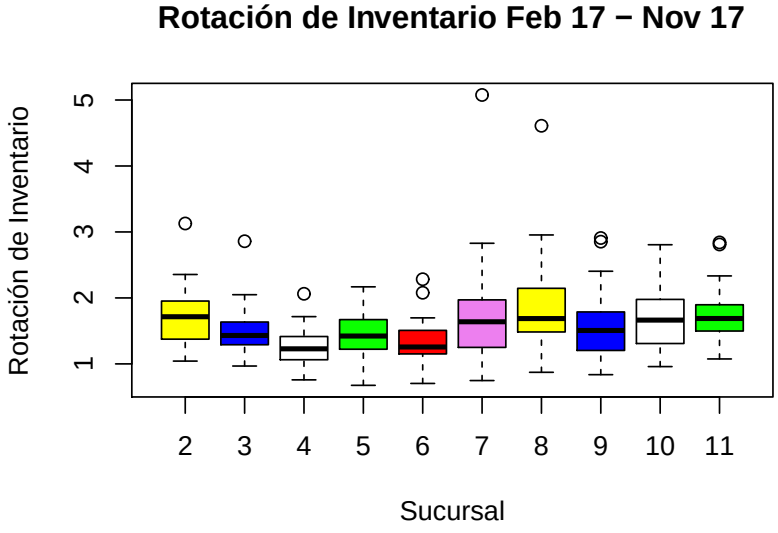
Donde $\mu_{Ranking}$, $\sigma_{Ranking}$ es la media y varianza de la variable Ranking respectivamente.

Así, la variable que modelaremos para la clasificación será la variable construida *Desempeño* y se modelará a partir de las variables: Rotación de Inventario, Nro transacciones, Ganancia.

- Variable rotación de inventario

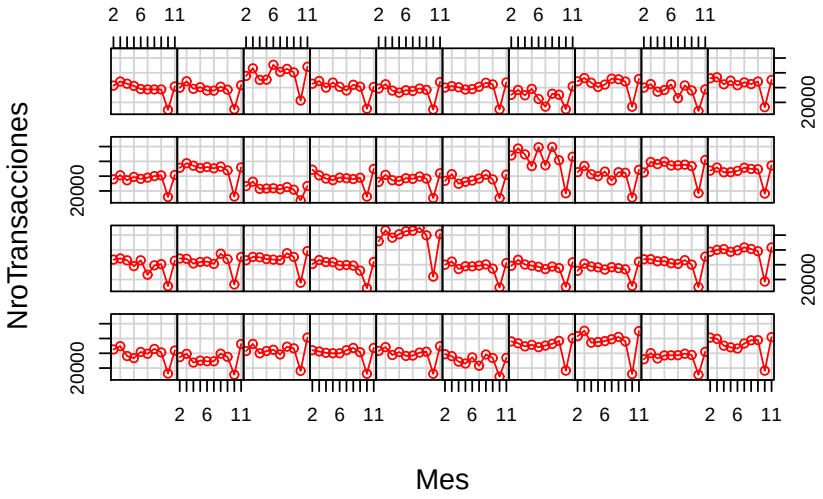


Podemos observar que la serie temporal para la variable rotación de inventario es bantante regular a través de los meses por cada sucursal. A excepsión de la sucursal Chacaito, cuya gráfica se encuentra en la fila 4 columna 8, se ve un gran salto para los meses de julio y agosto de 2017. La gráfica de la sucursal IPSFA Los próceres (file 3, columna 5) también tiene un pico un poco prominente para el mismo periodo.

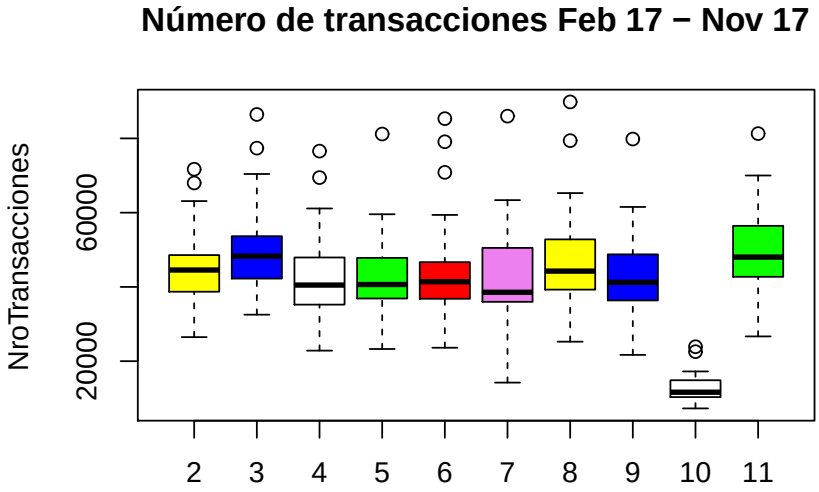


En esta gráfica podemos ver que la media de la variable Rotación de Inventario a nivel cadena a través de los meses estudiados también se mantiene bastante regular.

- Variable número de transacciones

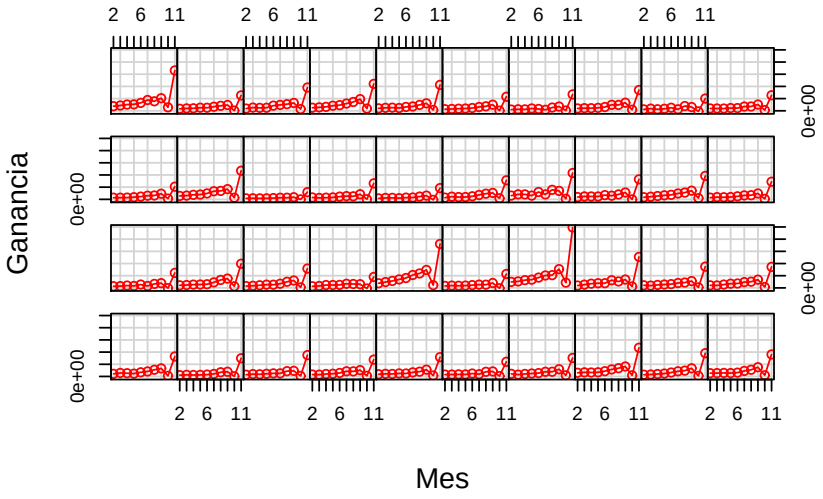


En cuanto a la serie temporal de las transacciones podemos ver que la sucursal IPSFA Los próceres es la que más transacciones genera, siguiendole la sucursal de Montalbán cuya gráfica está en la fila 1 - columna 3. Sin embargo las sucursales presentan una caída para el mes de octubre, suponemos que ese comportamiento se debe a la alza inflacioanria que comenzó a presentarse en ese mes.

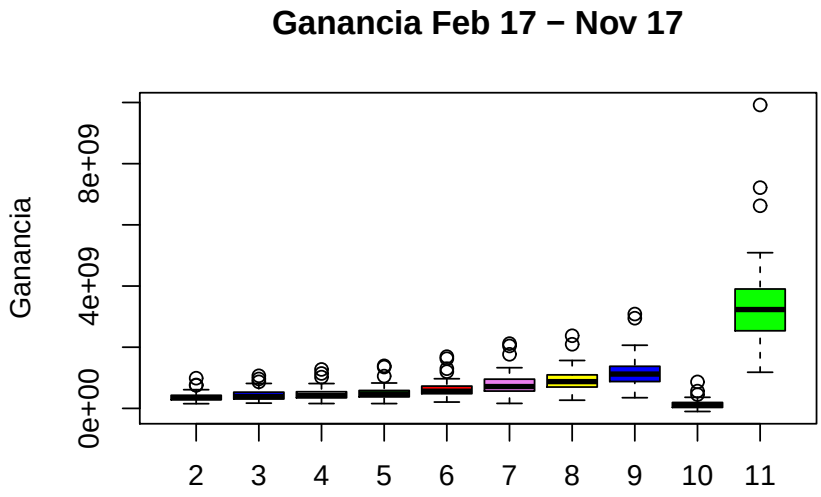


Este fenómeno en el mes de octubre también se evidencia a nivel total cadena.

- Variable ganancia



En cuanto a la ganancia, como esta es calculada a partir de las ventas, como la variable Ticket Promedio, ella también es sensible al fenómeno inflacionario, esto se evidencia el salto que tuvo en el mes de noviembre cuando el porcentaje de variación en los precios comenzó a ser mucho más alto.



Las ganancias para el mes de octubre fueron las más bajas hasta ese entonces, pero tuvo un alza en el mes de noviembre. Lo bajo del mes de octubre se cree es por las bajas transacciones que se tuvo en ese mes y la alza de noviembre por el fenómeno inflacionario.

7. Modelos de predicción para las series temporales de ventas y transacciones

Sucursal 01 - La Urbina

Serie de ventas en bolívares

Comenzaremos los análisis con la serie temporal de las ventas de la sucursal 01-La Urbina, ubicada en la ciudad de Caracas, la serie va desde enero de 2012 hasta marzo de 2018 para predecir abril 2018. Podemos apreciar en la esquina superior izquierda de la *Figura 1* la serie temporal de los datos crudos, la cual se acerca a una distribución exponencial por la manera acelerada que ha venido creciendo las ventas desde mitad del año 2016 apróximadamente, pudiendo deberse a distintos factores económicos entre ellos el fenómeno inflacionario de los precios como lo comentamos en el capítulo anterior.

Ahora bien, para establecer un modelo sobre la media seguiremos la metodología Box-Jenkings quienes han desarrollado modelos estadísticos para series temporales que tienen en cuenta la dependencia existente entre los datos, esto es, cada observación es modelada en función de los valores anteriores. Para ello debemos realizar transformaciones previas a los datos debido a que, al observar la serie, esta evidentemente no es estacionaria sobre su media, una de las hipótesis que debemos cumplir para aplicar dicha metodología. La serie que debemos hallar es aquella que siga la definición de serie estacionaria débil, cómo se definió en el capítulo 2.

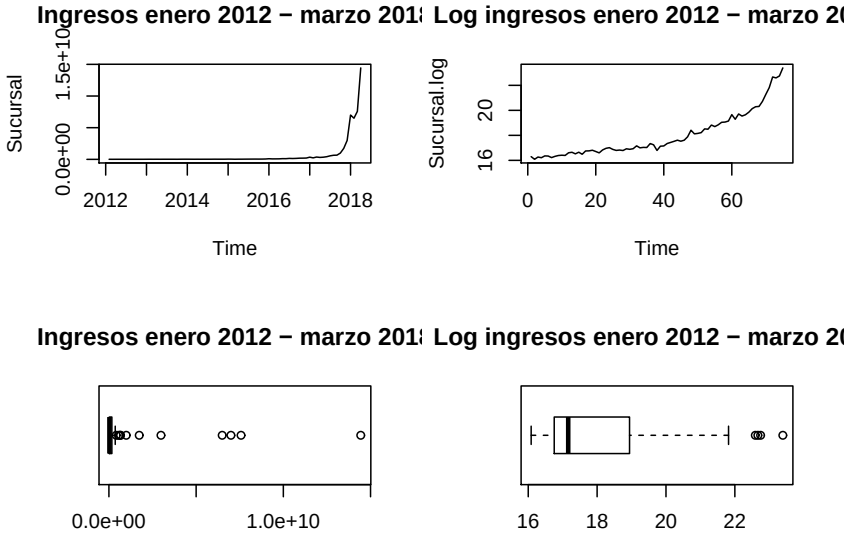
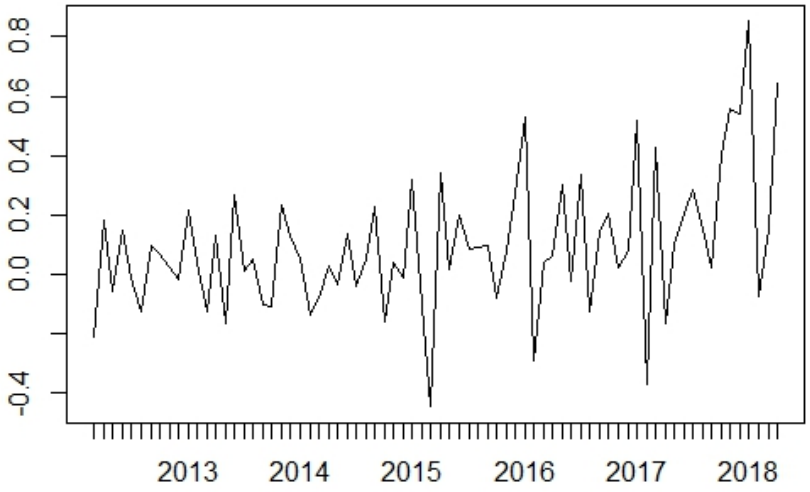


