UNIVERSIDAD NACIONAL AGRARIA LA MOLINA

Facultad de Economía y Planificación

Departamento Académico de Estadística e Informática

Trabajo Monográfico

MODELO PREDICTIVO DE QUIEBRE DE STOCK EN UN SUPERMERCADO COMPARANDO DOS MÉTODOS DE SELECCIÓN DE VARIABLES

Presentado para Optar el Título de Ingeniero Estadístico e Informático

BEATRIZ DEL CARMEN LIDIA MONTAÑO MIRANDA

Modalidad Examen Profesional

LIMA - PERÚ
2013

UNIVERSIDAD NACIONAL AGRARIA LA MOLINA

Facultad de Economía y Planificación

Departamento Académico de Estadística e Informática

MODELO PREDICTIVO DE QUIEBRE DE STOCK EN UN SUPERMERCADO COMPARANDO DOS MÉTODOS DE SELECCIÓN DE VARIABLES

Trabajo Monográfico presentado para Optar el Título de Ingeniero Estadístico e Informático

Autor: BEATRIZ DEL CARMEN LIDIA MONTAÑO MIRANDA

Modalidad Exámen Profesional

Mg. Sc. Clodomiro Fernando Miranda Villagómez
Presidente

MS. Rino Nicanor Sotomayor Ruiz
Miembro

Ms. Luz Jeanet Bullón Camarena
Miembro

Mg. Raphael Félix Valencia Chacón
Miembro

Mg. Raquel Margot Gómez Oscorima
Representante de Consejo de Facultad

ÍNDICE

Pág.

RESUMEN

El presente estudio tuvo como objetivo principal verificar la disponibilidad de productos en el almacén de un Supermercado, así las decisiones a tomar ante la falta de productos serían más certeras y se tendría un mejor panorama al respecto de la situación del abastecimiento del Supermercado.

El trabajo consiste en un modelo de predicción de quiebres de stock para un supermercado. Se analizaron las ventas en unidades, el stock de los productos, los despachos de proveedores, los días del mes de Agosto (con ventas y sin ellas) de un grupo de productos de diferentes gerencias del supermercado (Abarrotes Comestibles, Abarrotes no Comestibles y Bebidas). Con la información recopilada se consideró el valor de la variable dependiente (en quiebre con valor 1 y 0 en caso contrario).

La selección de variables por Boruta permitió obtener un modelo con menos cantidad de variables y con un mejor ajuste al realizar el análisis usando la regresión logística en comparación con la selección de variables por Stepwise.

1. INTRODUCCION

El número de supermercados en el país ha incrementado, generándose una competencia agresiva entre las diversas cadenas de consumo, incidiéndose así en un mayor número de novedosas promociones, acuerdos publicitarios, espacios administrativos, etc. originándose el crecimiento del número de compras por parte de los clientes. Durante este proceso de compra se tiene que cumplir la función de abastecimiento, que es la encargada de suministrar los recursos y de adquirir la importancia fundamental del desempeño del supermercado, condicionado a los costos productivos y la capacidad de respuesta para los clientes. Es por eso que el abastecimiento se ha vuelto un motivo de competitividad entre los Supermercados.

La gestión de abastecimiento es un área muy poco atendida en muchas organizaciones y por lo tanto presenta un gran potencial de mejora. Las compañías que han comprendido el valor estratégico del abastecimiento no sólo han reestructurado esta función, sino que han comenzado a replantearse las formas tradicionales de las compras y su relación con los proveedores, dando lugar a una visión más integradora de la cadena de abastecimiento del Supermercado con relaciones de colaboración entre sus distintos actores(proveedores, fabricantes, transporte, detallistas, distribuidores, los clientes, comunicación y compradores) implementando mejoras conjuntas, y redefiniendo roles a lo largo de la cadena. Buscando que se cumplan las especificaciones requeridas de las cantidades, dimensiones y/o calidad de las compras solicitadas, que incurrirán en mayores costos por devoluciones, reproceso o desperdicios, repercutiendo negativamente en el precio final del producto y en el nivel de servicio al cliente. De la misma forma, el mantener altos niveles de inventarios implica soportar altos costos de mantenimiento, incurrir en costos de oportunidad y asumir riesgos de roturas, robos u obsolescencia.
Entonces así los Supermercados han podido generar un valor superior y posicionarse de manera más competitiva entre los mercados de consumo detallista.

Se puede resumir que todas las actividades que se llegan a cumplir y a finalizar adecuadamente en la cadena de suministros, hasta llegar al cliente final generan

utilidades y beneficios, esto justifica el desarrollo de un modelo que permita identificar los elementos más importantes en proceso de abastecimiento de supermercados y se pueda llegar a tomar las decisiones pertinentes para reducir al mínimo la falta de mercadería en almacén, teniendo siempre una opción de respuesta inmediata para la mayor satisfacción de los clientes.

2. EL PROBLEMA DE INVESTIGACIÓN

2.1 Fundamentación del problema de investigación

El mundo del retail en el Perú se encuentra en cúspide; según la última encuesta internacional realizada por Consensus Economics. El Perú lideraría en el 2014 el crecimiento económico de América Latina con 6,2% y tendría la inflación más baja de la región con 2,1%. Dicha encuesta es realizada entre bancos de inversión, analistas económicos y empresas consultoras en los diferentes países de América Latina. La economía peruana tendría un crecimiento de 6.2% para el 2014, tasa mayor a la que se alcanzaría este año. Por lo mismo entonces que los empresarios peruanos son más optimistas respecto al crecimiento económico nacional, informó el Banco Central de Reserva del Perú al precisar que de "agosto a setiembre" el indicador de expectativas saltó de 48 a 53 puntos. Esa es la conclusión de la Encuesta de Expectativas macroeconómicas elaborada por la institución, la cual incluye los diez indicadores más monitoreados, los cuales crecieron en su totalidad.

Se percibe así un cambio en el estilo de vida de la mayoría de peruanos, que se encuentran en búsqueda ya no de los productos sino de las promociones que brindan los Supermercados y la misma que origina una rivalidad entre los Supermercados por los consumidores y por como fidelizarlos (Arellano, marketing). Así se ha logrado una dinamización de todos los retailers (supermercados, bodegas y mercados) generando una competencia y con esto un beneficio para los consumidores en precio, calidad y servicio (Comercio). En estos aspectos, se ha avanzado mucho en los últimos años, los supermercados han liberado una lucha constante por conseguir mayor cantidad de clientes y ofrecerles los mejores productos, estupendas promociones, un servicio de calidad, etc. evidenciándose la necesidad de un adecuado abastecimiento de los productos en la cantidad precisa, en el momento exacto y al mínimo costo. Los supermercados conseguirían la reducción de sus costos totales en el sistema mediante acciones efectivas de marketing, abastecimiento, producción y distribución, asegurando la respuesta efectiva a la demanda del cliente, minimizando el lead time (el menor tiempo) entre la producción y la venta al cliente, maximizando el flujo de caja,

reduciendo el nivel de inventario, mejorando los plazos de pago a proveedores, asegurando las ventajas competitivas en el tiempo de introducción de los nuevos productos y mejorando el servicio del cliente. Si se consigue una exitosa gestión de la cadena de suministros, para la entrega del producto apropiado al cliente final, en el lugar correcto, en el tiempo exacto, al precio requerido y con el menor costo posible. Estamos logrando clientes satisfechos y mayores ingresos para el Supermercado.

Como parte de la cadena de suministros tenemos a los proveedores y compradores, los cuales hacen las negociaciones para que se atiendan los pedidos que deben llegar a cada supermercado, abasteciéndolo y cumpliendo con las órdenes de compra de productos requeridos por el comprador del negocio. En muchas ocasiones los mismos compradores del negocio no llegan a solicitar la mercadería suficiente para satisfacer la demanda de los puntos de venta, ocasionando el desabastecimiento en las góndolas donde se colocan los productos solicitados, siendo el almacén de la tienda el principal lugar en el cual no se encuentran los productos requeridos. Esto conlleva a que los clientes cuando visiten la tienda no encuentren los productos en dicho lugar. Para dar solución a la falta de stock es necesario tener conocimiento de los quiebres en el supermercado y la manera como predecir estos quiebres, como encargados de la negociación con los proveedores de tal manera que nos puedan atender a tiempo y no presenciar problemas de quiebres de stock en el almacén del supermercado. Por lo cual, es importante contar con un modelo predictivo de propensión de quiebres de Stock en el almacén de la tienda que nos proporcione la probabilidad alta, media y baja de los productos, pertenecientes a los distintos proveedores, a punto de quebrar.

