-Presentación del equipo

Predicción de Ventas en las Sucursales de una Cadena de Supermercados

Integrantes

Mentor

-Objetivo del trabajo

El objetivo central de la mentoría es implementar modelos predictivos univariados y multivariados de series temporales convencionales y de aprendizaje automático (y posiblemente combinaciones de ambos). Por medio de dichos modelos se intentará predecir las ventas en las sucursales de una importante cadena de supermercados, con una frecuencia diaria.

Trataremos de responder algunas de las siguientes preguntas: Cuáles variables proveen más información para predecir las ventas, si hay productos o sucursales similares, si los días de la semana, los días feriados y la ubicación de las sucursales impactan en las ventas, etc

-Intro/Contexto de Dataset

Se pretende obtener una primera aproximación a los datos bajo análisis y realizar una análisis exploratorio profundo de las series.

La información central se encuentra en el archivo train.csv porque que el mismo contiene las series históricas con las ventas de cada sucursal para cada familia de productos. Entonces, en primer lugar se cargan y presentan brevemente los datos del archivo train.csv y luego se proponen las consignas a seguir para la resolución del trabajo práctico.

Como las observaciones se encuentran indexadas en el tiempo, este tipo de datos se conocen como **series de tiempo**, por lo que algunas de las herramientas estadísticas habituales deben interpretarse con cierto cuidado y se pueden usar algunos recursos analíticos adicionales. El siguiente [tutorial](https://www.dataquest.io/blog/tutorial-time-series-analysis-with-pandas/) puede ser de gran ayuda para introducirse en la confección de algunos gráficos para series de tiempo y funcionalidades básicas de pandas para dicho tipo de datos.

Serie Temporal:

Contamos con un dataset que refiere a cantidad de ventas de una cadena de supermercados de Ecuador a lo largo de casi 5 años, de Enero del año 2014 a Agosto de 2017, agrupadas en 54 sucursales. Los valores de cada registro refieren a una cantidad total de unidades o kilogramos vendidos.

También contamos con 2 datasets con información externa, uno con datos de feriados y eventos especiales del país, y otro con el cambio del valor del petróleo a lo largo del tiempo, los cuales incluiremos a la hora de realizar predicciones en caso de que se encuentre una relación entre estos datos y las ventas de la cadena.

-Analisis y visualización de los datos (análisis de variables)

Se proponen las siguientes actividades motivadoras:

* Determinar el tipo de variables y el período de tiempo bajo análisis.
* Explorar si hay valores faltantes.
* Indagar sobre la existencia de ceros. ¿Hay valores nulos? ¿Son muchos? ¿En qué días ocurren?
* Analizar las distribuciones por medio de gráficos(por ejemplo: histogramas y boxplots) y medidas de estadística descriptiva (por ejemplo: media, varianza, coeficiente de variación, coeficientes de asimetría y kurtosis).
* ¿Se mantienen inalteradas en el tiempo las distribuciones? En este caso se podría seleccionar una serie (o el total) y graficar un boxplot para cada año.
* ¿Hay outliers? Si es así, ¿qué sugiere hacer con ellos? (Tener en cuenta las características de la estacionalidad)
* Explorar la existencia de estacionalidad.
* Estudiar la evolución temporal de las series. ¿En qué sentido tienden a moverse?
* ¿Las series tienden a moverse en forma conjunta? ¿Siempre se da dicho patrón?
* Seleccionar una serie y confeccionar un diagrama de dispersión entre la misma y su valor retardado un período. Repetir el mismo procedimiento entre la serie y sus valores retardados 2, 3 y 4 períodos.
* Grafique e interprete las funciones de autocorrelación (ACF) y autocorrelación parcial (PACF).

Datos totales

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Boxplot

Ventas de todas las sucursales de la muestra agrupadas por dia de la semana.

Boxplot con valores extremos

Diagrama, Escala de tiempo

Descripción generada automáticamente

Boxplot sin valores extremos

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

Se observa para todo el periodo considerado en forma agregada como es esperable que los días sábados y domingos son días de mayores ventas. Los días miércoles presentan a su vez, un pico de ventas que podría deberse a promociones o descuentos.

Explorando los datos sobre promociones, confirmamos que los días miércoles son los días que mas productos poseen promociones de ventas (ver la siguiente tabla).