Figura 1: (Izquierda) gráficos de la serie temporal de las ventas y gráfico Boxplot para la sucursal La Urbina. (Derecha) Serie temporal del logaritmo de las ventas de la sucursal La Urbina con su respectico gráfico de Boxplot

Es por ello que aplicamos la función logaritmo a los datos, esto porque suponemos que el aumento acelerado de las ventas se asemeja a una distribución exponencial. La gráfica de esta nueva serie la podemos ver en la esquina superior derecha de la *Figura 1*, también podemos observar que la serie aún

no es estacionaria y que presenta una tendencia alcistas a través de los años, es por esto que aplicaremos una primera diferenciación a los datos en busca de hacerla estable sobre su media, una vez hecho esto, le aplicamos pruebas de hipótesis Portmanteau de raíces unitarias para probar si es estacionaria, entre las pruebas a aplicar están: Prueba Phillip-Perron, Prueba Dickey-Fuller Aumentado,Prueba Kwiatkowski-Phillips-Schmidt-Shin.



Serie temporal del logaritmo de las ventas diferenciada una vez.

- Prueba de raíces unitarias Dickey-Fuller

$$H_0 : \phi = 1 \Rightarrow x_t \sim I(1)$$

$$H_1 : |\phi| < 1 \Rightarrow x_t \sim I(0)$$

Al correr la prueba Dickey - Fuller el valor del p-valor da $0,34$, lo cual es mayor a $0,05$ bajo la hipótesis de $95\,\%$ en el nivel de confianza, por lo que aceptamos H_0, esto es, la serie del logaritmo de ventas diferenciada una vez es proximado a un $I(1)$, una serie integrada de orden 1, lo que significa el mínimo orden que se necesita ser diferenciada la serie para que esta sea estacionaria.

- Prueba de raíces unitarias Phillips-Perron

$$H_0 : \phi = 1 \Rightarrow x_t \sim I(1)$$

$$H_1 : |\phi| < 1 \Rightarrow x_t \sim I(0)$$

La corrida de la prueba Phillips-Perron nos da un p-valor de 0,01, el cual es menor a 0,05, bajo la hipótesis de 95 % en el nivel de confianza, por lo que rechazamos H_0 y aceptamos la alternariva H_1, esto es, la serie del logaritmo de ventas diferenciada una vez es aproximado a un $I(0)$, una serie estacionaria.

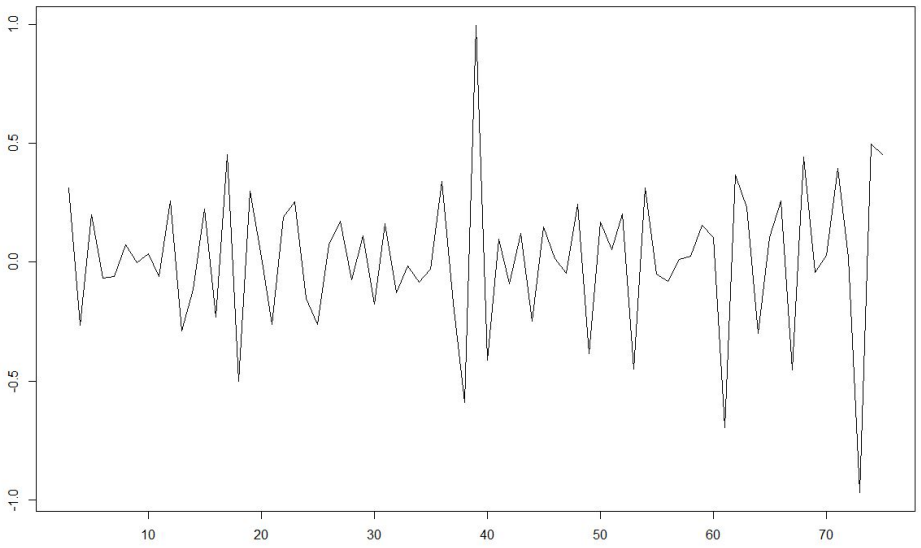
- Prueba de Kwiatkowski-Phillips-Schmidt-Shin

$$H_0 : \sigma_\epsilon^2 = 0 \Rightarrow x_t \sim I(0)$$

$$H_1 : \sigma_\epsilon^2 > 0 \;\; x_t \;\; no \;\; es \;\; estacionaria$$

La prueba Kwiatkowski-Phillips-Schmidt-Shin dio un p-valor de 0,01, el cual es menor a 0,05, bajo la hipótesis de 95 % en el nivel de confianza, por lo que rechazamos H_0 y aceptamos la alternariva H_1, esto es, la serie del logaritmo de ventas diferenciada una vez no es una serie estacionaria.

Podemos ver que, dos de las tres pruebas, no aceptan la hipótesis nula sobre estacionariedad, es por esto se decide diferenciar de nuevo la serie y ver si las tres pruebas aceptan la estabilidad de la serie sobre su media una vez hecho esta segunda diferenciación.



Serie temporal del logaritmo de las ventas diferenciada por segunda vez.

A continuación se presenta los resultados de las pruebas Portmanteau a partir de la serie diferenciada por segunda vez:

Pruebas de estacionariedad para la serie de logaritmo de las ventas diferenciada por segunda vez.

Prueba	P - Valor
Dickey-Fuller	0.01
Phillips-Perron	0.01
Kwiatkowski-Phillips-Schmidt-Shin	0.1

Estas vez las tres pruebas aceptan la estacionariedad de la serie del logaritmo de ventas difereniada por segunda vez. Por lo tanto, es la que usaremos para el proceso de modelaje.

Para determinar un modelo AR, MA o $ARMA$ para la serie estacionaria, es decir, encontrar los órdenes p y q de su estructrura autorregresiva y media móvil, observaremos las gráficas de la Función Autocorrelación Parcial (PACF) y la gráfica de la Función de Autocorrelaciones Simples (ACF).

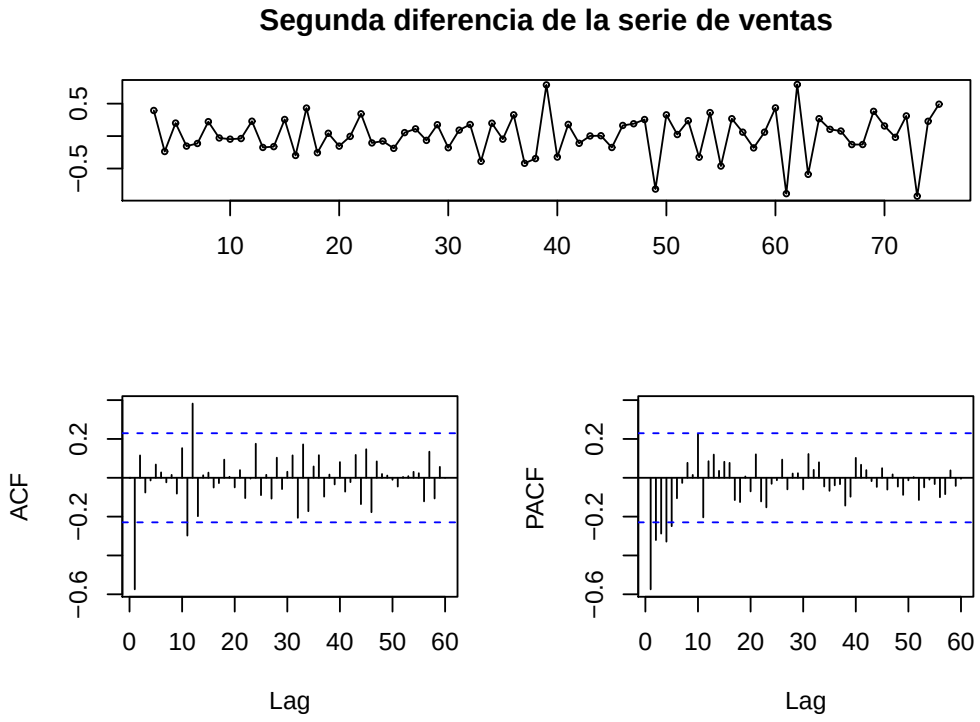


Figura 2: Graficos de de la serie diferenciada 2 veces con sus repectivos gráficos de autocorrelaciones simples y parciales.

En la gráfica del PACF podemos observar autocorrelaciones significativas hasta el cuarto retardo, en el caso de la gráfica ACF existe una autocorrelación significativa en el primer retardo. Además de los modelos que propondremos por medio de las gráficas ACF y PACF estará el modelo que resulte de la función *auto.arima*, el comando hace la elección del modelo en base al mejor modelo ARIMA según

el valor de AIC, AICc o BIC, dependiendo del criterio de información elegido, en este caso usaremos el BIC (Criterio de información bayesiano) y el orden de la primera diferencia en caso de ser necesario, se escogerá de acuerdo a la prueba de Dickey-Fuller Aumentado. Por lo que los modelos a comparar son los siguientes:

```
Arima(Sucursal2, order=c(4,0,1), seasonal=list(order = c(0,0,1), period = 12))
Arima(Sucursal2, order=c(4,0,1), seasonal=list(order = c(1,0,0), period = 12))
Arima(Sucursal2, order=c(4,0,1), seasonal=list(order = c(1,0,1), period = 12))
Arima(Sucursal2, order=c(0,0,2))
```

El modelo con menor BIC fue ARIMA(4,0,1)(1,0,0)[12] con media distinta de cero.

```
Series: Sucursal2
ARIMA(4,0,1)(1,0,0)[12] with non-zero mean

Coefficients:
          ar1      ar2      ar3      ar4      ma1    sar1    mean
      -0.7224  -0.5943  -0.5344  -0.3110  -0.4300  0.4477  0.0074
s.e.   0.2087   0.2239   0.2066   0.1532   0.2064  0.1215  0.0064

sigma^2 estimated as 0.03735:  log likelihood=17.75
AIC=-19.49   AICc=-17.24   BIC=-1.17
```

Quedando la ecuación de la siguiente manera:

$$\phi_4(B)\Phi_1(B^{12})(X_t - \mu) = \theta_1(B)\epsilon_t$$

$$(1 - \phi_1 B - \phi_2 B^2 - \phi_3 B^3 - \phi_4 B^4)(1 - \Phi_{12} B^{12})(X_t - \mu) = (1 + \theta_1 B)\epsilon_t$$

$$(1 + 0,772B + 0,594B^2 + 0,534B^3 + 0,311B^4)(1 - 0,447B^{12})(X_t - 0,07) = (1 - 0,430B)\epsilon_t$$

Ahora evaluaremos el ajuste de este modelo analizando sus residuos.

- Prueba Shapiro-Wilks

$$H_0 : \epsilon_t \sim N(0, \sigma^2)$$

$$H_1 : \epsilon_t \text{ no sigue una distribución normal}$$

La corrida de la prueba Shapiro - Wilks nos da un p-valor de 0,80, el cual es mayor a 0,05 bajo la hipótesis de 95 % en el nivel de confianza, por lo que aceptamos H_0, los residuos siguen una distribución normal.