Para llegar a obtener un buen modelo es necesario contar con variables de predicción apropiadas que nos permitan estimar la existencia o no de los productos en almacén. Para esto el supermercado cuenta con una gran base transaccional con indicadores de ventas, compras, stocks, pedidos y márgenes, etc. De todos los productos que se ofrecen en piso de venta de cada supermercado. Debido a la gran cantidad de información de los productos por cada proveedor, es importante contar con una metodología que permita hallar el mejor subconjunto de variables de la base total y así implementar un modelo de predicción de quiebres con las mejores variables predictoras.

Las técnicas de selección de variables y de reducción de dimensiones son fundamentales en este contexto debido a que modelos más parsimoniosos son deseables desde el punto de vista de la interpretación así como de la reducción en los errores de predicción. Para dicha selección tenemos los métodos secuenciales (forward selección, backwar delimination y el stepwise), que pueden resultar fuertemente inestables o directamente inaplicables cuando el número de variables es similar o incluso ampliamente superior al número de observaciones. Debido a esto, las nuevas metodologías han desarrollado en las últimas décadas soluciones que permitan enfrentar el problema o maldición de la dimensionalidad. Un conjunto amplio de estas técnicas puede plantearse agregando a la función objetivo, un ajuste que mida los datos en un determinado modelo con un término de penalización.

2.2 Formulación de las preguntas de investigación

- ✓ ¿Cuál es el modelo para detectar quiebres de stock para los distintos productos en supermercados mediante sus datos y reportes transaccionales?
- ✓ ¿Cuáles son las mejores variables derivadas de las bases de datos de las transacciones del SAP (sistema informático basado en módulos integrados que abarca todas las áreas de una empresa)y el Business (modulo del SAP) que ayuden a predecir los faltantes de mercadería en almacén (FMA)?
- ✓ ¿Cuáles la técnica de selección de variables, que presenta un mayor aporte?

2.3 Objetivos de la investigación

Objetivo General

- ✓ Construir un modelo para detectar quiebres de stock para distintas categorías en supermercados mediante sus datos y reportes transaccionales.

Objetivos Específicos:

- ✓ Identificar las variables derivadas de las bases de datos de las transacciones del SAP y del Business que ayuden a predecir los faltantes de mercadería en almacén (FMA).
- ✓ Evaluar y comparar los dos modelos de selección de variables.
- ✓ Identificar las técnicas de selección de variables que brindan un mayor aporte.

2.4 Justificación de la investigación

El objetivo principal es la detección de quiebres de stock consiguiendo así el nivel de servicio para los Sku (stock keeping unit) de todos los productos de la base de datos revisada, de esta manera se obtendría una visión más global de lo que está sucediendo en los almacenes de los supermercados y cuáles son los productos con mayor número de problemas (quiebres).

Lo cual nos permitiría analizar los que tengan mayores faltantes y tener productos con características similares, ventas similares para cubrir dichos faltantes disminuyendo así las ventas perdidas.

Se detectaría en el momento justo de la falta de mercadería en almacén (FMA), dicho en otras palabras, cuando un SKU no está disponible en el almacén. Es importante la solución de este problema dados los altos costos en que se ven afectados los actores involucrados (retailer, proveedores, consumidores, etc.). La forma de resolver esta situación, será a partir de modelos de predicción computacionales, para así llegar finalmente a un modelo que entregue en sus "outputs" pronósticos de cuándo ocurrirá la FMA, poder anticiparse a dicha falta y evitarla en el almacén.

Evitando la falta de mercadería en almacén (FMA), ya no se originarían diversas respuestas por parte de los clientes, siendo algunas de ellas cambio de marca, tipo o tamaño, búsqueda del producto en otro supermercado, aplazamiento de la compra (volver otro día por el producto) y cancelación de la compra. Dados estos posibles escenarios, el resultado de ellos se puede traducir en pérdida de clientes, reducción de ventas, baja rentabilidad, mala imagen para el supermercado, etc.

3. MARCO TEÓRICO

3.1 Modelo Estadístico

El objetivo de trabajo se centra en predecir el comportamiento de una variable categórica con dos niveles la disponibilidad de productos en los almacenes de un Supermercado o la no disponibilidad de productos en los almacenes de un Supermercado.

En el presente estudio se utiliza la técnica de Regresión Logística.

3.2 Regresión Logística

Es una de las herramientas estadísticas con mejor capacidad para el análisis de datos en distintas áreas de la investigación.

El objetivo primordial que resuelve esta técnica es el de modelar como influye en la probabilidad de aparición de un suceso, habitualmente dicotómico, la presencia o no de diversos factores y el valor o nivel de los mismos. También puede ser usada para estimar la probabilidad de aparición de cada una de las posibilidades de un suceso con más de dos categorías. Este tipo de situaciones se aborda mediante técnicas de regresión. Sin embargo la metodología de la regresión lineal no es aplicable ya que ahora la variable respuesta solo presenta dos valores.

Definición

Sea Y una variable dependiente binaria que toma dos valores posibles etiquetados como 0 y 1.

Sean $x_1,x_2,\ldots,x_p$ un conjunto de variables independientes observadas con el fin de explicar o predecir el valor de Y.

El objetivo es determinar $P(Y=1/x_1,x_2,\ldots x_p)$, donde p número de variables independientes.

Por lo tanto $P(Y=0/x_1,x_2,\ldots x_p)$.

Se construye un modelo de la forma:

$$P(Y=1/x_1,x_2,...x_p)_{=} P(Y=x_1,x_2,...x_p;\beta) \quad (1)$$

Donde:

$P(Y=x_1,x_2,...x_p;\beta)$: Rp [0,1]

Es una función que recibe el nombre de función de enlace cuyo valor depende de un vector de parámetro

$\beta=(\beta_1, \beta_2, \beta_3,\ldots, \beta_p)$

3.3 Modelo de Regresión Logística Binaria

Sea:

$$P(Y=x_1,x_2,...x_p;\beta)= G(\beta_0+\beta_1X_1+ \ldots.+\beta_pX_p) \quad (2)$$

Donde: $G(x)=\frac{e^x}{1+e^x}$

Es la función de densidad acumulada correspondiente a la función logística. El modelo es:

$$\log\left(\frac{p(x_1,\ldots\ldots, x_p;\beta)}{1-p(x_1,\ldots\ldots, x_p;\beta)}\right)=\beta_0+\beta_1x_1+\ldots..+\beta_px_p \quad (3)$$

3.4 Modelo de Regresión Logística Simple

Para construir el modelo matemático es necesario tener valores numéricos, los cuales se obtienen considerando P (Y=1) en relación con la dependencia de que dicha probabilidad no ocurra 1-P (Y=1).

La probabilidad es un número que oscila entre 0 y 1, que proporciona predicciones consistentes y de fácil interpretación de los resultados en términos de razón de probabilidades llamado "odds ratio".

Sea la función:

$$f(x) = \frac{1}{1+e^{-x}} \qquad (4)$$

El modelo logit será $\log\left(\frac{p_i}{1-p_i}\right) = \beta o + \beta_1 X$

Donde log significa logaritmo en base diez, β_0y β_1 son constantes y x es una variable explicativa que puede ser continua o discreta. El campo de la variación de $\frac{p_i}{(1-pi)}$ es todo el campo real, mientras para p el campo es solo de 0 a 1.

3.5 Estadístico Wald

Evalúa el coeficiente estimado en la población y se define como un cociente entre el coeficiente y el error estándar del coeficiente en la hipótesis.

H0: $\beta_1=0$

H1: $\beta_1\neq0$

Estadístico de prueba

Wald= $\frac{\beta_i}{S_{\beta_i}} \approx N(0,1)$ o lo que es equivalente a $\left(\frac{\beta_i}{S_{\beta_i}}\right)^2 \approx \lambda^2_1$. Esta distribución del estadístico de Wald sirve para aceptar o rechazar la hipótesis nula establecida sobre el j-esimo parámetro.

Decisión: Si Wald $>\lambda^2_1$ se rechaza H_0 con un nivel de significación de α y se concluye que la variable independiente influye en la probabilidad de las características de la variable dependiente. Si la variable independiente es cualitativa los grados de libertad son iguales al número de categorías menos 1.

3.6 Modelo de Regresión Logística Múltiple

Es una generalización del modelo simple, relaciona la probabilidad de que ocurra un determinado suceso independiente denotado por el vector $x=(x_1,\ldots,x_p)$ con probabilidad condicional P(Y=1/x) en función de p variables independientes que pueden ser cuantitativas, cualitativas o ambos tipos según sea el tipo de diseño de estudio.

