Grado en Estadística

TRABAJO FIN DE GRADO

Modelización estadística de ventas en el sector retail

Marta Venegas Pardo

Sevilla, Noviembre de 2021

Índice general

Prólogo

Escrito colocado al comienzo de una obra en el que se hacen comentarios sobre la obra o su autor, o se introduce en su lectura; a menudo está realizado por una persona distinta del autor.

También se podrían incluir aquí los agradecimientos.

Resumen

Resumen...

Aquí comienza mi resumeeenennenenenenenenneennenemn

Abstract

Abstract...

Introduccion

Esta es la introducción de mi trabajo

- Definición

El sector retail, o también conocido como comercio minorista, es un sector económico que se encarga de proveer de bienes y servicios al consumidor final, ya que integra al grueso de comerciantes de un determinado lugar.

Este sector aglutina a comerciantes y empresas encargadas de la comercialización, ofreciendo de una gran variedad de productos y servicios a los consumidores.

(Una tienda, un supermercado, una librería, son claros ejemplos de lo que es el sector retail.)

- Objetivos
- Estimación de la demanda
- Asociación entre los dos productos

Índice de figuras

Índice de tablas

Capítulo 1

La ciencia de datos en el sector retail

Modelización estadística de ventas en el sector retail

- Objetivo: predicción de ventas, es decir, predicción de la demanda de productos

1.1. Big Data en el sector retail (COMO TITULAR)

Con R...

Capítulo 2

Análisis de cesta de la compra (Market basket análisis)

Este capítulo está por ver

2.1. Introducción

2.2. Definición

2.3. Desarrollo

Capítulo 3

Modelos estadísticos

A continuación se exponen los modelos estadísticos que estudiaremos de cara a predecir las ventas de los productos.

3.1. Análisis de series temporales

Aplicaremos este modelo de predicción para tratar de identificar los patrones de la demanda anterior a lo largo del tiempo y luego proyectar (predecir) los patrones en el futuro.

Se define una serie temporal como una sucesión de datos ordenados en el tiempo que corresponden a una misma variable. Los datos son suelen ser tomados en intervalos regulares de tiempo.

Nuestro objetivo dentro del análisis de series temporales será identificar el proceso estocástico que ha sido capaz de generar la serie de estudio.

(No se si añadirlo) Se dice proceso estocástico a una colección o familia de variables aleatorias $\{X_t, \text{ con } t \in T\}$ que siguen la misma ley de distribución y están relacionadas entre sí, pudiendo por este motivo, describir la información de estas variables en términos de medias, variaciones y covarianzas.

A continuación encontramos las cuatro etapas en un análisis descriptivo de series temporales para elegir un modelo que se adecue a nuestros datos:

- **Representación gráfica de la serie**. Para tener así una primera aproximación del comportamiento de la serie y la existencia de posibles tendencias.
- **Modelización**: Se trata de encontrar el modelo que mejor se ajuste a los datos.
- **Validación de los modelos**: Es necesario saber si el modelo ajustado es adecuado o no, por lo que es muy importante el estudio de los residuos.
- **Predicciones**: Una vez construido y validado un modelo, realizaremos estimaciones del futuro con nuevas observaciones.

En un enfoque clásico de series temporales, asumiremos que el comportamiento de la variable con respecto al tiempo se compone de cuatro componentes:

1. **Tendencia**: Se trata del movimiento suave y regular de la serie a largo plazo. La tendencia existe cuando hay un aumento o disminución a largo plazo de los datos.