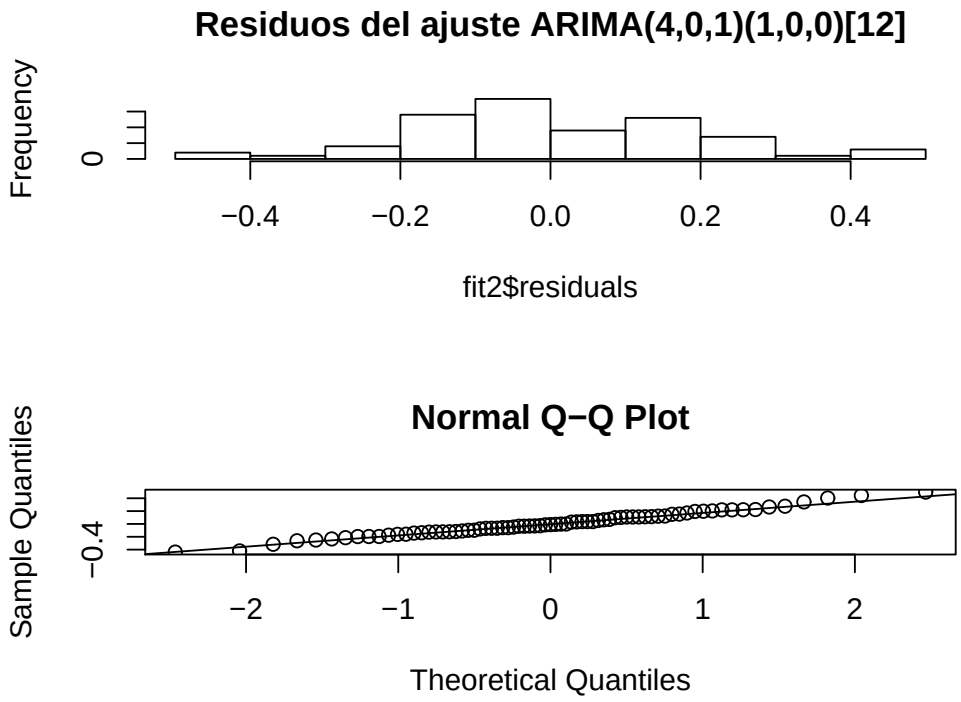


Figura 3: Histograma de los residuos del ajuste con el gráfico de cuantiles muestrales vs. los cuantiles teóricos de la distribución Normal. .

Podemos ver en el histograma que los residuos se asemejan a una campana de Gauss, y en efecto siguen una distribución normal, ya que el p-valor para la prueba de Shapiro- Wilks es de 0,80 que es mayor que 0,05, bajo la hipótesis de 95 % en el nivel de confianza, por lo que aceptamos la normalidad de los residuos.

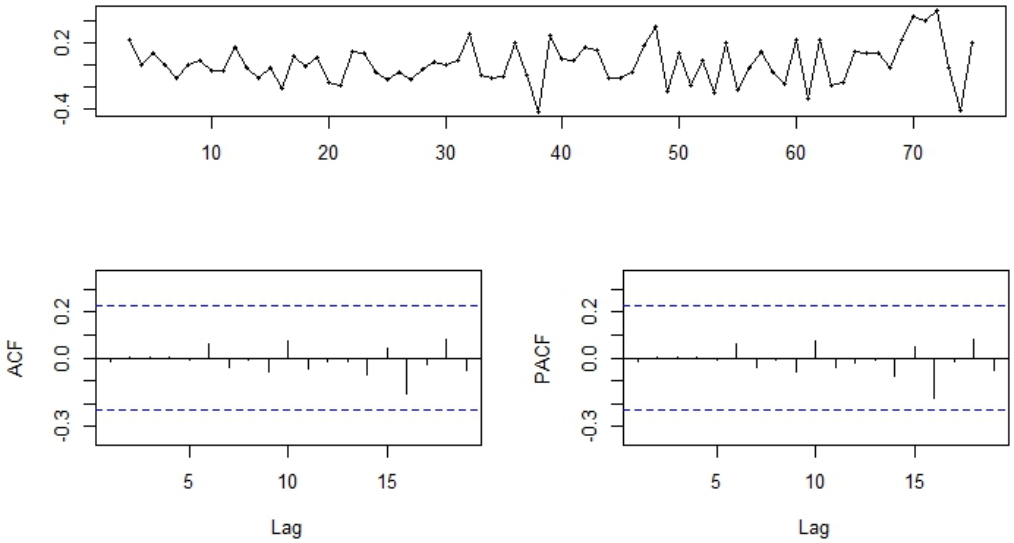
Sometemos los residuos a otras pruebas como la de Box-Ljung en busca de una estructura de ruido blanco y de existencia o no de correlaciones significativas.

- Prueba Box-Ljung

$$H_0 : \epsilon_t \sim RB(0, \sigma^2)$$

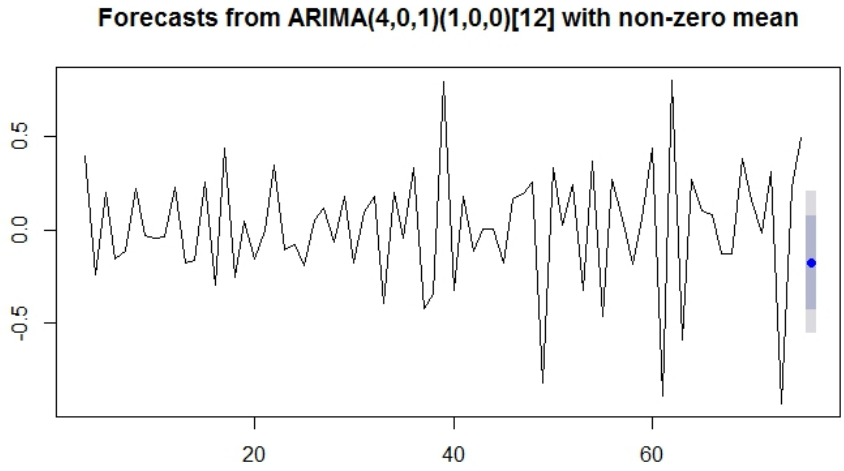
$$H_1 : \epsilon_t \text{ no son ruido blanco}$$

La corrida de la prueba Box - Ljung nos da un p-valor de 0,99, el cual es mayor a 0,05 bajo la hipótesis de 95 % en el nivel de confianza, por lo que aceptamos H_0, esto es, los residuos son ruido blanco. En el siguiente gráfico vemos que los residuos tampoco poseen correlaciones significativas.



ACF y PACF de los residuos del ajuste.

Procedemos entonces a hacer predicción con el modelo ARIMA(4,0,1)(1,0,0)[12] con media distinta de cero.



Predicción del mes de abril 2018 para la serie de ventas diferenciada dos veces.

Una vez obtenida la predicción de las las segundas diferencias de la serie logaritmo de las ventas en bolívares, devolvemos los cambios realizados previamente en los datos, esto es, integraremos dos

veces los resultados de las predicciones y luego de hacer eso aplicaremos la función exponencial de lo que resulte de la integración para así obtener la predicción de los valores originales.

Para la primera diferencia se realizó lo siguiente:

$$x'_t = x_t - x_{t-1}$$

Diferenciando por primera vez permite la eliminación de la tendencia y estacionalidad y como consecuencia estabiliza la media de la serie, sin embargo en nuestro caso se necesito diferenciar por segunda vez para que las pruebas Portmanteau coincidieran en aceptar las hipótesis de estacionariedad, para ello se realizó lo siguiente:

$$x^*_t = x'_t - x'_{t-1} = (x_t - x_{t-1}) - (x_{t-1} - x_{t-2}) = x_t - 2x_{t-1} + x_{t-2}$$

A continuación, el resultado de la predicción junto con sus errores absolutos y relativos.

Resultados

Predicción serie	Predicción Bs	Venta real abril 2018
23,86	23.061.709.271,00	24.062.467.707,00

Error absoluto	Error relativo
1.000.758.436,00	4,15 %

Ahora bien, el error promedio para el pronóstico de abril del presente año es de 4,15 %, en promedio. Lo cual es un modelo bastante aceptable ya que es un error bajo y cumple con los supuestos teóricos.

Sin embargo, el modelo ARIMA(4,0,1)(0,0,1)[12] con media distinta de cero también cumple con los supuestos y el error de predicción es más bajo, como lo veremos a continuación.

```
Series: Sucursal2
ARIMA(4,0,1)(0,0,1)[12] with non-zero mean

Coefficients:
          ar1      ar2      ar3      ar4      ma1    sma1    mean
      -0.6487  -0.5339  -0.5079  -0.2986  -0.4873  0.3555  0.0070
s.e.   0.1886   0.2018   0.1965   0.1514   0.1838  0.1284  0.0051

sigma^2 estimated as 0.03962:  log likelihood=16.09
AIC=-16.18   AICc=-13.93   BIC=2.14
```

Quedando la ecuación de la siguiente manera:

$$\phi_4(B)(X_t - \mu) = \theta_1(B)\Theta_1(B^{12})\epsilon_t$$

$$(1 - \phi_1 B - \phi_2 B^2 - \phi_3 B^3 - \phi_4 B^4)(X_t - \mu) = (1 + \theta_1 B)(1 - \Theta_{12} B^{12})\epsilon_t$$

$$(1 + 0,648 + 0,533B^2 + 0,507B^3 + 0,298B^4)(X_t - 0,07) = (1 - 0,487B)(1 - 0,355B^{12})\epsilon_t$$

Cuyos residuos pasan las pruebas de ajuste como se puede ver a continuación.

```
        Shapiro-Wilk normality test

data:  fit1$residuals
W = 0.98657, p-value = 0.6334
```

La corrida de la prueba Shapiro - Wilks nos da un p-valor de 0,63, el cual es mayor a 0,05 bajo la hipótesis de 95 % en el nivel de confianza, por lo que aceptamos H_0, los residuos siguen una distribución normal.

```
        Box-Ljung test

data:  fit1$residuals
X-squared = 6.4484, df = 20, p-value = 0.9981
```

La corrida de la prueba Box - Ljung nos da un p-valor de 0,99, el cual es mayor a 0,05 bajo la hipótesis de 95 % en el nivel de confianza, por lo que aceptamos H_0, esto es, los residuos son ruido blanco. En el siguiente gráfico vemos que los residuos tampoco poseen correlaciones significativas.

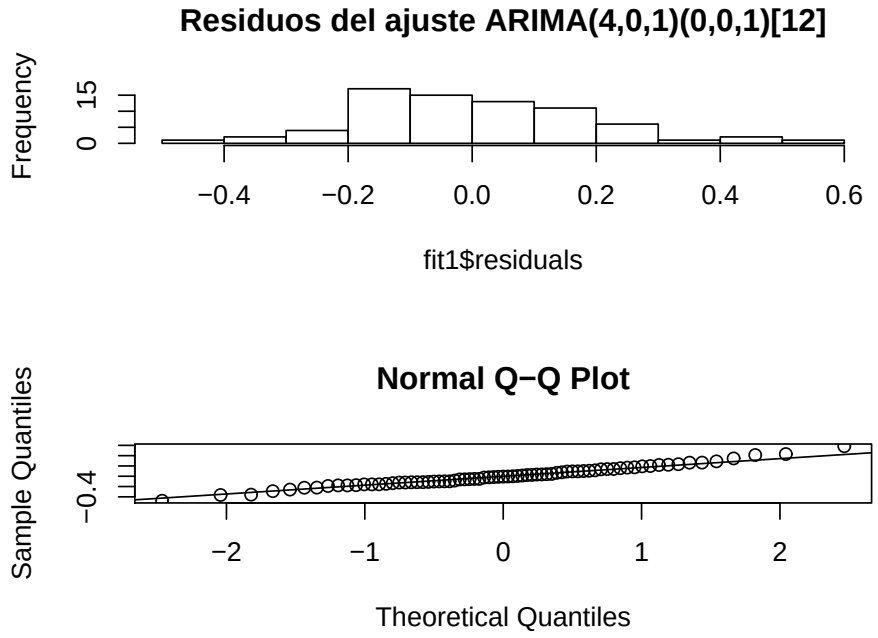


Figura 6: Histograma de los residuos del ajuste con el gráfico de cuantiles muestrales vs. los cuantiles teóricos de la distribución Normal. .