El modelo logístico múltiple es:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$$

Estadístico de Wald

Evalúa la significancia de los coeficientes, se define como el vector matriz de los coeficientes estimados, según hipótesis:

H_0: $\beta_i=0$

H_1: Para algún $\beta_i \neq 0$; i=1,….,p

Estadístico de Prueba:

Donde X es la matriz de datos de los productos que se encuentran disponibles o no y V_{nxn} es una matriz diagonal cuyo elemento general es $p_i(1-\hat{p}_i)$.

Siendo I(β) la matriz de información con variancias y covariancias de los coeficientes estimados β_i

$$W = \left(\hat{\beta}'[I(\hat{\beta})]^{-1}\hat{\beta}\right)^2 = \left(\hat{\beta}(X'VX)\hat{\beta}\right)^2 \approx \lambda^2_{(k+1)}$$

La regla de decisión evalúa si se acepta o se rechaza la hipótesis nula con un nivel de significancia determinado.

Razones para la selección de variables

- ✓ La selección de variables predictoras es un proceso estadístico importante porque no todas tienen igual importancia.
- ✓ Algunas variables predictoras pueden perjudicar la confiabilidad del modelo, especialmente si están correlacionadas con otras.
- ✓ Computacionalmente es más fácil trabajar con un conjunto de variables predictoras pequeño.
- ✓ Es más económico recolectar información para un modelo de pocas variables.
- ✓ Si se reduce el número de variables entonces el modelo cumple con el principio de parsimonia.

- ✓ La idea de estos métodos es elegir el mejor modelo en forma secuencial pero incluyendo o excluyendo una sola variable predictoras en cada paso de acuerdo a ciertos criterios.
- ✓ El proceso secuencial termina cuando se satisface una regla de parada establecida.

3.7 Métodos de Selección de Variables

Resumiendo, cabe afirmar que tres son los principales métodos de selección de variables en un modelo. El método manual, el automático y una combinación de ambos. Se estudiara el automático que será plasmado en tres métodos: Allsubsets regressions, Backwar elimination y Forward selection. El objetivo en los tres métodos es el mismo, el establecimiento de un modelo que sobre la base de un mismo conjunto de datos parsimonioso y a la vez, eficiente en la estimación de los coeficientes y en la predicción ajustada de la variable respuesta. Cada uno de estos procedimientos realiza su función por etapas, utilizando un determinado criterio para decidir sobre la inclusión o exclusión de una determinada variable, así como para determinar el momento de finalizar el proceso.

En el caso desarrollado veremos el criterio AIC que es *"an information criterion" (AIC).* ***Akaike's information criterion,*** consiste en obtener la distancia entre dos modelos, en una estimación de la distancia relativa esperada entre el modelo estimado. El AIC sirve para seleccionar el mejor modelo dentro de un conjunto de estos, obtenidos con los mismos datos. En principio el criterio de selección será el escoger los modelos con valores más bajos de AIC.

3.7.1 Métodos Stepwise:

Si se reduce el número de variables entonces el modelo cumple con el principio de la parsimonia.
La idea de estos métodos es elegir el mejor modelo en forma secuencial pero incluyendo o excluyendo una sola variable predictora en cada paso de acuerdo a ciertos criterios.

El proceso secuencial termina cuando se satisface una regla de parada establecida.

Hay tres algoritmos más usados Backward Elimination, Forward Selection y el Stepwise Selección.

3.7.1.1 Stepwise Selección (Selección Paso a Paso)

- ✓ Se empieza con un modelo de regresión simple y en cada paso se puede añadir una variable en forma similar al método forward, pero se coteja si alguna de las variables que ya están presentes en el modelo puede ser eliminada.
- ✓ Aquí se usan el F-out y el F-in con F-out > F-in.
- ✓ El proceso termina cuando ninguna de las variables fuera del modelo tiene importancia suficiente como para ingresar al modelo.

3.8 Descripción en R (Algoritmo Stepwise):

La descripción de la selección de un modelo basado en AIC.
Step (object, scope, scale = 0, direction = c("both", "backward", "forward"), steps = 1000, k = 2, ...)

3.9 Random Forests

Es una combinación de árboles predictores, siendo ellos el conjunto de atributos que se tiene que relacionar con mayor fuerza con la variable objetivo. Entonces cada árbol depende de los valores de un vector aleatorio independientemente y con una misma distribución para cada uno de estos. Es una modificación sustancial de bagging que construye una larga colección de árboles no correlacionados y luego los promedia.El algoritmo es usado para inducir un Random Forest que fue desarrollado por Leo Breiman, Adele Cutler. El método combina la idea bagging de Breiman y la selección aleatoria de atributos, introducida independientemente por Ho, Amit y Geman, para construir una colección de árboles de clasificación con una variación controlada. La

selección de un subconjunto aleatorio de atributos es un ejemplo del método Random subspace.

3.10 Algoritmo Bagging

Esta herramienta tiene como base el método Bootstrap, que se constituye por servir para estimar el error estadístico y se fundamenta en el teorema de Glivenko-Cantelli, el cual establece una convergencia casi segura pero asintótica entre una distribución desconocida F (del parámetro de interés) y una empírica F_n, calculada a partir de una muestra cuando $n \rightarrow \infty$, Efron y Tibshirani (1996). Esta herramienta de re-muestreo permite analizar el rendimiento del estadístico de interés o las predicciones bajo una distribución empírica. Las muestras son tomadas con reemplazo lo que puede enmascarar el error estadístico ya que muestras diferentes pueden poseer las mismas respuestas o la misma respuesta varias veces.

A comparación del Boosting, este divide el conjunto de entrenamiento en varios subconjuntos, a partir de los cuales los clasificadores son entrenados, estos subconjuntos pueden ser disjuntos o no. Una de las ventajas de este tipo de sistemas es que pueden ser fácilmente paralelizados, o utilizar fácilmente distintos tipos de clasificadores.
La idea esencial del bagging es promediar muchos modelos ruidosos pero aproximadamente imparciales, y por tanto reducir la variación. Los árboles son los candidatos ideales para el bagging, dado que ellos pueden registrar estructuras de interacción compleja en los datos, y si crece lo suficiente en profundidad, tienen relativamente baja parcialidad. Producto de que los árboles son notoriamente ruidosos, ellos se benefician grandemente al ser promediados.

Cada árbol es construido usando el siguiente algoritmo:

1. Sea N el número de casos de prueba, M es el número de variables en el clasificador.
2. Sea m el número de variables de entrada a ser usado para determinar la decisión en un nodo dado; m debe ser mucho menor que M.

3. Elegir un conjunto de entrenamiento para este árbol y usar el resto de los casos de prueba para estimar el error.
4. Para cada nodo del árbol, se elige aleatoriamente m variables en las cuales se basara la decisión. Calcular la mejor partición a partir de las m variables del conjunto de entrenamiento.

3.11 Definición de Random Forests

Random Forests es una técnica de agregación desarrollada para mejorar la precisión en la clasificación mediante la incorporación de aleatoriedad en la construcción de cada clasificador individual. Esta aleatorización puede introducirse en la partición del espacio (construcción del árbol), así como en la muestra de entrenamiento. Las distintas variantes de Random Forest consisten en cómo se incorpora la aleatoriedad en la construcción.

El algoritmo tiene un parámetro $1 \leq v < d$ entero positivo fijado por el usuario. La raíz del árbol es R_d. En cada paso de la construcción del árbol, una hoja es elegida uniformemente aleatoria de v variables que son seleccionadas uniformemente aleatorias de las d candidatas $x_{(1)}, \ldots, x_{(d)}$. Entre las v variables, se selecciona la partición que minimiza el número de puntos mal clasificados de la muestra de entrenamiento. Teniendo como criterio que cada árbol nos da una clasificación, y nos dicen que árboles "votos" de esa categoría son los que presentan ser los más clasificados en el Random Forest (atributos a clasificarse) .Este criterio es repetido hasta obtener la mejor clasificación.

Las ventajas del Random Forests son:

- ✓ Ser uno de los más certeros algoritmos de aprendizaje disponible. Para muchos sets de datos que se producen de un clasificador preciso.
- ✓ Correr eficientemente en grandes bases de datos.
- ✓ Poder manejar cientos de variables entrantes sin excluir ninguna.
- ✓ Dar estimados de qué variables son importantes en la clasificación.