Las sucursales tienen diferentes comportamiento (distribución) en cantidades.

Histograma de frecuencias de ventas totales por sucursal de la muestra: a través de este grafico, también podemos visualizar la distribución de las ventas y la presencia de valores atípicos. Se puede observar que hay dos grupos que tienen distribucion similar entre sí:

Grupo A- Stores 13, 23  y 5 por un lado (leptokúrticas) y

Grupo B - Store 7 y 51 por otro,  (platykurticas).

Las primeras tienen un sólo pico marcado de ventas y las segundas poseen, en un caso dos picos de ventas y la ultima no presenta picos de ventas significativos.

Gráfico, Histograma

Descripción generada automáticamente

Evolución temporal de la serie

Los componentes de la serie de tiempo son:

-Tendencia: describe si la serie temporal es creciente a lo largo del tiempo.

-Estacionalidad: describe la señal periódica en su serie temporal.

-Ruido: describe la variabilidad en los datos lo que no puede ser explicado por el modelo.

Gráfico

Descripción generada automáticamente

Autocorrelación y Autocorrelación Parcial

ACF

Autocorrelación es una cuantificación del grado de asociatividad de variables continuas. En este caso la autocorrelacion se analiza en que medidad un valor en un tiempo t depende del valor anterior en un tiempo t-1.

A continuación calcularemos la función ACF que permite calcular la autocorrelación simple. La misma expone como ordenada el coeficiente de correlación e indica hasta que punto la variable depende de sí misma.

En esta gráfica, que los valores estén próximos a 1 implica que hay una dependencia fuerte respecto al valor anterior.

Se observa una mayor correlación para el retardo de 7 días, es decir la comparación con el mismo día de la semana.

Gráfico

Descripción generada automáticamente

-Curación de datos

Se busca profundizar el conocimiento de las series bajo estudio y realizar tareas de preprocesamiento.

En lo desarrollado en el primer práctico pudimos notar algunas características propias de los datos de series temporales y en particular de las ventas de las sucursales. Algunas de ellas son que se trata de series no estacionarias, o sea que sus distribuciones cambian a través del tiempo pero no parecen presentar raíces unitarias. Además, se aprecian patrones de estacionalidad. Lo antes descripto tiene consecuencias directas en la selección de las metodologías adecuadas para abordar las problemáticas de tratamiento de valores faltantes, detección de outliers y la construcción de modelos predictivos y/o causales.

Curación: Al observar el dataset identificamos algunos días faltantes para los cuales no se tiene registro, correspondientes al 25/12, y existencia de ceros para todos los primero de Enero de cada año, probablemente estas dos fechas la cadena cierra sus sucursales. También observamos que algunas sucursales empiezan a tener registrar ventas a partir de una fecha posterior al inicio del dataset, las cuales filtraremos para que no afecten al realizar cálculos de ventas acumuladas a lo largo del tiempo.

**Gráfico

Descripción generada automáticamente**

**Transformaciones**

Se indaga sobre la utilidad de aplicar transformaciones sobre las series. Por ejemplo: logaritmo natural, diferencias, tasa de crecimiento, estandarización, normalizacion, transformaciones de Box y Cox, etc.

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Gráfico, Gráfico de líneas

Descripción generada automáticamente

Gráfico, Gráfico de líneas

Descripción generada automáticamente

-Conclusiones

Los siguientes gráficos boxplot representan las ventas totales de todas las sucursales de nuestra muestra, desagregadas por mes para todo el periodo analizado. Se observa que existe un pico de ventas para el mes de junio y que se incrementan las ventas desde septiembre a diciembre.

Observaciones para cada año completo:

2013: Se presenta un pico de ventas en el mes de diciembre y el mes mas bajo en ventas es febrero.

2014: Se observa una fuerte caída en las ventas para los meses de febrero, abril, mayo y junio. El pico de ventas se produce en diciembre.

2015: Desde enero a mayo se registran periodos con bajas ventas, que se arrastra probablemente del comportamiento observado en 2014. A partir de junio los niveles de ventas se recuperan.

2016: Aunque se aprecian algunos picos, como el habitual de diciembre, este año tiene una estabilidad y resultan más parejas de mes a mes.