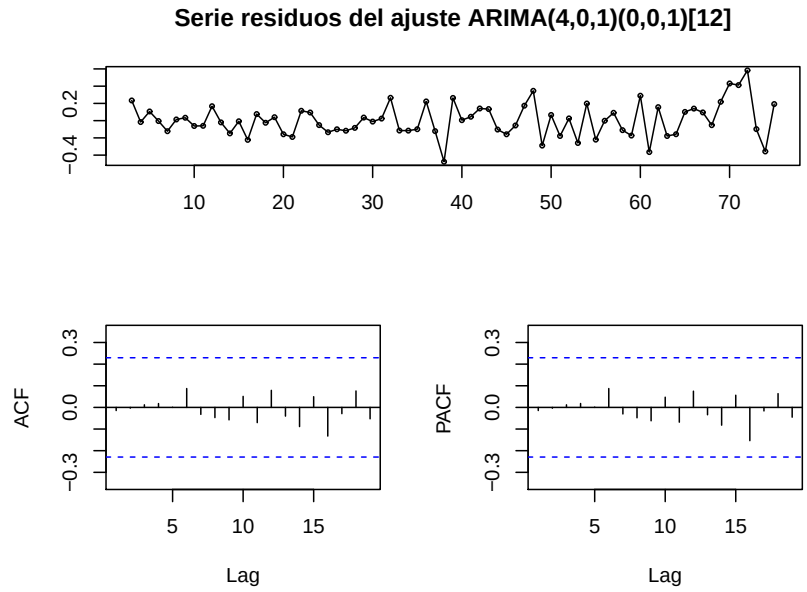


Figura 7: Graficos de de la serie de residuos del ajuste con sus repectivos gráficos de autocorrelaciones simples y parciales.

Podemos ver en el histograma que los residuos se asemejan a una campana de Gauss y que no existe correlaciones significativas en dichos residuos.

Luego procedemos a realizar el pronóstico.

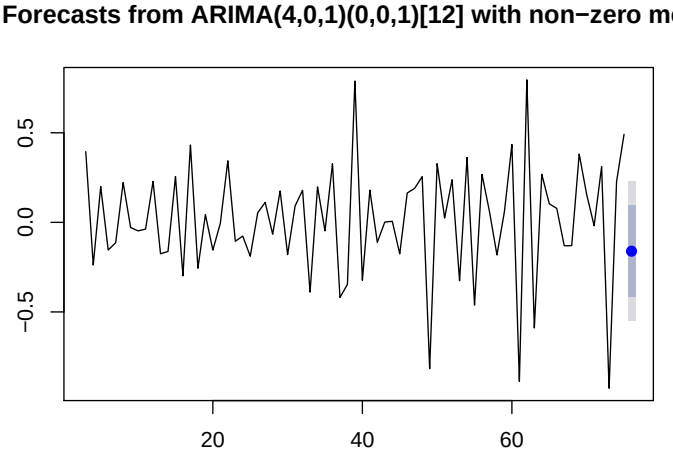


Figura 9: Grafico de la predicción para la serie de la segunda diferencia del logaritmo de las ventas en bolívares.

Resultados

Predicción serie	Predicción Bs	Venta real abril 2018
23,87	23.428.615.146,00	24.062.467.707,00

Error absoluto	Error relativo
633.852.561,00	2,63 %

El error de pronóstico promedio fue del 2,63 %, un error mucho más bajo que el modelo con menor BIC elegido anteriormente, por lo que elegimos a este ajuste como modelo para la media para la sucursal La Urbina.

Serie número de transacciones.

Continuaremos el análisis de la sucursal 01-La Urbina, pero esta vez analizaremos las transacciones obtenidas mensualmente desde enero de 2012 hasta junio de 2017. Podemos apreciar en la *Figura 10* la serie temporal de los datos crudos, donde se puede ver que a tavés de los años ha estado decayendo el número de transacciones.

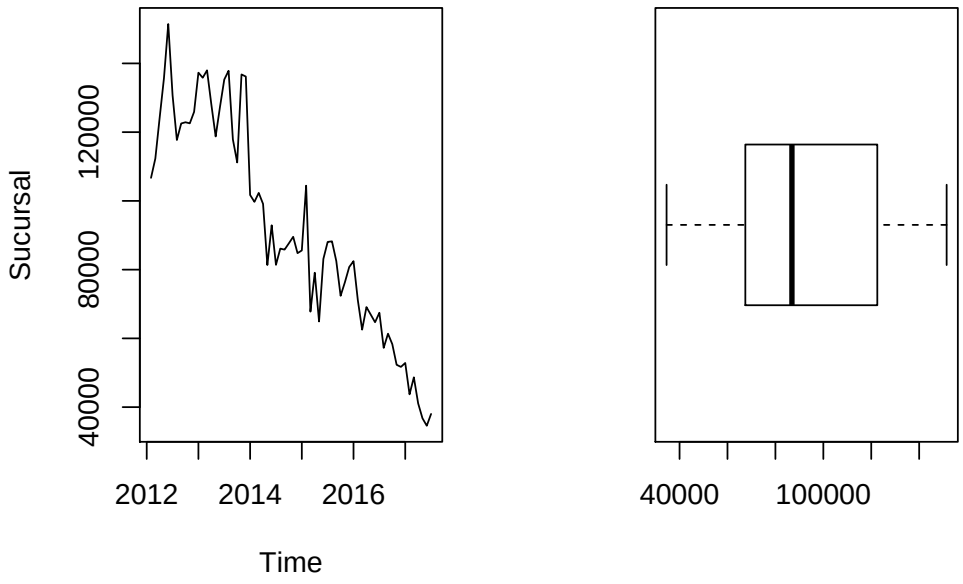


Figura 10: Serie de tiempo de las transacciones durante enero 2012 - junio 2017 y su respectivo gráfico de boxplot

Se puede observar también que existe una especie de ruptura a partir del segundo semestre del año 2014, por lo que gráficamente se pueden ver como si existierán dos series en una a través del tiempo, de hecho al trabajar con la serie completa desde enero 2012 hasta abril 2018 se pudieron encontrar modelos con residuos siguiendo un proceso ruido blanco pero el error de predicción eran grandes.

Por lo que se decide trabajar con la serie a partir de enero de 2014 para modelarlo y hacer la predicción. Es decir, trabajaremos con la siguiente serie.

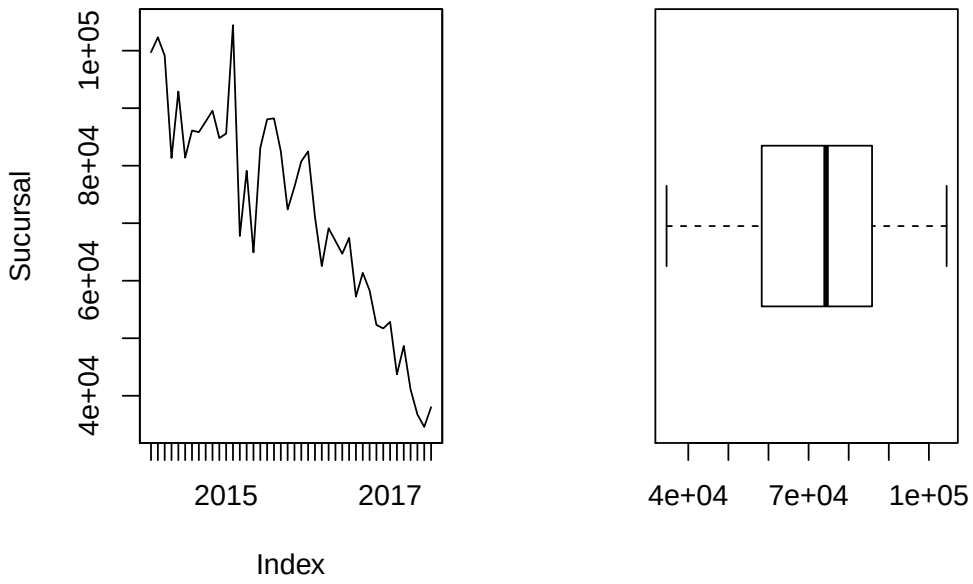


Figura 11: Serie de tiempo de las transacciones durante enero 2014 - abril 2018 y su respectivo gráfico de boxplot

Ahora, con esta serie se puede ver que los datos no tienen periodos tan abruptos a través de los años y en la gráfica boxplot se puede ver también que los datos están un poco más centrados sobre su media.

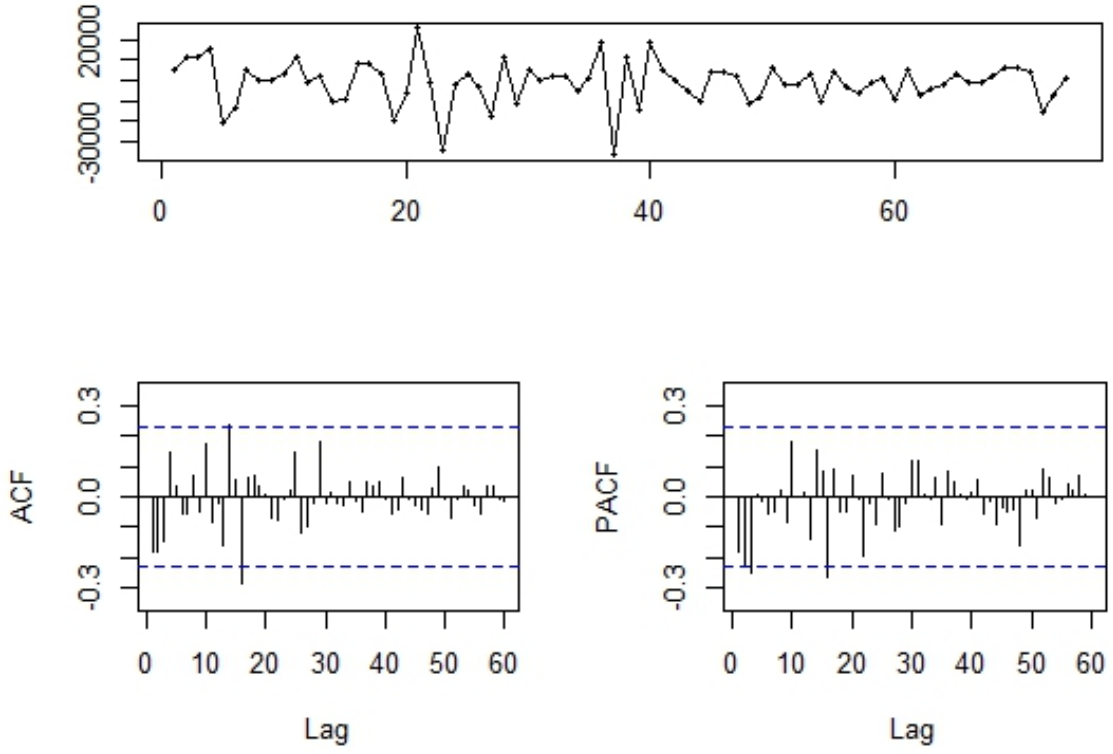
Ahora bien, para establecer un modelo sobre la media seguiremos la metodología Box-Jenkins, como en el análisis de la serie de las ventas. Para ello debemos también realizar transformaciones previas a los datos debido a que al observar la serie no es estacionaria sobre su media, una de las hipótesis que debemos cumplir para aplicar dicha metodología, como ya hemos mencionado.