- ✓ Tener un método eficaz para estimar datos perdidos y mantener la exactitud cuando una gran proporción de los datos está perdida.
- ✓ Computar los prototipos que dan información sobre la relación entre las variables y la clasificación.
- ✓ Computar las proximidades entre los pares de casos que pueden usarse en los grupos, localizando outliers, o (ascendiendo) dando vistas interesantes de los datos.
- ✓ La construcción de los arboles es rápida, al explorar solo unos pocos atributos.
- ✓ Los nodos que tienen la suerte de cazar atributos muy discriminativos compensan a los otros.

3.12 El Algoritmo Boruta

La importancia de la puntuación por sí sola no es suficiente para identificar las correlaciones de significancia entre las variables y el atributo de la decisión.

Breiman asumió que, debido a las bajas correlaciones entre los árboles individuales, la importancia tiene distribución normal y por lo tanto, la puntuación score se puede utilizar para evaluar la importancia de las variancias. Desafortunadamente, se ha demostrado que la suposición de Breiman es falsa y por lo tanto se necesita una referencia que pueda ayudar a discernir los atributos verdaderamente importantes de los atributos no importantes. Para hacer frente a este problema, Breiman estableció los criterios para la selección de los atributos importantes. La importancia de las variables es vista desde el aspecto de clasificación de los árboles, por medio del índice de Gini que se utiliza en el criterio de impureza.

El algoritmo surge de Random Forest y se hace frente a los problemas mediante la adición de más aleatoriedad al sistema. La idea básica es muy sencilla hacemos una copia aleatoria del sistema, para combinar la copia con el original y construir el clasificador en un sistema extendido.

- ✓ Así se construye un sistema extendido, donde se replica cada variable descriptiva. Los valores de estas variables son luego permutados al azar a través

de los objetos, por consiguiente, todas las correlaciones entre las variables replicadas y el atributo tomado son aleatorias para el diseño.

- ✓ Se realizan varias salidas del Random Forest, las variables replicadas se asignan al azar antes de cada salida y por lo tanto la parte aleatoria del sistema es diferente para cada ejecución de Random Forest.

Para cada serie se calcula la importancia de todos los atributos.

- ✓ Los atributos se consideran importantes para una única prueba, si su importancia es mayor que la máxima importancia de todos los atributos asignados aleatoriamente.
- ✓ Se realiza una prueba estadística para todos los atributos. La hipótesis nula es la importancia de la variable que es igual a ***"la importancia máxima de los atributos aleatorios"*** **(MIRA).** La prueba de hipótesis es de dos colas y puede ser rechazada o aceptada.

  Cuando la importancia del atributo es significativamente más alta o mucho más baja que MIRA en cada atributo se tiene en cuenta que tanta importancia tiene el atributo que es superior a MIRA.

La esperanza se estima cuando el número de accesos es N igual a:
E(p) = 0.5N, la desviación estándar es igual a $S(p)=\sqrt{0.25\,N}$ (distribución binomial con p =q=0.5)

- ✓ Las variables que se consideran importantes son aceptadas, cuando el número de accesos es significativamente más alto que el valor esperado, y se considera poco importante (rechazado), cuando el número de accesos es significativamente más bajo que el valor esperado. Es sencillo calcular los límites para aceptar y rechazar las variables para cualquier número de iteraciones con un nivel de confianza deseado.
- ✓ Las variables que se consideran poco importantes se eliminan del sistema de información con sus respectivas aleatorizaciones para una mejor conformidad.

En algunos casos las variables aleatorizadas se pueden mantener en el sistema, que pueden ayudar en la reducción del número de variables consideradas importantes, sin reducir la exactitud del clasificador Random Forest.

3.13 Descripción del Paquete estadístico Boruta en R:

Boruta(x, y, confidence = 0.999, maxRuns = 100, light = TRUE, doTrace = 0, getImp = getImpRf, ...)

El paquete Boruta es un modelo que tiene una serie de datos variables predictoras y un vector de respuesta.

X	Base de entrenamiento con las variables predictoras.
Y	Variable respuesta. Es nominal para procesos de clasificación y numérica para procesos de regresión.
getImp	Argumento utilizado para obtener la importancia de los atributos. Por defecto es getImpRf, un argumento de la librería Randon Forest que genera los scores para estimar la importancia.
confidence	Nivel de confianza, toma por defecto 0.999. Un valor reducido se refleja en la reducción del tiempo de corrida del Boruta.
maxRuns	Es el máximo número para las corridas finales del random forest que puede incrementar el número de atributos provisionales o tentativos.
doTrace	0 significa que se realiza un seguimiento y 1 que se muestra la importancia en reportes consecutivos de la prueba.
Light	Si es TRUE, Boruta se ejecuta en un modo estándar, en donde los atributos se prueban si serán rechazados; si es FALSE, Boruta se ejecuta en modo forzando, en donde todas las variables agregadas en la durante la presente corrida..
	Argumentos adicionales del Random Forest

La función utilizada para obtener la importancia de los atributos es el valor predeterminado *getImpRf*, que sale del paquete Random Forest y recoge las puntuaciones del score del promedio a disminuir en la medida de precisión.

4. METODOLOGIA DE LA INVESTIGACIÓN

4.1 Tipo de investigación

La presente investigación es exploratoria y correlacional. Exploratoria porque consiste de un problema de investigación con pocos estudios preliminares y correlacional ya que tiene la finalidad de medir la relación entre las variables.

4.2 Formulación de hipótesis

4.2.1 Existe un modelo predictivo para detectar los quiebres de stock para los distintos productos del supermercado, mediante sus datos y reportes transaccionales.
4.2.2 Las variables derivadas de las bases de datos de las transacciones del SAP y del Business tienen la capacidad para predecir los faltantes de mercadería en almacén (FMA).
4.2.3 La técnica de selección de variables Boruta tiene la capacidad para brindar un mayor aporte.
4.2.4 La técnica de selección de variables Stepwise (paso a paso) tiene la capacidad para brindar un mayor aporte.

4.3 Identificación de Variables

Variable Dependiente:

La variable que se desea predecir, es disponibilidad para cada Sku (productos del supermercado). Esta sería una variable binaria que vale uno si existe quiebre de stock o cero en caso contrario. Los valores binarios se consideraron a partir de encontrarnos sin stock hasta 3 días a la semana y 12 días al mes, lo cual nos originaria perdidas de venta (los días de la semana con mayores "ventas" que son desde el jueves a domingo).
Durante el mes de Agosto del 2013 se consideró para la variable dependiente: Y

12 días sin stock	12 días con stock
Y= 1 Existe quiebre de stock	Y= 0 No existe quiebre de stock.

Variables Independientes:

Se calcularon las siguientes variables, de transacciones que pasaran a ser variables predictivas para los modelos.

En total considerando todas las variables, obtendríamos 67 variables predictoras entre los días del mes de agosto con ventas semanales y mensuales (0,1), los proveedores si les pertenece el sku y si no les pertenece el sku (0,1), las variables provenientes de las unidades vendidas y las variables de los días con ventas (semanales y mensuales).

Unidades vendidas día anterior $= V_{t-1}$
Unidades vendidas mismo día semana anterior $= V_{t-7}$
Unidades vendidas promedio semana anterior $= \frac{\sum_{s=1}^{7} V_{t-s}}{7}$
Unidades vendidas promedio últimos 30 días $= \frac{\sum_{s=1}^{30} V_{t-s}}{30}$
Crecimiento de las ventas = **Unidades promedio vendidas a diaria, última semana / unidades promedio vendidas a diario último mes.**
Coeficiente de variación de las ventas (semana anterior) $= \frac{\sigma_{semana}}{\mu_{semana}}$
Coeficiente de variación de las ventas (mes anterior) $= \frac{\sigma_{mes}}{\mu_{mes}}$
Historial de quiebres semanal, % días con ventas = **0 últimos7 días**

Historial de quiebres mensual, % días con ventas = **0 últimos 30 días**
Variables dummies para los días de la semana: **Por cada uno de ellos (según el día perteneciente al mes de agosto del 2013).**
Variables dummies para los proveedores de los Skus. **Variable que tomara el valor 1 si un sku pertenece al proveedor X y 0 sino.**

Se desarrolló una base de datos con todas las variables antes mencionadas, es decir para cada día producto se tiene su disponibilidad, demanda del día anterior, etc.
Se normalizaron las variables de acuerdo a la media y varianza de cada Sku. Esto quiere decir que para cada sku se calculó la media y la desviación estándar histórica de las ventas (del mes de agosto), y se normalizo con respecto a los valores de todas las variables de ventas absolutas (unidades vendidas día anterior, semana anterior, mes anterior y mismo día semana anterior).