Puede ser lineal (ajuste mediante una recta) o no lineal (aproximación mediante una curva, como por ejemplo logarítmica o exponencial)

2. **Ciclo**: Componente de tipo oscilante caracterizada por movimientos recurrentes en torno a la tendencia de la serie y que se repiten cada año pero sin una frecuencia fija.
3. **Componente estacional**: Se trata de movimientos regulares dentro de la serie con una periodicidad menor a un año, es decir, aquello que ocurre generalmente y con la misma intensidad año tras año en los mismos períodos, por ejemplo, en la misma época del año o día de la semana. Vamos a denotar por L al número de estaciones.
4. **Componente irregular**: Se trata de las variaciones de la serie sin un comportamiento sistemático y que no son explicadas por las otras tres componentes

Existen diferentes modelos de combinación de las componentes. Para describir los modelos necesitamos primero una nomenclatura básica. Denotando por X_t al valor de la variable en el instante t, se tiene:

$$X_t = f(T_t, E_t, I_t)$$

donde:

- T_t: Valor de la tendencia en el instante t
- E_t: Valor de la componente estacional en el instante t
- I_t: Valor de la componente irregular en el instante t (ruido).

Por tanto, los modelos que puede adoptar la función f son los siguientes:

- **Modelo multiplicativo**: La composición de la serie se realiza mediante el producto de sus componentes.
$$X_t = T_x \times E_t \times I_t$$
- **Modelo aditivo**: Las componentes se agregan para formar la serie temporal.
$$X_t = T_x + E_t + I_t$$
- **Modelo mixto**: La composición de la serie de la parte irregular viene de forma aditiva y la parte regular de forma multiplicativa.
$$Xt = Tx \times Et + It$$

http://www5.uva.es/estadmed/datos/series/series2.htm

Tras haber detectado el modelo mas adecuado, podremos conocer el comportamiento de la serie a largo plazo.

El siguiente paso realizar una estimación de la tendencia, T_t, habiendo eliminado previamente la componente estacional para impedir que estas oscilaciones perturben la identificación de la tendencia.

Para estimar T_t, debemos hacer una hipótesis sobre su forma:

- **Tendencia determinista**: Se supone que la tendencia es una función determinística del tiempo:
$$T_t = a + bt \quad a, b \in \mathbb{R}$$

Siendo a y b constantes, que se estimarán mediante un modelo de regresión lineal.

Sin embargo, el método que aplicaremos será el que exponemos a continuación:

- **Tendencia evolutiva (método de medias móviles)**: Este método consiste en definir la tendencia como una serie suavizada. Sunpondremos que la tendencia de la serie es una función que evoluciona lentamente y que podremos aproximar función simple del tiempo, suponiendo así una recta.

Una vez identificada la tendencia, procedemos a hacer un análisis de la estacionalidad de la serie, con el objetivo de:

- **Desestacionalizar la serie**, es decir, eliminar las oscilaciones periodicas que se repiten a lo largo de los años, haciendo así que los datos de distintas estaciones sean comparables. La serie desestacionalizada la conseguimos diferenciando la serie.
- **Realizar predicciones**, ya que si nuestros datos están afectados por una componente estacional, necesitaremos una estimación de esta de cara a realizar una predicción

Para desestacionalizar la serie, emplearemos los índices de variación estacional asociados a cada estación, ya que se suponen constantes año a año. Con esta técnica, se evidencian las diferencias en cada período, por ejemplo, podemos ver la diferencia del volumen de ventas en función de la época del año (mes, día de la semana, estación,...) Estos índices reflejan la cantidad fija o proporción en la que se modifica la tendencia en cada estación.

Una vez calculados estos índices, se desestacionaliza la serie, eliminando así el efecto de cada estación.

Por último, procedemos a realizar las predicciones. Para ello, necesitamos que se cumpla la condición de estacionariedad, es decir, la media y la varianza permanecen constantes en el tiempo (no tiene raíces unitarias). En el caso de no imponer esta condición de estacionariedad, predeciríamos carasterísticas que no serán las mismas en el futuro que en el pasado.

Se tiene que todo proceso lineal es estacionario, por tanto, obtendremos trabajaremos con series estacionarias, y de lo contrario, podremos aplicar los mismos métodos a series no estacionarias realizando las transformaciones pertinentes para conseguir la estacionariedad.