Podemos observar que la serie no es estacionaria ya que presenta una tendencia bajista a través de los años, es por esto que aplicaremos una primera diferenciación a los datos en busca de hacerla estable sobre su media, una vez hecho esto le aplicamos pruebas de hipótesis Portmanteau de raíces unitarias para probar si es estacionaria.

Prueba	P - Valor
Dickey-Fuller	0.01
Phillips-Perron	0.01
Kwiatkowski-Phillips-Schmidt-Shin	0.1

Las tres pruebas aceptan las hipótesis de estabilidad de la serie sobre su media, por lo que trabajaremos con esta serie diferenciada una vez para construir el modelo y revisaremos las gráficas ACF y PACF para establecer las órdenes del mismo.

ACF Y PACF



Gráficos ACF y PACF de la serie temporal de las transacciones diferenciada una vez.

En la gráfica del PACF podemos observar autocorrelaciones significativas hasta el primer retardo, en el caso de la gráfica ACF existe una correlación significativa en el primer retardo. Además de los modelos que propondremos por medio de las gráficas ACF y PACF estará el modelo que resulte de la función *auto.arima*.Por lo que los modelos a comparar son los siguientes:

```
Arima(Sucursal1, order=c(0,0,0)) #Auto Arima
Arima(Sucursal1, order=c(0,0,2), seasonal=list(order = c(1,0,0), period = 12))
Arima(Sucursal1, order=c(2,0,0), seasonal=list(order = c(2,0,0), period = 12))
Arima(Sucursal1, order=c(0,0,2), seasonal=list(order = c(1,0,1), period = 16))
```

El modelo con menor coeficiente de Akaike (AIC) fue ARIMA(0,0,2)(1,0,1)[16] con media distinta de cero.

```
Series: Sucursal1
ARIMA(0,0,2)(1,0,1)[16] with non-zero mean

Coefficients:
          ma1      ma2     sar1     sma1       mean
      -0.4164  -0.2469  -0.0949  -0.2657  -1297.663
s.e.   0.1199   0.1095   0.3807   0.3732    290.831

sigma^2 estimated as 88993796:  log likelihood=-780.97
AIC=1573.94   AICc=1575.19   BIC=1587.76
```

Los residuos pasan la prueba de Shapiro-Wilks, con un p-valor es de 0,16 lo que es mayor que 0,05 por lo que aceptamos la normalidad de los residuos.

Las siguientes son otras pruebas de ajuste a los datos:

Los residuos también pasan las hipótesis nulas de las pruebas Box-Ljung y Dickey-Fuller Aumentado de ruido blanco y estacionariedad de los residuos con 0,80 y 0,98 respectivamente.

Procedemos por lo tanto a realizar el pronóstico.

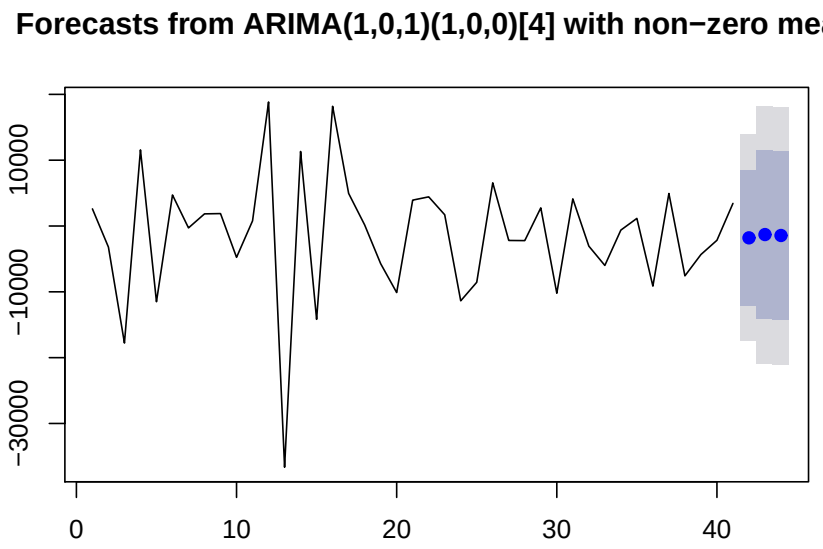


Figura 13: Gráfico de la predicción para la serie diferenciada una vez de los datos transaccionales.

Una vez obtenida la predicción de la primera diferencia de la serie de transacciones, devolvemos los cambios realizados previamente en los datos, esto es, integraremos una vez los resultados de las predicciones para así obtener la predicción de los valores originales.

Para la primera diferencia se realizó lo siguiente:

$$x'_t = x_t - x_{t-1}$$

Diferenciando por primera vez permite la eliminación de la tendencia y estacionalidad y como consecuencia estabiliza la media de la serie. Así,

$$x_t = x'_t + x_{t-1}$$

A continuación presentamos los resultados para la serie de transacciones de la sucursal La Urbina.

Resultados

Predicción serie	Predicción Bs	Transacción real abril 2018
1.006	32.377	30.685

Error absoluto	Error relativo
1.692	5,51 %

El error de pronóstico promedio fue del 5,51 %.

Ahora bien, para el resto de las sucursales se siguió la misma metodología para contrucción de los modelos de pronósticos de las ventas y transacciones, a continuación se presenta la tabla resumen de los modelos por sucursal:

Tabla resumen con los modelos para las ventas y transacciones por sucursal de la cadena Central Madeirense

Modelo Ventas	**%Error**	**Modelo Transacción**	**%Error**
1. ARIMA(4,0,1)(1,0,0)[12]	2,65 %	ARIMA(0,0,2)(1,0,1)[16]	5,51 %
2. ARIMA(4,0,0)(1,0,0)[12]	2,45 %	ARIMA(2,0,0)(1,0,1)[11]	8,77 %
3. ARIMA(4,0,1)(0,0,1)[12]	9,25 %	ARIMA(1,0,1)(1,0,1)[18]	1,20 %
4. ARIMA(4,0,2)(0,0,1)[12]	10,27 %	ARIMA(1,0,1)(1,0,0)[5]	4,16 %
5. ARIMA(2,0,1)(0,0,2)[12]	8,55 %	ARIMA(1,0,1)(1,0,0)[5]	7,93 %
6. ARIMA(3,0,0)(1,1,0)[4]	3,20 %	ARIMA(0,0,1)(1,0,0)[6]	7,76 %
7. ARIMA(4,0,0)(1,0,1)[11]	7,35 %	ARIMA(1,0,1)(0,0,1)[14]	3,33 %
8. ARIMA(2,0,0)(1,0,0)[12]	4,73 %		
9. ARIMA(4,0,0)(1,1,0)[4]	6,84 %	ARIMA(2,0,1)(2,0,1)[14]	3,01 %
11. ARIMA(0,0,2)(1,0,0)[12]	17,78 %	ARIMA(3,2,0)	7,94 %
13. ARIMA(1,0,2)(1,1,0)[8]	9,59 %	ARIMA(2,0,1)(0,1,1)[12]	1,19 %
14. ARIMA(0,0,1)(2,1,0)[9]	12,58 %	ARIMA(1,4,0)(2,1,0)[14]	7,01 %
15. ARIMA(4,1,1)(0,1,0)[9]	1,65 %	ARIMA(1,0,0)(2,0,0)[4]	8,02 %
16. ARIMA(4,1,0)(0,0,1)[12]	9,58 %	ARIMA(3,0,0)(0,0,1)[3]	4,45 %
17. ARIMA(4,0,0)(0,0,2)[12]	5,71 %	ARIMA(2,2,2)	4,85 %
18.		ARIMA(0,0,2)	8,47 %
19. ARIMA(0,0,0)(2,1,2)[12]	2,65 %	ARIMA(1,0,1)	6,91 %
20. ARIMA(3,0,0)(0,0,1)[12]	1,82 %	ARIMA(1,0,1)(1,0,1)[6]	6,99 %
21. ARIMA(3,0,0)(0,1,1)[12]	8,84 %	ARIMA(2,0,1)	4,76 %
22. ARIMA(3,2,2)(1,0,0)[5]	14,31 %	ARIMA(1,0,2)	7,87 %
23. ARIMA(0,1,2)(0,0,2)[24]	10,96 %	ARIMA(1,0,1)(1,0,0)[6]	1,87 %
24. ARIMA(4,0,0)(1,0,0)[6]	11,37 %	ARIMA(1,0,1)(1,0,0)[6]	5,52 %
25. ARIMA(4,0,0)(0,0,1)[12]	9,36 %	ARIMA(1,0,1)	1,73 %

Modelo Ventas	**%Error**	**Modelo Transacción**	**%Error**
26. ARIMA(1,0,2)(0,1,1)[12]	9,44 %	ARIMA(1,1,1)(1,0,2)[3]	2,17 %
28. ARIMA(4,0,0)(0,1,1)[12]	2,75 %	ARIMA(1,0,1)	7,27 %
30.		ARIMA(0,0,3)	2,09 %
31.		ARIMA(1,0,1)	5,81 %
32. ARIMA(4,0,1)(1,0,0)[11]	3,75 %	ARIMA(1,1,1)	0,48 %
33. ARIMA(3,0,0)(1,1,0)[11]	7,98 %		
34. ARIMA(0,0,0)(2,1,0)[12]	4,39 %	ARIMA(1,1,1)	5,24 %
35. ARIMA(2,1,4)(0,1,2)[4]	2,16 %	ARIMA(3,0,0)	3,70 %
36. ARIMA(2,1,2)(0,1,2)[4]	9,26 %	ARIMA(2,1,1)(0,1,1)[4]	6,55 %
37. ARIMA(0,1,2)(1,1,0)[4]	6,00 %		
38. ARIMA(3,1,0)(0,1,0)[12]	8,88 %	ARIMA(2,0,1)(1,0,0)[6]	4,22 %
39. ARIMA(0,0,0)(2,0,4)[10]	9,95 %	ARIMA(2,0,1)(1,0,0)[6]	2,33 %
40.		ARIMA(1,0,0)	9,64 %
41. ARIMA(3,0,0)(0,0,4)[12]	3,22 %	ARIMA(2,1,0)	9,08 %
42. ARIMA(4,0,0)(1,0,1)[36]	6,01 %	ARIMA(0,2,2)	0,42 %
43. ARIMA(0,0,0)(2,0,3)[12]	9,46 %	ARIMA(1,0,1)	11,27 %
44.		ARIMA(3,1,0)	11,08 %
45. ARIMA(2,0,0)(1,1,0)[4]	2,65 %		
46. ARIMA(0,1,0)(1,0,0)[11]	3,01 %	ARIMA(2,1,1)	9,27 %
47. ARIMA(2,0,2)(1,0,0)[4]	7,05 %	ARIMA(1,0,0)	9,03 %
48.		ARIMA(1,0,1)(1,0,0)[3]	7,0 %
49. ARIMA(4,1,1)	3,03 %		
50. ARIMA(0,0,1)(1,0,1)[12]	1,58 %	ARIMA(1,1,1)(0,1,1)[12]	5,24 %
51. ARIMA(0,0,2)(0,1,0)[12]	19,20 %		
52. ARIMA(3,1,1)	4,46 %	ARIMA(1,0,0)	1,96 %

7.1. Modelo de clasificación

Regresión logística ordinal

Para la construcción del modelo de clasificación, usaremos en primera instancia las hipótesis del modelo de Regresión Logística Ordinal revisado en el capítulo Clasificación, esto debido a que la variable a predecir la cual es "Desempeño.[es] una variable del tipo categórica con 3 niveles, estos niveles a su vez poseen un orden los cuales son ALTO, MEDIO y BAJO, para efecto de los cálculos denotaremos por 3,2 y 1 respectivamente, además, como vimos en el apartado de Descripción General de los Datos, las variables explicativas no poseen una correlación significativa.