4.4 Definición Operacional de las Variables:

Las variables a definir la variable dependiente es cualitativa (0 y 1) e independientes son cuantitativas, dummies. De los días solo se colocó uno de los días (todos los días del mes de Agosto, son variables dummies).

Variable Dependiente	Y	Cualitativa
Variable Independiente	D_20130801	Cualitativa
Variable Independiente	VtaDia_N	Cuantitativa
Variable Independiente	VtaDia7_N	Cuantitativa
Variable Independiente	VtaProm7_N	Cuantitativa
Variable Independiente	VtaProm30_N	Cuantitativa
Variable Independiente	DiasCero30	Cuantitativa
Variable Independiente	DiasCero7	Cuantitativa
Variable Independiente	CreciVta	Cuantitativa
Variable Independiente	CoefVarProm7	Cuantitativa
Variable Independiente	CoefVarProm30	Cuantitativa
Variable Independiente	Proveedor_0000004542	Cualitativa
Variable Independiente	Proveedor_0000004780	Cualitativa
Variable Independiente	Proveedor_0000005980	Cualitativa
Variable Independiente	Proveedor_1007082395	Cualitativa
Variable Independiente	Proveedor_1016523694	Cualitativa

Variable Independiente	Proveedor_2010000394	Cualitativa
Variable Independiente	Proveedor_2010003083	Cualitativa
Variable Independiente	Proveedor_2010005205	Cualitativa
Variable Independiente	Proveedor_2010005523	Cualitativa
Variable Independiente	Proveedor_2010006791	Cualitativa
Variable Independiente	Proveedor_2010007372	Cualitativa
Variable Independiente	Proveedor_2010007402	Cualitativa
Variable Independiente	Proveedor_2010008522	Cualitativa
Variable Independiente	Proveedor_2010009545	Cualitativa
Variable Independiente	Proveedor_2010011922	Cualitativa
Variable Independiente	Proveedor_2010012716	Cualitativa
Variable Independiente	Proveedor_2010019079	Cualitativa
Variable Independiente	Proveedor_2010034400	Cualitativa
Variable Independiente	Proveedor_2010208033	Cualitativa
Variable Independiente	Proveedor_2025464087	Cualitativa
Variable Independiente	Proveedor_2039595940	Cualitativa
Variable Independiente	Proveedor_2041737891	Cualitativa
Variable Independiente	Proveedor_2041966599	Cualitativa
Variable Independiente	Proveedor_2045161412	Cualitativa
Variable Independiente	Proveedor_2046597656	Cualitativa
Variable Independiente	Proveedor_2046886208	Cualitativa
Variable Independiente	Proveedor_2050197352	Cualitativa

4.5 Diseño de investigación

El diseño de investigación es transversal correlacional ya que se describirá las relaciones entre dos o más variables en un momento determinado.

4.6 Población y muestra

El estudio se realizó en el área del almacén de un Supermercado ubicada en el distrito de Miraflores. El Supermercado es uno de los más importantes por encontrarse ubicado muy cerca de las oficinas principales.

El Supermercado cuenta con varios sku, por cada uno de los diversos productos en toda la tienda. Solo se estará revisando los productos pertenecientes a las gerencias de Abarrotes Comestibles, Abarrotes no Comestibles y Bebidas.

La investigación se trabajó con una muestra representativa de 1,867 sku, obteniendo todos los datos requeridos y ya mencionados, se analizaron los datos por un intervalo de 1 mes, para los historiales de quiebres mensual y semanal se tomaron los 30 días y 7 días respectivamente sin venta para colocar cero o 1 y formar las variables respectivas.

4.7 Instrumento de Colecta de Datos

El instrumento empleado en la obtención de la información requerida, para la investigación fueron salidas en Sap y en Bussiness por medio de transacciones de ventas, históricos de stock, productos de proveedores, productos con stock cero (quiebres).

5. PROCEDIMIENTO DE ANÁLISIS DE DATOS

Para realizar la obtención de los datos se utilizó el Sap y el Business. Luego para el análisis exploratorio se utilizó el Excel, para conseguir las medias, variancias y pasar después a normalizarlas (las VtaDia_N, VtaDia7_N, VtaPrm7_N y la VtaProm30_N). Las variables dummies se evaluaron por medio de fórmulas del mismo programa para ser transformadas de numéricas a dummies, por criterio del negocio se consideró la Y=1 (en quiebre), si se encuentra más de 12 días al mes sin despacho de sus productos, mes de Agosto. Se consideraron con venta de ese día con valor 1 (quiebre) y si no tuvo venta lo contrario (sin quiebre).

t: considerado el día de inicio: El 31 de agosto del 2013.
Las variables de forma más detallada:

$Vta_{(t)}$: Las ventas registradas el día 31 de agosto.
$VtaProm_{(t-7)}$: Suma de las ventas de una semana/7
$VtaProm_{(t-30)}$: Suma de las ventas del mes/ 30
Historial de quiebre semanal (%): $\frac{\text{Numero de días sin venta por una semana}}{7}$ (valores ceros)
Historial de quiebre al mes (%): $\frac{\text{Numero de días sin venta por un mes}}{30}$ (valores ceros)
CreciVeta: $\frac{\text{VtaProm7}}{\text{VtaProm3(}}$

Las variables de los días del mes, proveedores de los productos revisados son valores 0 y 1 sea el caso y los coeficientes de variancias de las medias y variancias, ya realizadas para las normalizaciones de las variables. Con toda la información recopilada en Excel, se realizó la selección de variables en el software R, previa instalación de las library("mlbench"), library("randomForest") y library("Boruta") para el Boruta.

Teniendo las dos selecciones de variables por Boruta, por Stepwise se consideró el ajuste de los modelos en la regresión logística, con el software SPSS.

6. RESULTADOS

6.1 Paquete estadístico Boruta:

```
Buru<- Boruta(Y~.,data=Data[,-1],doTrace=2,ntree=500)
plot(Buru,las=3)
```

Gráfico N.1: Boxplot

Interpretación:

Se puede observar por medio de un gráfico de cajas. Las azules sonlas que presentan un mínimo puntaje promedio y un máximo puntaje promedio (z-scores) y los rojos y verdes representan respectivamente los atributos rechazados y aceptados.

6.2 Paquete estadístico Stepwise:

Mod_1 <- glm(Y ~ . ,family = binomial)

Seleccion <- step(Mod_1,direction = c("both"))

summary(Seleccion)

Call:

glm(formula = Y ~ D_20130801 + D_20130808 + D_20130812 + D_20130813 + ...+ D_20130815, family = binomial)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.7927	-0.7559	-0.4265	0.6642	3.0783

Coefficients:

Coef	Estimate	Std. Error z	value	Pr(>|z|)	
(Intercept)	-2.8121	0.6179	-4.551	5.34E-06	***
D_20130801	-0.9926	0.2792	-3.556	0.000377	***
D_20130808	-0.5892	0.2996	-1.966	0.049244	*
D_20130812	0.5982	0.3312	1.806	0.070878	.
D_20130813	-0.5593	0.3067	-1.824	0.068207	.
D_20130814	-0.8219	0.3318	-2.477	0.013259	*
D_20130816	0.6255	0.3209	1.949	0.051297	.
D_20130822	0.8645	0.3175	2.723	0.00647	**
D_20130823	0.9573	0.3144	3.045	0.002327	**
VtaDia_N	6.174	1.4511	4.255	2.09E-05	***
VtaDia7_N	13.7472	1.8516	7.424	1.13E-13	***
VtaProm7_N	-23.3742	2.9222	-7.999	1.25E-15	***
CoefVarProm30	-0.3122	0.1708	-1.829	0.067474	.
Proveedor_0000004542	-0.6591	0.3079	-2.14	0.032331	*
Proveedor_0000005980	1.0321	0.4967	2.078	0.037703	*
Proveedor_2010000394	-0.4182	0.2337	-1.789	0.073537	.
Proveedor_2010005523	-1.3739	0.7657	-1.794	0.072777	.