En nuestro caso, aplicaremos la metodología Box-Jenkis como método predictivo.

3.1.1. Metodología Box-Jenkis

Esta metodología tiene en cuenta la dependencia existente entre los datos, es decir, cada observación en el instante t será modelada a partir de los valores pasados. Los modelos se conocen con el nombre de ARIMA (modelos integrados autorregresivos de medias móviles), que deriva de las siguientes componentes: AR (Autorregresivo) , I (integrado), MA(Medias móviles)

El siguiente paso es identificar el modelo más adecuado a través del estudio de la función de autocorrelación (FAC) y la función de autocorrelación parcial (FAP).

Nota: el método recomienda como mínimo 50 observaciones en la serie temporal.

Fases de la metodología Box-Jenkis:

1. Identificar el la estructura ARIMA que sigue la serie a través del estudio de la función de autocorrelación simple (FAS) y la función de autocorrelación parcial (FAP). Determinar el modelo arima consiste en identidicar los órdenes p y q de su estructura autoregresiva y de medias móviles
2. Estiamción de parámetros: Una vez tenemos identificado el modelo, estimamos los parámetros AR y MA del modelo por el método de máxima verosimilitud, obteniendo el error estándar y los resíduos del modelo

   Nota: Es muy importante comprobar que las estimaciones son significativamente no nulas.
3. Diagnosis del modelo: Comprobamos que los resíduos sigan un proceso de ruido blanco mediante el Test de Ljung-Box.

   Si hemos identificado varios modelos y todos ellos pasan la diagnosis, nos quedaremos con uno de ellos según el criterio del menor AIC
4. Predicción: una vez identificado y validado el mejor modelo, se realizan las predicciones con éste.

3.2. Modelo 2 (decidir)

Capítulo 4

Data science process

4.1. Conocimiento del negocio (Knowledge of Bussiness)

Problem definition (this is the introduction of my report) Definición de objetivos y metodología para cumplirlos Muy importante la definición del objetivo: VOLUMEN DE VENTAS

4.2. Adquisición de los datos (Collect the data)

Explicación de los datos, fuente, explicación de las variables,...

4.3. Preparación de los datos (data preparation)

Fase del data cleaning:

- Eliminación de duplicados (filas y columnas)
- Datos erróneos (ej: precios negativos)
- Detección de valores faltantes (¿eliminación o imputación?)
- Detección de outliers (decidir si mantener, quitar o tratar a parte)
- Unificación de variables (unificación de unidades,...)
- Creación de variables si fuera necesario

Preparación de los datos

- Reformateo
- Categorización,...
- Feature selection (muy importante)

4.4. Análisis exploratorio de los datos (EDA)

- Distribución de las variables
- Relación entre las variables
- ¿Es necesario reducir la dimensionalidad?

4.5. Modelado

- Clasificación o regresión
- Librería caret
- Probar diferentes hiperparametrizaciones en los datos de entrenamiento
- Aplicar el modelo con la mejor hiperparametrización a los datos de validación y decidir un modelo final
- Evalucación del error con visualización (centrándome en los objetivos)

4.5.1. SVM

4.5.2. KNN

4.5.3. Árboles de regresión (XGBoost Model)

4.6. Evaluación y presentación de resultados (+análisis del error)

- Predicciones con el mejor modelo
- Final de la historia de una forma ordenada y resumida
- Señalar posibles mejoras y recomendaciones para proyectos futuros

Apéndice A

Apéndice: Título del Apéndice

A.1. Primera sección

Apéndice B

Apéndice: Título del Apéndice

B.1. Primera sección

\nocite{Luque2017,Luque2019,RStudio,R-base,R-knitr,R-rmarkdown,R-dplyr,R-ggplot2,MestaTFG,DefSectorRetail,ST_Apuntes # , ST_Art,ST_LibroUbiqum }