Tomaremos el mes de septiembre de 2017 para el proceso de entrenamiento. Una vez seleccionado el mes con que el modelo realizará el proceso de aprendizaje, procedemos a estandarizar a las variables cuantitativas debido a que estas están registradas con diferentes escalas. Para ello seleccionamos sólo las variables cuantitativas y las transformamos a formato matriz y con la función *scale* en R realizamos la estandarización.

Visualizamos las primeras 4 filas con las variables estandarizadas.

```
      Sucursales         Mes       RotacionInv  NroTransacciones   Ganancia
1     LA URBINA          SEPTIEMBRE   0.1909810   -0.5538661      -0.7044737
2     AV. VICTORIA       SEPTIEMBRE  -0.3912982   -0.2204793      -0.3310937
3     PLAZA LAS AMERICAS SEPTIEMBRE   0.2858150   -0.2331807      1.2528279
4     LOS RUICES         SEPTIEMBRE   1.2061613   0.9078068       0.3529330
     Desempeno
1     MEDIO
2     MEDIO
3     MEDIO
4     MEDIO
```

Una vez preparado los datos, construiremos el modelo de Regresión Logística Ordinal mediante el método logarítmico acumulativo visto en el capítulo 5.

En este caso de estudio, la variable respuesta "Desempeño"tiene tres niveles: ALTO, MEDIO y BAJO el cual tiene un orden ordinal de 1,2 y 3 respectivamente, siendo 3 el más alto y 1 el más bajo.

La serie de funciones logísticas acumulativas están definidas de la siguiente forma para nuestro caso donde $Y = 1, 2, 3$:

$$L1 = \log(\frac{\pi_1}{\pi 2 + \pi_3}) \text{ y } L_2 = \log(\frac{\pi_1 + \pi_2}{\pi_3})$$

L_j es la proporción de caer dentro o fuera de la categoría j. Ahora en notación lineal el modelo queda de la siguiente forma para las variables regresoras Rotación de inventario, Nro de transacciones y Ganancia:

$$L_1 = \alpha_1 + \beta_1 RotacionInv + \beta_2 NroTransacciones + \beta_3 Ganancia$$

$$L_2 = \alpha_2 + \beta_1 RotacionInv + \beta_2 NroTransacciones + \beta_3 Ganancia$$

Al correr la función *clm* (Cumulative Link Model) en R y calculado los parámetros $\alpha's$ junto a los coeficientes de las variables nos da el siguiente resultado:

```
formula: Desempeno ~ RotacionInv + NroTransacciones + Ganancia
data:    b

link  threshold nobs logLik AIC   niter max.grad cond.H
logit flexible  40   -2.48  14.95 14(0) 1.72e-13 3.4e+03

Coefficients:
Estimate Std. Error z value Pr(>|z|)
RotacionInv        16.672     17.108   0.974    0.330
NroTransacciones    7.494      7.986   0.938    0.348
Ganancia           26.537     27.088   0.980    0.327

Threshold coefficients:
Estimate Std. Error z value
1|2   -35.75      37.10  -0.963
2|3    19.60      18.99   1.032
```

Sustituyendo los coeficientes calculados en el modelo anterior nos queda:

$$L_1 = -35,75 + 16,67 RotacionInv + 7,49 NroTransacciones + 26,53 Ganancia$$

$$L_2 = 19,60 + 16,67 RotacionInv + 7,49 NroTransacciones + 26,53 Ganancia$$

La interpretación de los p-valores es similar a la de un modelo lineal. Podemos ver que ninguna de las variable es significativa para el modelo (cada una tiene un p-valor mucho mayor de 0.05), sin embargo en cuanto a los coeficientes eso cambia debido a que, por ejemplo, el desempeño aumenta 16,67 unidades por cada unidad que aumenta la variable Rotación de inventario; 7,49 unidades cada vez que aumenta la variable número de transacciones y 26,53 cada que aumenta la variable Ganancia, esta última es la que más peso tiene.

Para este modelo los interceptos pueden variar, pero la pendiente para cada variable se mantiene para las diferentes ecuaciones, podemos pensar esto como un conjunto de líneas paralelas con diferentes interceptos. En nuestro caso los interceptos son 1|2 y 2|3, los cuales corresponden a BAJO|MEDIO y MEDIO|ALTO y cuyos coeficientes son -35,75 y 19,60 respectivamente.

Una vez calculado los parámetros y definido el modelo, seleccionamos los datos que probarán el modelo y estos datos serán los del mes de Octubre 2017.

Ya seleccionados los datos de este mes y realizado la estandarización nos queda la base datos como sigue:

	Sucursales	Mes	RotacionInv	NroTransacciones	Ganancia	Desempeno
1	LA URBINA	OCTUBRE	0.38762948	-0.0950931	-0.6596458	2
2	AV. VICTORIA	OCTUBRE	0.09413968	-0.1110659	-0.1919139	2
3	PLAZA LAS AMERICAS	OCTUBRE	-0.02249636	-0.3437727	1.2182247	3
4	LOS RUICES	OCTUBRE	0.84695439	1.1937447	0.3492491	3

Procedemos a realizar la predicción con la función *predict* en R y calculamos la matriz de confusión para evaluar el desempeño de la predicción así como también sus estadísticas.

```
Confusion Matrix and Statistics

      BAJO  MEDIO  ALTO
BAJO    5     0     0
MEDIO   1    23     0
ALTO    0     4     7
```

Al observar la matriz de confusión vemos que de las 6 sucursales que tienen un BAJO desempeño el modelo clasificó bien 5 de 6, en el caso de desempeño MEDIO clasificó bien 23 de 27 y en el caso del desempeño ALTO las clasificó todas de forma correcta.

```
Overall Statistics
Accuracy : 0.875
95\% CI : (0.732, 0.9581)
No Information Rate : 0.675
P-Value [Acc > NIR] : 0.003455

Kappa : 0.7633
Mcnemar's Test P-Value : NA

Statistics by Class:
                     Class:BAJO Class:MEDIO Class:ALTO
Sensitivity             0.8333     0.8519    1.0000
Specificity             1.0000     0.9231    0.8788
Pos Pred Value          1.0000     0.9583    0.6364
Neg Pred Value          0.9714     0.7500    1.0000
Prevalence              0.1500     0.6750    0.1750
Detection Rate          0.1250     0.5750    0.1750
Detection Prevalence    0.1250     0.6000    0.2750
Balanced Accuracy       0.9167     0.8875    0.9394
```

Revisando las estadísticas vemos que tuvo un porcentaje de exactitud (Accuracy) para la predicción en ese mes del 87,5 %, lo cual es bueno ya que la exactitud indica la probabilidad global de que una sucursal sea correctamente clasificado, y el intervalo de confianza para este indicador está entre el 73,2 % y 95,87 %.

De 6 registros que eran del tipo de desempeño BAJO, se clasificó incorrectamente 1, es decir, tuvimos una Sensibilidad (Sensitivity) de 83,33 %. De 27 casos que eran de desempeño MEDIO, clasificamos correctamente 23, esto es, 92.31 % de Especificidad (Specificity) y para el caso de desempeño ALTO se tuvo una sensibilidad del 100 % y especifidad del 87,8 %.

Otra medida útil es el estadístico Kappa. Este nos da una medida de qué tanto mejora nuestro modelo una predicción, contra las probabilidades observadas. Entre más cercano a 1 es el valor de Kappa, nuestro modelo es mejor para predecir que la probabilidad esperada. En general, valores arriba de 0,6 se consideran "buenos", para nuestro caso el valor fue de 76,3 %.

El valor predictivo positivo (Pos Pred Value) indica la probabilidad de que un dato que ha sido predicho como perteneciente a la categoría "BAJO", realmente pertenezca a ella (Class 1: BAJO , en este ejemplo). En este caso, la probabilidad es de 100 %. Por complemento, el valor predictivo negativo

(Neg Pred Value) indica la probabilidad de que un dato predicho como perteneciente a la categoría negativa ("BAJO", "MEDIO.º .ªLTO"), en efecto pertenezca a ella. Esta fue de 97,14 % para BAJO, 75 % para MEDIO y 100 % para ALTO .

Finalmente, la precisión balanceada, indica qué tan bien predice nuestro modelo tanto a la categoría BAJA, MEDIA o BAJA. Esto es muy importante con datos como los nuestros, en los que tenemos clases no balanceadas, es decir, que una es más abundante y tiene más probabilidades de aparecer que la otra. En conjuntos de datos como estos, es fácil obtener una precisión alta para la clase más probable, aunque tengamos poca para la clase menos probable.

Nuestra precisión balanceada es de 91,6 % para BAJO, 88,7 % para MEDIA y 93,9 % para ALTO lo cual está muy bien dedido a que no hay mucha diferencia porcentual entre ellas.

Bayes Ingenuo

Vamos a estudiar otro modelo de clasificación, el cuál es el Bayes Ingenuo y veremos si mejora las estadísticas de predicción del modelo anterior.

Usaremos el mismo conjunto de datos que en el modelo anterior para el aprendizaje, el cual fue el mes de septiembre 2017, y usaremos el mismo mes de octubre 2017 como conjunto de pruebas del modelo.

El clasificador de Bayes ingenuo es un algoritmo basado en la aplicación del teorema de Bayes con la suposición ïngenua"de independencia entre cada par de características o clases. Dada una variable de clase Y y un vector de característica dependiente o atributos X_1 a X_n, el teorema de Bayes establece la siguiente relación:

$$P(Y|X_1, ..., X_n) = \frac{P(Y)P(X_1, X_2, ..., X_n)}{P(X_1, ..., X_n)}$$

Usando la hipótesis ingenua de independencia tenemos:

$$P(X_i|Y, X_1, X_2, ..., X_{i-1}, X_{i+1}, ..., X_n) = P(X_i|Y)$$

para todo i, esta relación se simplifica a:

$$P(Y|X_1, ..., X_n) = \frac{P(Y)\prod_{i=1}^{n} P(X_i|Y)}{P(X_1, ..., X_n)}$$

donde $P(X_1, ..., X_n)$ es constante independiente de los datos de entrada, podemos usar la siguiente regla de clasificación o mejor llamado el clasificador de Bayes ingenuo:

$$P(Y|X_1, ..., X_n) \propto P(Y)\prod_{i=1}^{n} P(X_i|Y)$$

$$\downarrow$$

$$\hat{Y} = argmax_Y P(Y)\prod_{i=1}^{n} P(X_i|Y)$$

En nuestro caso, las probabilidades condicionales $P(X_i|Y)$ son calculados con la función de densidad Gaussiana debido a que los atributos o predictores son variables aleatorias continuas. La función que usaremos en R para estos cálculos se llama *naiveBayes*, esta función cálcula la media y la desviación estándar de cada atributo usando los datos de las clases BAJO, MEDIO y ALTO, así como sus funciones a priori como podemos ver :

```
Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace, type = ..1)

A-priori probabilities:
Y
1     2     3
0.175 0.650 0.175

Conditional probabilities:
RotacionInv
Y         [,1]      [,2]
1 -0.7550917 0.3871597
2 -0.1788568 0.7204528
3  1.4194168 1.0163866

NroTransacciones
Y          [,1]      [,2]
1 -1.05871424 0.4837620
2  0.05336948 0.7465144
3  0.86048473 1.3266830

Ganancia
Y          [,1]      [,2]
1 -1.06347184 0.2708073
2 -0.09604548 0.5062906
3  1.42021219 1.2829691
```

Luego el algoritmo cálcula las probabilidades condicionales a partir de la función de densidad normal de la siguiente forma:

$$P(X_i|Y_j) = \frac{1}{\sqrt{2\pi}\sigma_{j,i}} \epsilon^{\frac{(x_i - \mu^2_{j,i}}{2\sigma^2_{j,i}}}$$

Una vez calculado el modelo, se procede a realizar la predicción del mes de Octubre 2017 con este modelo.

```
Confusion Matrix and Statistics


BAYE_PREDICT  1  2  3
           1  4  0  0
           2  2 25  4
           3  0  2  3

Overall Statistics

Accuracy : 0.8
95\% CI : (0.6435, 0.9095)
No Information Rate : 0.675
P-Value [Acc > NIR] : 0.0602

Kappa : 0.5455
Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class1: BAJO  Class2: MEDIO  Class3:ALTO
Sensitivity                0.6667         0.9259        0.4286
Specificity                1.0000         0.5385        0.9394
Pos Pred Value             1.0000         0.8065        0.6000
Neg Pred Value             0.9444         0.7778        0.8857
Prevalence                 0.1500         0.6750        0.1750
Detection Rate             0.1000         0.6250        0.0750
Detection Prevalence       0.1000         0.7750        0.1250
Balanced Accuracy          0.8333         0.7322        0.6840
```

En las estadísticas de esta predicción vemos que tuvo un porcentaje de exactitud (Accuracy) para la predicción en ese mes del 80,0 %, un poco menos que el modelo anterior, sin embargo sigue siendo un muy buen valor y el intervalo de confianza para este indicador está entre el 64,3 % y 90,5 %.