Proveedor_2010007372	0.516	0.28	1.843	0.065314	.
Proveedor_2010007402	-0.4381	0.2502	-1.751	0.079921	.
Proveedor_2010008522	0.5791	0.3222	1.797	0.072277	.
Proveedor_2010009545	-0.987	0.3409	-2.895	0.003787	**
Proveedor_2010011922	0.5865	0.3104	1.889	0.058855	.
Proveedor_2010012716	0.9752	0.2064	4.724	2.31E-06	***
Proveedor_2010019079	0.9315	0.3084	3.02	0.002525	**
Proveedor_2010034400	-0.6369	0.3997	-1.594	0.111029	
Proveedor_2010208033	2.0064	0.9471	2.119	0.034129	*
Proveedor_2025464087	2.0059	0.7188	2.791	0.005263	**
Proveedor_2039595940	-15.9497	366.8911	-0.043	0.965325	
Proveedor_2041737891	-15.9915	792.0164	-0.02	0.983891	
Proveedor_2041966599	-15.9789	603.7753	-0.026	0.978886	
Proveedor_2046597656	-0.8237	0.2861	-2.88	0.003983	**
Proveedor_2046886208	0.8186	0.2479	3.303	0.000957	***
D_20130830	0.7271	0.355	2.234	0.025499	*
D_20130815	0.4526	0.3083	1.468	0.142049	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2133.2 on 1860 degrees of freedom

Residual deviance: 1689.5 on 1827 degrees of freedom

AIC: 1757.5

Number of Fisher Scoring iterations: 15

6.3 Selección de Variables:

Variables seleccionadas de ambos modelos (Boruta y Stepwise). Se seleccionaron por Boruta 28 variables y por Stepwise 33 variables.

Stepwise	Boruta
CoefVarProm30	CoefVarProm30
D_20130801	D_20130801
D_20130808	D_20130808
D_20130816	D_20130816
D_20130822	D_20130822
D_20130823	D_20130823
Proveedor_0000004542	Proveedor_0000004542
Proveedor_2010008522	Proveedor_2010008522
Proveedor_2010009545	Proveedor_2010009545
Proveedor_2010012716	Proveedor_2010012716
Proveedor_2010019079	Proveedor_2010019079
Proveedor_2025464087	Proveedor_2025464087
Proveedor_2039595940	Proveedor_2039595940
Proveedor_2046886208	Proveedor_2046886208
VtaDia_N	VtaDia_N
VtaDia7_N	VtaDia7_N
VtaProm7_N	VtaProm7_N
D_20130812	D_20130812
D_20130813	D_20130813
Proveedor_0000005980	Proveedor_0000005980
Proveedor_2010007372	Proveedor_2010007372
Proveedor_2010208033	Proveedor_2010208033
D_20130814	
Proveedor_2010000394	
Proveedor_2010005523	
Proveedor_2010007402	
Proveedor_2010011922	
Proveedor_2010034400	
Proveedor_2041737891	
Proveedor_2041966599	
Proveedor_2046597656	
D_20130830	
D_20130815	

Aplicando Regresión Logística en la selección de variables por Boruta:

	B	E.T.	Wald	gl	Sig.	Exp(B)	I.C. 95% para EXP(B)	
							Inferior	Superior
CoefVarProm30	-.419	.165	6.413	1	.011	**.658**	.476	.910
D_20130801(1)	1.003	.271	13.738	1	.000	2.727	1.604	4.635
D_20130808(1)	.607	.281	4.680	1	.031	1.835	1.059	3.181
D_20130816(1)	-.673	.307	4.817	1	.028	.510	.280	.931
D_20130822(1)	-.795	.303	6.893	1	.009	.451	.249	.817
D_20130823(1)	-.994	.305	10.639	1	.001	.370	.204	.673
Proveedor_0000004542(1)	.367	.294	1.560	1	.212	1.444	.811	2.569
Proveedor_2010008522(1)	-.832	.308	7.286	1	.007	.435	.238	.796
Proveedor_2010009545(1)	.659	.326	4.087	1	.043	1.932	1.020	3.660
Proveedor_2010012716(1)	-1.252	.186	45.201	1	.000	.286	.199	.412
Proveedor_2010019079(1)	-1.207	.294	16.854	1	.000	.299	.168	.532
Proveedor_2025464087(1)	-2.322	.726	10.234	1	.001	.098	.024	.407
Proveedor_2039595940(1)	20.335	6182.617	.000	1	.997	6.7e8	.000	.
Proveedor_2046886208(1)	-1.024	.231	19.647	1	.000	.359	.228	.565
VtaDia_N	6.062	1.339	20.503	1	.000	**429.098**	31.121	5916.414
VtaDia7_N	13.720	1.895	52.406	1	.000	9.3e5	2.2e5	3.1e6
VtaProm7_N	-23.412	2.823	68.782	1	.000	.000	.000	.000
D_20130812(1)	-.489	.314	2.416	1	.120	.613	.331	1.136
D_20130813(1)	.601	.291	4.280	1	.039	1.825	1.032	3.226
Proveedor_0000005980(1)	-1.306	.489	7.143	1	.008	.271	.104	.706
Proveedor_2010007372(1)	-.815	.265	9.444	1	.002	.443	.263	.744
Proveedor_2010208033(1)	-2.263	.944	5.750	1	.016	.104	.016	.662
Constante	-12.138	6182.617	.000	1	.998	.000		

Interpretaciones:

Y=0; se encuentra disponible el producto en el almacén. (no quiebre).
Y=1; no se encuentra disponible el producto en el almacén. (quiebre).

CoefVarProm30: $e(\beta n)$=**0.658**

Ante un incremento en una unidad de medida del coeficiente de variación promedio mensual que provocara un incremento multiplicativo por un factor de 0.658de la ventaja que no se encuentre disponible el producto en el almacén (quiebre).

VtaDia_N: 429.098

Ante un incremento en una unidad de medida de la venta diaria normalizada que provocara un incremento multiplicativo por un factor de 429.098 de la ventaja de queno se encuentre disponible el producto en el almacén (quiebre).

Variables Dummies:

X= 0; no se vende productos durante un día de la semana.
X= 1; si vende productos durante un día de la semana.

D_20130801(1): $e(\beta n)$:2.727

En el caso del día20130801, el valor estimado es de 2.727 significa que manteniendo constante el resto de las variables, el no encontrar disponibilidad del producto en el almacén es de 2.727 veces más desventajoso de que se vendan los productos durante ese día de la semana que de otros días de la semana que no se venden.

Proveedor_0000005980 (1): $e^{(\beta n)}$: .271

X= 0; al proveedor no le pertenece el sku.

X= 1; al proveedor le pertenece el sku.

En el caso del proveedor, el valor estimado de .271 significa que, manteniendo constante el resto de las variables, el no encontrar disponibilidad del producto en el almacén es .271 veces más desventajoso que al proveedor que pertenece el sku que otros proveedores.

Tabla de clasificación				
		Pronosticado		
		Y		
Observado		0	1	%
Y	0	1283	94	93.2
	1	296	188	38.8
	Total			79.0

Y=0; se encuentra disponible el producto en el almacén. (no quiebre).

Y=1; no se encuentra disponible el producto en el almacén. (quiebre).

1,283 +94 productos que se encuentran disponibles en el almacén, 1,283 han sido pronosticados como productos que no se encuentran quebrados, es decir un porcentaje de aciertos del 93.2%.

De las 296+188 productos que no se encuentran disponibles en el almacén, 188 han sido pronosticados como productos quebrados, es decir un porcentaje de aciertos es de 38.8%.El porcentaje global de aciertos es del 79.0%.

Regresión Logística con Selección de variables por Stepwise:

	B	E.T.	Wald	Gl	Sig.	Exp(B)	I.C. 95% para EXP(B)	
							Inferior	Superior
CoefVarProm30	-0.312	0.171	3.343	1	0.067	0.732	0.524	1.023
D_20130801(1)	0.993	0.279	12.644	1	0	2.698	1.561	4.663
D_20130808(1)	0.589	0.3	3.867	1	0.049	1.803	1.002	3.243
D_20130816(1)	-0.626	0.321	3.799	1	0.051	0.535	0.285	1.004
D_20130822(1)	-0.865	0.317	7.415	1	0.006	0.421	0.226	0.785
D_20130823(1)	-0.957	0.314	9.272	1	0.002	0.384	0.207	0.711
Proveedor_0000004542(1)	0.659	0.308	4.581	1	0.032	1.933	1.057	3.535
Proveedor_2010008522(1)	-0.579	0.322	3.231	1	0.072	0.56	0.298	1.054
Proveedor_2010009545(1)	0.987	0.341	8.383	1	0.004	2.683	1.376	5.234
Proveedor_2010012716(1)	-0.975	0.206	22.317	1	0	0.377	0.252	0.565
Proveedor_2010019079(1)	-0.932	0.308	9.122	1	0.003	0.394	0.215	0.721
Proveedor_2025464087(1)	-2.006	0.719	7.787	1	0.005	0.135	0.033	0.55
Proveedor_2039595940(1)	20.602	6192.076	0	1	0.997	8.50E+08	0	.
Proveedor_2046886208(1)	-0.819	0.248	10.908	1	0.001	0.441	0.271	0.717
VtaDia_N	6.174	1.451	18.103	1	0	480.09	27.936	8250.449
VtaDia7_N	13.747	1.852	55.122	1	0	9.30E+05	2.40E+05	3.40E+06
VtaProm7_N	-23.374	2.922	63.984	1	0	0	0	0
D_20130812(1)	-0.598	0.331	3.263	1	0.071	0.55	0.287	1.052
D_20130813(1)	0.559	0.307	3.326	1	0.068	1.749	0.959	3.191
Proveedor_0000005980(1)	-1.032	0.497	4.318	1	0.038	0.356	0.135	0.943
Proveedor_2010007372(1)	-0.516	0.28	3.397	1	0.065	0.597	0.345	1.033
Proveedor_2010208033(1)	-2.006	0.947	4.488	1	0.034	0.134	0.021	0.861
D_20130814(1)	0.822	0.332	6.134	1	0.013	2.275	1.187	4.359
D_20130815(1)	-0.453	0.308	2.156	1	0.142	0.636	0.348	1.164
D_20130830(1)	-0.727	0.325	4.99	1	0.025	0.483	0.255	0.915

Proveedor_2010000394(1)	0.418	0.234	3.202	1	0.074	1.519	0.961	2.402
Proveedor_2010005523(1)	1.374	0.766	3.219	1	0.073	3.951	0.881	17.721
Proveedor_2010007402(1)	0.438	0.25	3.067	1	0.08	1.55	0.949	2.53
Proveedor_2010011922(1)	-0.587	0.31	3.569	1	0.059	0.556	0.303	1.022
Proveedor_2010034400(1)	0.637	0.4	2.54	1	0.111	1.891	0.864	4.139
Proveedor_2041737891(1)	20.687	13663.331	0	1	0.999	9.60E+08	0	.
Proveedor_2041966599(1)	20.608	10076.144	0	1	0.998	8.70E+08	0	.
Proveedor_2046597656(1)	0.824	0.286	8.292	1	0.004	2.279	1.301	3.992
Constante	-59.333	18070.91	0	1	0.997	0		

Y=0; se encuentra disponible el producto en el almacén. (no quiebre).

Y=1; no se encuentra disponible el producto en el almacén. (quiebre).

CoefVarProm30: $e(\beta n) = 0.732$

Interpretaciones:

Ante un incremento en una unidad de medida del coeficiente de variación promedio mensual que provocara un incremento multiplicativo por un factor de 0.732 de la ventaja de que no se encuentre disponible el producto en el almacén (quiebre).

VtaDia_N: $e(\beta n) = 480.090$

Ante un incremento en una unidad de medida en la venta diaria normalizada que provocara un incremento multiplicativo por un factor de 480.090 de la ventaja de que no se encuentre disponible el producto en el almacén (quiebre).

Variables Dummies:

D_20130801(1): e(βn):2.698

X= 0; no se vende productos durante un día de la semana.
X= 1; si vende productos durante un día de la semana.

En el caso del día20130801, el valor estimado de 2.698 significa que, manteniendo constantes el resto de las variables, el no encontrar disponibilidad del producto en el almacén es 2.698 veces más desventajoso de que se venda productos durante ese día de la semana que de otros días de la semana que no se vende.

Proveedor_0000004542(1): e(βn):1.933

X= 0; al proveedor no le pertenece el sku.
X= 1; al proveedor le pertenece el sku.

En el caso del proveedor, el valor estimado de 1.933significa que, manteniendo constante el resto de las variables, el no encontrar disponibilidad del producto en el almacén es 1.933 veces más desventajoso que al proveedor que pertenece el sku que otros proveedores.

Tabla de clasificación				
		Pronosticado		
		Y		
Observado		0	1	%
Y	0	1265	112	91.9
	1	283	201	41.5
	Total			78.8

Y=0; se encuentra disponible el producto en el almacén. (no quiebre).
Y=1; no se encuentra disponible el producto en el almacén. (quiebre).

1,265 + 112 productos se encuentran disponibles en el almacén, 1,265 han sido pronosticados como productos no quebrados, es decir un porcentaje de aciertos del 91.9%.

De las 283+201 productos que no se encuentran disponibles en el almacén, 201 han sido pronosticados como productos quebrados, es decir un porcentaje de aciertos es de 41.5%.El porcentaje global de aciertos es del 78.8%.

7. CONCLUSIONES

1. Se puede concluir que la selección de variables por Boruta en el paquete R, nos brinda un menor número de variables seleccionadas, lo cual reduciría el costo de inversión por analizar a profundidad dichas variables. Podemos decir entonces que el mejor modelo es el Boruta por el menor número de variables seleccionadas que ayudaran a predecir los faltantes de mercadería en almacén (FMA). Ahorrando costo e inversión de tiempo siendo por lo mismo el modelo que más se ajusta a un modelo predictivo para el quiebre de stock en un supermercado.

2. Las variables derivadas de las bases de datos transaccionales (Sap y Business) si se encuentran en la capacidad de brindarnos aportes para el modelo predictivo.

3. La selección de variables por Stepwise, no se estaría considerando ya que presenta un mayor número de variables, lo cual ocasiona una mayor inversión.

4. Las dos técnicas brindaron aporte. Pero el que presenta una menor inversión y tiempo de trabajo al recolectar la información es la técnica Boruta.

5. La tasa de respuesta en ambos modelos es casi la misma, sin embargo el algoritmo Boruta es más parsimonioso con solo 28 variables seleccionadas.

6. Las variables seleccionadas nos apoyaran en saber que productos se encontraran disponibles o no se encuentran en almacén para tenerlas presentes en las gestiones que realizan los compradores con la finalidad de no presentar faltantes o que el cliente presente una opción alternativa hacia algún faltante. Son gestiones que a largo plazo reducirían el costo de las ventas perdidas y no solo eso; sino la posible pérdida de un cliente al tener que dirigirse a un Supermercado de la competencia por la falta de mercadería en los ambientes de venta requeridos.

8. RECOMENDACIONES

1. Se podría analizar a partir de la investigación realizada a otras gerencias como la gerencia de bazar (librería, juguetería, etc.).

2. Dependiendo del estudio que se realice se pueden aplicar estudios de otros árbolesde clasificación como el CART, el CHAID.

9. REFERENCIAS BIBLIOGRÁFICAS

- ✓ Agustín Alonso Rodríguez (2005), Selección de variables en el modelo lineal nuevos procedimientosCentro Universitario «Escorial-María Cristina» San Lorenzo del Escorial.

- ✓ Leo Breiman and Adele Cultler, Random Forest, classification and clustering

- ✓ Miron B. Kursa and Witold R. Rudnicki Package 'Boruta' August 29, 2013

- ✓ Miron, B. Kursa (2010), Feature Selection with the Boruta Package, Journal of Statistical Software.

- ✓ Sebastián Castro, Análisis de Datos en grandes dimensiones. Estimación y selección de variables en regresión.

10. ANEXOS

Salida delBaruta

setwd("D:/TesinaBeatriz/TesinaDatos")

Data<-read.csv("DataR_BEA.csv",header=T)

set.seed(1)

Data<-na.omit(Data)

Buru <- Boruta(Y~.,data=Data[,-1],doTrace=2,ntree=500)

Initial round 1: .....

8 attributes rejected after this test: D_20130814 Proveedor_0000004780 Proveedor_1016523694 Proveedor_2010003083 Proveedor_2010005205 Proveedor_2010005523 Proveedor_2010006791 Proveedor_2010034400

Initial round 2: .....

1 attributes rejected after this test: D_20130817

Initial round 3: .....

7 attributes rejected after this test: D_20130818 D_20130819 D_20130825 D_20130826 D_20130828 D_20130831 Proveedor_2010000394

Final round: .....

19 attributes confirmed after this test: D_20130801 D_20130805 D_20130806 VtaDia_N VtaDia7_N VtaProm7_N VtaProm30_N DiasCero30 CreciVta CoefVarProm7 CoefVarProm30 Proveedor_0000004542 Proveedor_2010008522 Proveedor_2010009545 Proveedor_2010012716 Proveedor_2010019079 Proveedor_2025464087 Proveedor_2039595940 Proveedor_2046886208

1 attributes rejected after this test: D_20130815

....

2 attributes rejected after this test: D_20130824 Proveedor_2045161412

....