De 6 registros que eran del tipo de desempeño BAJO, se clasificó correctamente 4, es decir, tuvimos una Sensibilidad (Sensitivity) para esta clase de 66,67 %. De 27 casos que eran de desempeño MEDIO, clasificamos correctamente 25, con un 53,85 % de Especificidad (Specificity) y para el caso de desempeño ALTO se tuvo una sensibilidad del 42,86 % y especifidad del 93,94 %.

El estadistico Kappa nos dió 0,545 y la precisión balanceada, el cuál indica qué tan bien predice nuestro modelo tanto a la categoría BAJA, MEDIA o BAJA, nos dió 83,33 %, 73,22 %, 68,40 % respectivamente.

Ambos modelos tienen buen desempeño con estadíticas bastante similares sin embargo el que hasta ahora tiene una ligera ventaja es el modelo de regresión logística ordinal. En la siguiente sección realizaremos predicción con otros meses y compararemos el desempeño de ambos modelos.

7.2. Comparación entre los modelos de clasificación Regresión Logística Ordinal y Bayes Ingenuo.

En esta sección compararemos ambos modelos Regresión Logística Ordinal y Bayes Ingenuo para ver cómo predicen otros meses del año, una vez expuesto los resultado decidiremos cual es el modelo que mejor se ajusta a los datos. Hasta los momentos vimos que el clasificador de la regresión logística ordinal tiene un mejor índice de exactitud pero corroboraremos si también lo tiene al predecir cualquier otro mes.

Estadísticas de la predicción mes de febrero 2017 para las sucusales con formato SUPER

	CLM			NAIVE BAYES		
Accuracy		0,875			0,825	
Kappa		0,781			0,653	
	| BAJO	MEDIO	ALTO |	| BAJO	MEDIO	ALTO |
Sensitivity	0,857	0,840	1,000	0,571	0,920	0,750
Specificity	0,939	0,933	0,937	0,969	0,666	0,968
Pos Pred Value	0,750	0,954	0,800	0,800	0,821	0,857
Neg Pred Value	0,968	0,777	1,000	0,914	0,833	0,939
Prevalence	0,175	0,625	0,200	0,175	0,625	0,200
Detection Rate	0,150	0,525	0,200	0,100	0,575	0,150
Detection Prevalence	0,200	0,550	0,250	0,125	0,700	0,175
Balanced Accuracy	0,898	0,886	0,968	0,770	0,790	0,859

Los dos modelos siguen teniendo un buen índice de exactitud para el pronósticos del mes de febrero 2017, en este caso el que va teniendo un ligero mejor desempeño en el modelo de la regresión logística ordinal.

Estadísticas de la predicción mes de julio 2017 para las sucursales con formato SUPER

		CLM			NAIVE BAYES	
Accuracy		0,900			0,925	
Kappa		0,808			0,833	
	|BAJO	MEDIO	ALTO |	| BAJO	MEDIO	ALTO|
Sensitivity	1,000	0,857	1,000	1,000	0,964	0,600
Specificity	1,000	1,000	0,885	1,000	0,833	0,971
Pos Pred Value	1,000	1,000	0,555	1,000	0,931	0,750
Neg Pred Value	1,000	0,750	1,000	1,000	0,909	0,944
Prevalence	0,175	0,700	0,125	0,175	0,700	0,125
Detection Rate	0,175	0,600	0,125	0,175	0,675	0,075
Detection Prevalence	0,175	0,600	0,225	0,175	0,725	0,100
Balanced Accuracy	1,000	0,928	0,942	1,000	0,898	0,785

En este caso, para la predicción del mes de junio 2017, el modelo con mejores estadísticas en su pronóstico fue el Bayes ingenuo.

Finalmente, ponemos a prueba ambos modelos, pero esta vez tomamos las 51 sucursales que hay a nivel nacional, es decir, añadiremos a las sucursales con los otros formatos con que cuenta la empresa (formato Mini e Hiper) y veremos cómo reacciona cada modelo al agregar sucursales que no estuvieron en la etapa de entrenamiento y evaluaremos la clasificación, esto lo haremos para el mes de noviembre 2017.

Estadísticas de la predicción mes de noviembre 2017 las 51 sucursales de todos los formatos

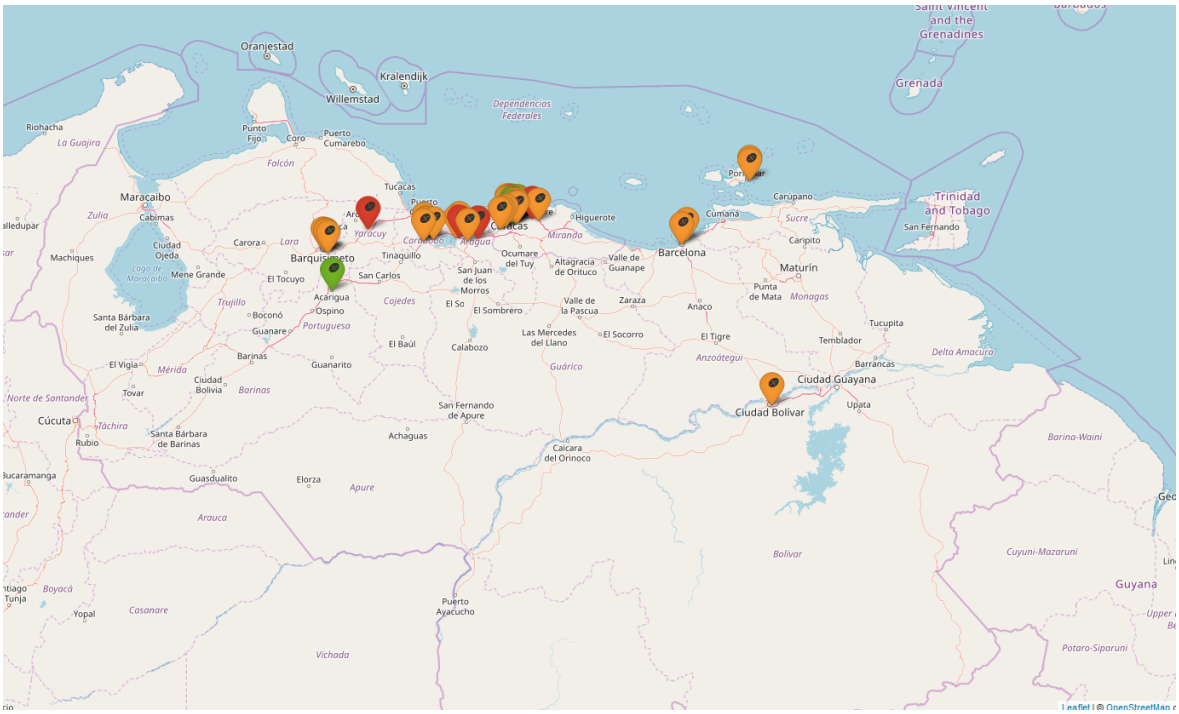
	CLM			NAIVE BAYES		
Accuracy		0,8039			0.8235	
Kappa		0,5518			0.5846	
	|BAJO	MEDIO	ALTO |	| BAJO	MEDIO	ALTO|
Sensitivity	0,571	0,842	0,833	0,857	0,868	0,500
Specificity	0,977	0,692	0,888	1,000	0,692	0,888
Pos Pred Value	0,800	0,888	0,500	1,000	0,891	0,375
Neg Pred Value	0,934	0,600	0,975	0,977	0,642	0,930
Prevalence	0,137	0,745	0,117	0,137	0,745	0,117
Detection Rate	0,078	0,627	0,098	0,117	0,647	0,058
Detection Prevalence	0,098	0,705	0,196	0,117	0,725	0,156
Balanced Accuracy	0,774	0,767	0,861	0,928	0,780	0,694

A nivel general ambos modelos tuvieron estadísticas similires y reaccionaron de buena manera ante los nuevo registros añadidos. Cómo vimos en esta y en la sección anterior, cualquiera de los dos modelos pueden ser tomados para clasificar al conjunto de datos estudiados, esto debido a que no hay diferencia significativas entre ambas, pero sólo eligiremos a una de ellas como el modelo definitivo de claficiación para este caso de estudio. Ese modelo será el modelo de Bayes Ingenuo.

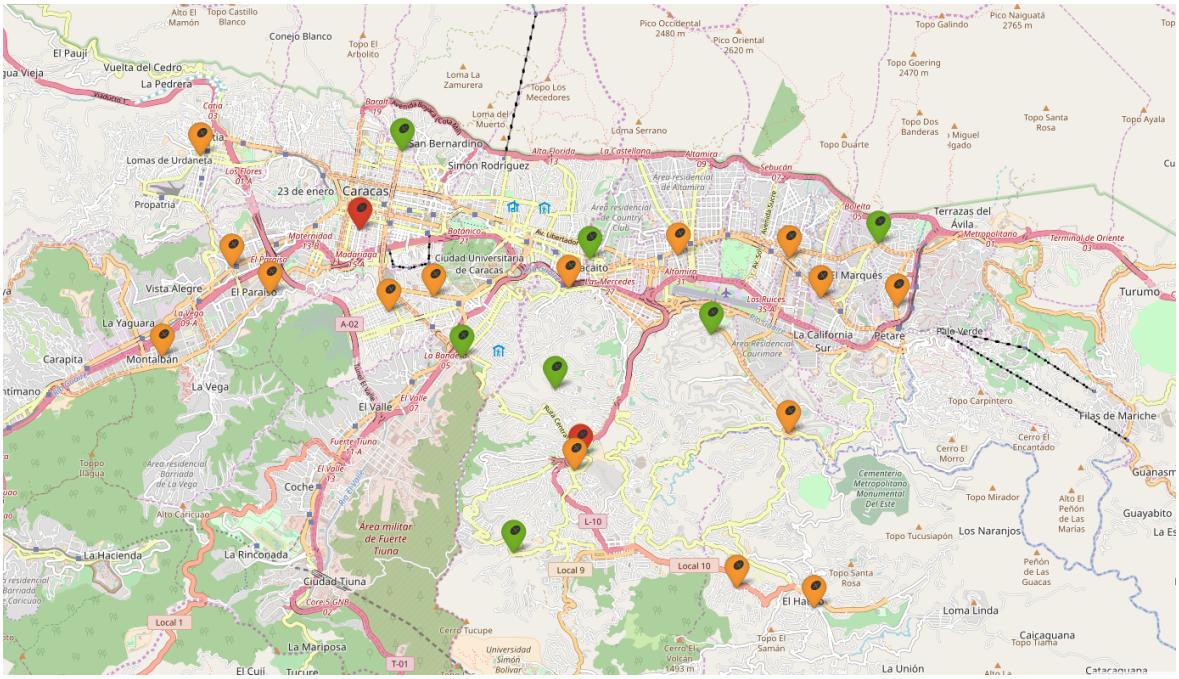
A continuación, presentamos la predicción de la clasificación de las 51 sucursales a través del modelo Bayes Ingenuo para el mes de noviembre 2017, por medio de mapas de geolocalización donde se representa a su vez qué tipo de desempeño tuvo en ese mes según el color de la etiqueta: verde para desempeño alto, naranja para desempeño medio y rojo para desempeño bajo.

Ubicación de las sucursales geográficamente y su nivel de desempeño

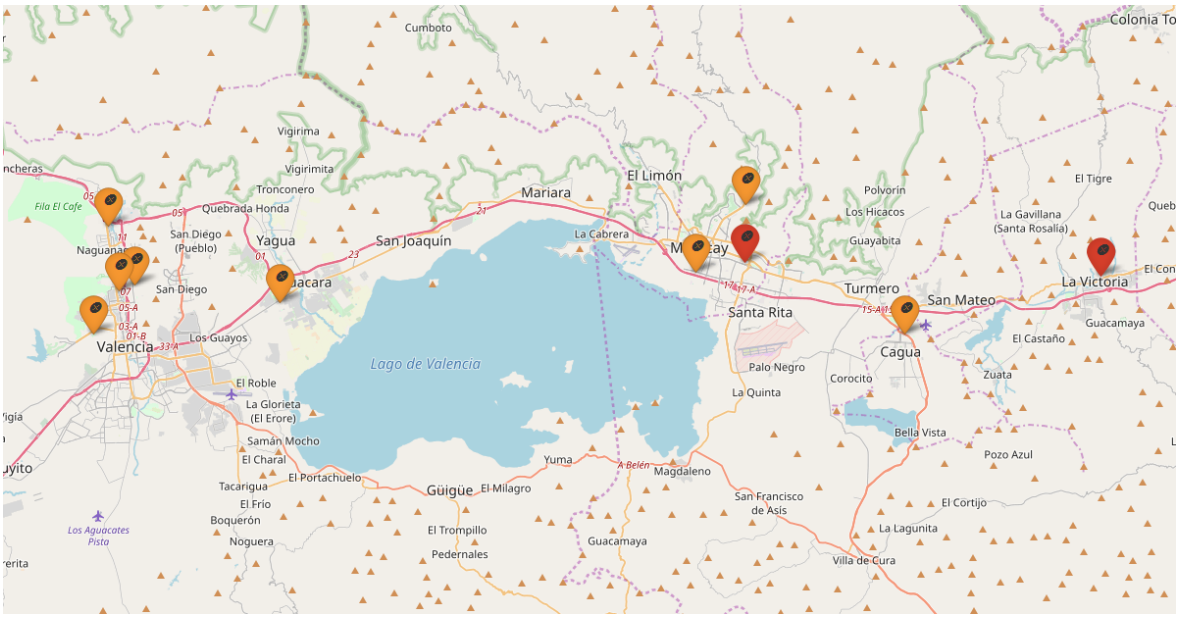
Venezuela



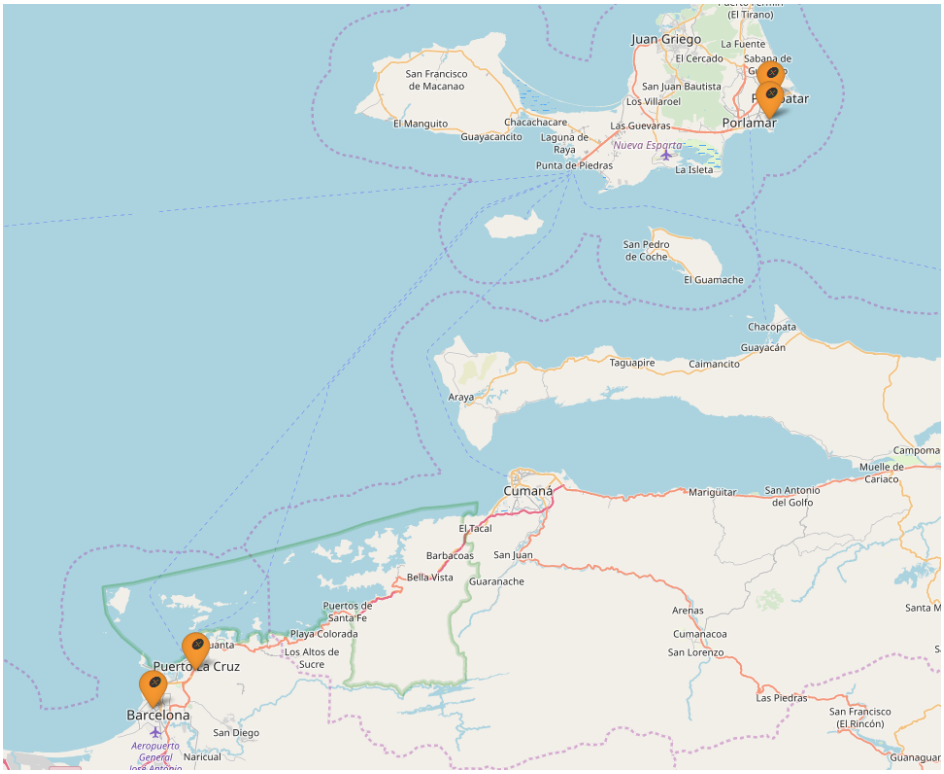
Región capital -Venezuela



Región central -Venezuela



Región oriental -Venezuela



8. Conclusión

Tras realizar los estudios detallados a la serie temporal de las ventas para cada una de las sucursales de la cadena Central Madeirense durante enero 2012 hasta abril de 2018, se puede concluir que en efecto las ventas de la cadena han tenido un aumento acelerado luego del primer trimestre de 2016, sin embargo, al analizar de manera simultánea la serie de las transacciones, esto es, el número de compras registradas por las cajas registradoras de las sucursales durante el mismo periodo de estudio y ver que, el comportamiento de estas es opuesto a la serie de las ventas, deja en evidencia de que ese incremento en las ventas es debido al fenómeno inflacionario que ha venido en aumento mes a mes. Una de las razones del comportamiento de las transacciones se puede deber al hecho de las regulaciones en los precios de los productos, cuya medida entró en vigencia para el primer trimestre del 2015, se puede ver en las gráficas que justo en ese periodo de tiempo las series por sucursales coinciden en una caída abrupta en las transacciones y a partir de ahí las transacciones vienen teniendo una tendencia bajista. Otro factor influyente en el comportamiento de las transacciones es el desabastecimiento, la discontinuidad o baja producción de productos.

Al construir los modelos de ajustes para ambas series, se debieron realizar transformaciones previas a los datos para tener una serie estacionaria y poder aplicar la metodología Box-Jenkins, sin embargo, en el caso de la sucusal 10 (Prado de María) no se puedo encontrar un modelo de ajuste tanto para ventas como transacciones. En el caso de: 18 (La Boyera), 30 (Naguanagua), 31(Guacara), 33 (Puerto La Cruz), 40(IPSFA Maracay), 44 (La Lagunita) y 48 (Cagua) si se pudieron encontrar modelos para transacciones pero no para las ventas; y caso contrario ocurrió con la sucursal 37 (Barcelona). Luego de analizar los modelos encontrados por sucursal, no se pudieron encontrar patrones similares de manera de poder realizar una clasificación de las mismas por medio de los modelos de ajuste de series temporales, es por esto que en este trabajo de investigación estudiamos dos métodos de clasificación que nos pueda dar un panorama sobre el desempeño de cada una de las sucursales mes a mes.

Como vimos en la sección 7.2 del capítulo 7, donde comparabamos los dos modelos Regresión Logística y Bayes Ingenuo, ambos modelos tuvieron un buen aprendizaje sobre la variable Desempeño cuando se entrenó ambos modelos con las 40 sucursales con tipo de formato SUPER, sin embargo, el Bayes Ingenuo tuvo mejor acierto en la predicción para diferenciar cada un de los niveles que conformaba dicha variable una vez que se agregaron el resto de las sucursales de los otros dos formatos para la predicción. Con esta herramienta se permitiría conocer el desempeño de una sucursal a partir de la variaciación de sólo tres variables, las cuales son: rotación de inventario, número de transacciones y ganancia; es decir, que para el modelo, la estrategia está en tener bien controladas estas variables para alcanzar el objetivo: todas las sucursales tengan un alto desempeño. Una de las fortalezas de este modelo de clasificación es la rápida aplicatividad en un escenario con inflación media moderada.

Este desempeño de las sucursales, si lo vemos a nivel operacional y financiero, como lo hicimos en este estudio, depende de muchos factores y sobre todo es sensible a los distintos cambios y fenómenos económicos que vive hoy en día nuestro país. Pudimos ver que el comportamiento de las sucursales con formato Super poseen comportamientos un tanto similares a pesar de estar en zonas geográficas muy distintas. Sin embargo, podemos ver también por el estudio, que aún sucursales como La Alameda, Manzanares siguen estando en el top 5 en cuanto desempeño y en cuanto a ganancia son las que más aportan, esto podría deberse a que por estar posicionada geográficamente en una zona de estratos altos, la cadena pone especial atención en estas, tanto en tema de abastecimiento como en la variedad en su catálogo de productos hasta donde les permite la situación del país. En este análisis se demuetra que existen otras sucursales, como por ejemplo IPSFA Los Próceres y Chacaito que si bien no están localizadas en zonas de estratos altos son una buena oportunidad para revisar el catálogo de productos o estrategias de venta que permitan aumentarlas ya que estas presentan uno de los índices de transacciones más elevados y a nivel de desempeño se han posicionado, en algunos meses, en el top 5 por lo que la posición geografía si bien pudo haber sido un indicador para saber si una sucursal

tendía o no un buen desempeño, no es la que realmente está dando la pauta en la actual situación que se vive.

La situación que vive Venezuela crea patrones de compras muy distintos a los patrones tradicionales que se pudieron vivir en años anteriores, esto debido a que la escasez, desabastecimiento y la inflación han provocado que personas de estratos bajos se dirijan a zonas de estratos altos en busca de los productos básicos y que el poder adquisitivo del cliente haya disminuido. Es por ello que un estudio más a fondo es requerido si se quiere saber con certeza que o cuáles sucursales tienen una clientela de alto target o qué sucursales poseeen una alta fidelidad, y para ello, se debe obviar los productos básicos por cada departamento que pueden distorcionar los datos. Así mismo, los modelos de proyecciones que se puedan construir sobre algún indicador financiero se deben hacer por periodos muy cortos, debido a que la varianza de los precios es cada vez más grandes en periodos muy cortos.

Las grandes cadenas como Central Madeirense, si bien a nivel de negocio, están más enfocados en estos momentos en mantener y controlar los gastos operativos, deben apostar por estudios más profundos que le pueden proporcionar herramientas de desición para evaluar la rentabilidad de las sucursales de manera eficaz para la toma de desiciones que permitan aumentar a su vez la rentalibilidad del negocio, herramientas que las distintas metodologías dentro de la minería de datos puede brindar.

Referencias

[1] Carvajal, Alexander. *Series temporales: Modelos heterocedásticos condicionales. Una aplicación usando R.* Universidad de Granada.Departamento de estadística e investigación operativa.2014.

[2] Brockwell Peter J., Davis Tichard A.. *Introduction to Time Series and Forecasting.*Springer. Segunda edición.2002.

[3] Villavicencio, J.. *Introducción a Series de Tiempo.*2014.

[4] Hosmer, David W. y Lameshow, Stanley. Applied Logistic Regression . Second edition.

[5] Mitchell, Tom M.. Machine Learning . McGraw-Hill Science.

[6] Barber, David.Bayesian Reasoning and Machine Learning .

[7] Frisenfeldt T., Kristine. Analysis of ranked preference data. Technical University of Denmark.