1 attributes confirmed after this test: DiasCero7

3 attributes rejected after this test: D_20130820 D_20130830 Proveedor_1007082395

...

1 attributes confirmed after this test: D_20130822

3 attributes rejected after this test: D_20130810 D_20130811 D_20130829

...

1 attributes rejected after this test: Proveedor_2010007402

.....

2 attributes confirmed after this test: D_20130802 D_20130807

...

1 attributes confirmed after this test: D_20130808

..

2 attributes confirmed after this test: D_20130816 D_20130821

...

1 attributes confirmed after this test: D_20130823

.....

1 attributes confirmed after this test: D_20130809

.....

2 attributes rejected after this test: Proveedor_2041737891 Proveedor_2046597656

.....

1 attributes rejected after this test: Proveedor_2041966599

.....

2 attributes rejected after this test: D_20130827 Proveedor_2050197352

.....

1attributes rejected after this test: Proveedor_2010011922

.....

There were 50 or more warnings (use warnings() to see the first 50)

Grafico del resumen de las variables confirmadas, no confirmadas y tentativas.

attStats(Buru)

Variables	meanZ	medianZ	minZ	maxZ	normHits	decision
D_20130801	11.599	11.593	9.51	13.057	0.992	Confirmed
D_20130802	5.501	5.488	3.435	7.105	0.838	Confirmed
D_20130803	3.932	4.003	1.507	5.193	0.531	Tentative
D_20130804	3.571	3.572	2.428	4.857	0.392	Tentative
D_20130805	7.331	7.337	6.085	8.376	0.962	Confirmed
D_20130806	7.376	7.386	5.722	9.075	0.954	Confirmed
D_20130807	5.084	5.089	3.925	6.485	0.808	Confirmed
D_20130808	4.947	4.896	3.52	6.476	0.754	Confirmed
D_20130809	4.786	4.768	2.271	7.556	0.708	Confirmed
D_20130810	2.511	2.495	0.667	4.974	0.046	Rejected
D_20130811	2.688	2.721	-0.072	4.802	0.054	Rejected
D_20130812	3.425	3.499	1.905	5.316	0.338	Tentative
D_20130813	3.344	3.36	0.165	5.788	0.354	Tentative
D_20130814	0.445	0.633	-1.577	1.194	0	Rejected
D_20130815	2.837	2.668	1.3	4.144	0.038	Rejected
D_20130816	4.816	4.774	3.353	6.475	0.731	Confirmed
D_20130817	1.269	1.286	-0.435	2.375	0	Rejected
D_20130818	1.698	1.89	0.232	2.454	0	Rejected
D_20130819	1.98	2.017	0.133	3.505	0	Rejected
D_20130820	2.699	2.665	0.799	4.009	0.038	Rejected
D_20130821	4.881	4.866	3.43	6.485	0.731	Confirmed
D_20130822	5.404	5.391	3.767	7.148	0.831	Confirmed
D_20130823	4.521	4.503	2.981	5.759	0.7	Confirmed
D_20130824	2.316	2.402	0.372	3.482	0.015	Rejected
D_20130825	0.018	0.131	-2.676	2.08	0	Rejected
D_20130826	2.004	2.092	-0.251	3.274	0	Rejected
D_20130827	3.029	3.065	0.645	5.179	0.215	Rejected
D_20130828	1.908	2.088	0.378	2.931	0	Rejected
D_20130829	2.504	2.529	0.55	4.051	0.038	Rejected
D_20130830	2.745	2.77	1.675	3.699	0.038	Rejected
D_20130831	1.981	1.9	0.559	3.253	0	Rejected
VtaDia_N	20.975	20.899	19.36	22.973	1	Confirmed
VtaDia7_N	21.46	21.423	18.528	24.151	1	Confirmed
VtaProm7_N	28.377	28.256	25.961	30.725	1	Confirmed
VtaProm30_N	33.926	33.763	31.086	36.836	1	Confirmed

DiasCero30	15.709	15.729	14.203	17.425	1	Confirmed
DiasCero7	6.943	6.881	5.352	8.566	0.931	Confirmed
CreciVta	12.664	12.619	10.134	15.345	1	Confirmed
CoefVarProm7	16.686	16.683	13.499	19.349	1	Confirmed
CoefVarProm30	26.15	26.027	23.477	29.383	1	Confirmed
Proveedor_0000004542	12.719	12.576	9.399	16.327	1	Confirmed
Proveedor_0000004780	-2.37	-2.273	-4.116	-1.155	0	Rejected
Proveedor_0000005980	4.153	4.042	1.155	8.526	0.5	Tentative
Proveedor_1007082395	1.864	2.073	-0.607	4.949	0.031	Rejected
Proveedor_1016523694	-0.716	-0.505	-2.976	1.328	0	Rejected
Proveedor_2010000394	1.584	1.714	-1.032	3.812	0	Rejected
Proveedor_2010003083	-0.754	-0.802	-2.113	1.548	0	Rejected
Proveedor_2010005205	-1.616	-1.742	-2.927	-0.434	0	Rejected
Proveedor_2010005523	-0.757	-0.243	-2.733	1.145	0	Rejected
Proveedor_2010006791	-0.09	0.089	-1.486	1.026	0	Rejected
Proveedor_2010007372	4.505	4.441	1.637	7.815	0.623	Tentative
Proveedor_2010007402	2.773	2.732	0.226	5.1	0.062	Rejected
Proveedor_2010008522	7.49	7.501	4.231	10.373	0.931	Confirmed
Proveedor_2010009545	10.218	10.223	7.28	12.956	0.992	Confirmed
Proveedor_2010011922	3.218	3.366	-1.226	6.359	0.308	Rejected
Proveedor_2010012716	24.288	24.289	21.007	28.734	1	Confirmed
Proveedor_2010019079	8.518	8.571	4.435	12.254	0.985	Confirmed
Proveedor_2010034400	-1.213	-1.114	-2.759	0.41	0	Rejected
Proveedor_2010208033	4.26	4.347	2.207	6.56	0.585	Tentative
Proveedor_2025464087	8.9	8.873	6.491	11.516	0.985	Confirmed
Proveedor_2039595940	17.461	17.599	14.546	19.971	1	Confirmed
Proveedor_2041737891	3.114	3.194	1.019	4.63	0.177	Rejected
Proveedor_2041966599	2.895	2.911	1.088	4.639	0.177	Rejected
Proveedor_2045161412	1.217	1.387	-2.265	3.961	0.031	Rejected
Proveedor_2046597656	3.046	3.059	0.38	5.048	0.154	Rejected
Proveedor_2046886208	20.238	20.395	16.664	24.143	1	Confirmed
Proveedor_2050197352	3.088	2.915	0.318	6.521	0.231	Rejected

Se puede observar todas las variables seleccionadas "Confirmed" y las no seleccionadas "Rejected"y también estimadores como meanZ, medianZ, minZ,maxZ y normHits.

Para la descripción en R (Algoritmo Stepwise):

La descripción de la selección de un modelo basado en AIC.

Step (object, scope, scale = 0, direction = c("both", "backward", "forward"), steps = 1000, k = 2, ...)

Entre los argumentos utilizados tenemos *object* que es la representación de un modelo lineal o lineal generalizado ("lm" y "GLM"), el cual es utilizado como el modelo inicial en la búsqueda paso a paso de las variables a ser seleccionadas. El argumento *scope* es la búsqueda paso a paso que define una gama de modelos, esto aparece a partir de una formula o de una lista que contiene los componentes superiores (upper) o inferiores (lower) o ambos. El *scale* es utilizado en la definición de la estadística de AIC para la selección de los modelos *lm*, *aov* y *glm*, el valor predeterminado con 0 que indica una escala que debería ser estimada por extractAIC.La *direction* es el método de búsqueda por pasos, puede usar los métodos hacia atrás (*backward*), hacia adelante (*forward*) o un valor predeterminado de ambos (*both*).El *steps* es para tener el número máximo de pasos, el valor predeterminado es 1000. Se utiliza típicamente para detener el proceso antes de tiempo y por ultimo tenemos el *k* que es el múltiplo del número de grados de libertad que se utilizan para la penalidad, con un k = 2 se obtieneAIC: k = log (n).

✓ Para el Stepwise:

Data<-read.csv("DataR_BEA.csv",header=T)

attach(Data)

✓ Para el Boruta:

setwd("D:/TesinaBeatriz/TesinaDatos")

Data<-read.csv("DataR_BEA.csv",header=T)

set.seed(1)

Data<-na.omit(Data)