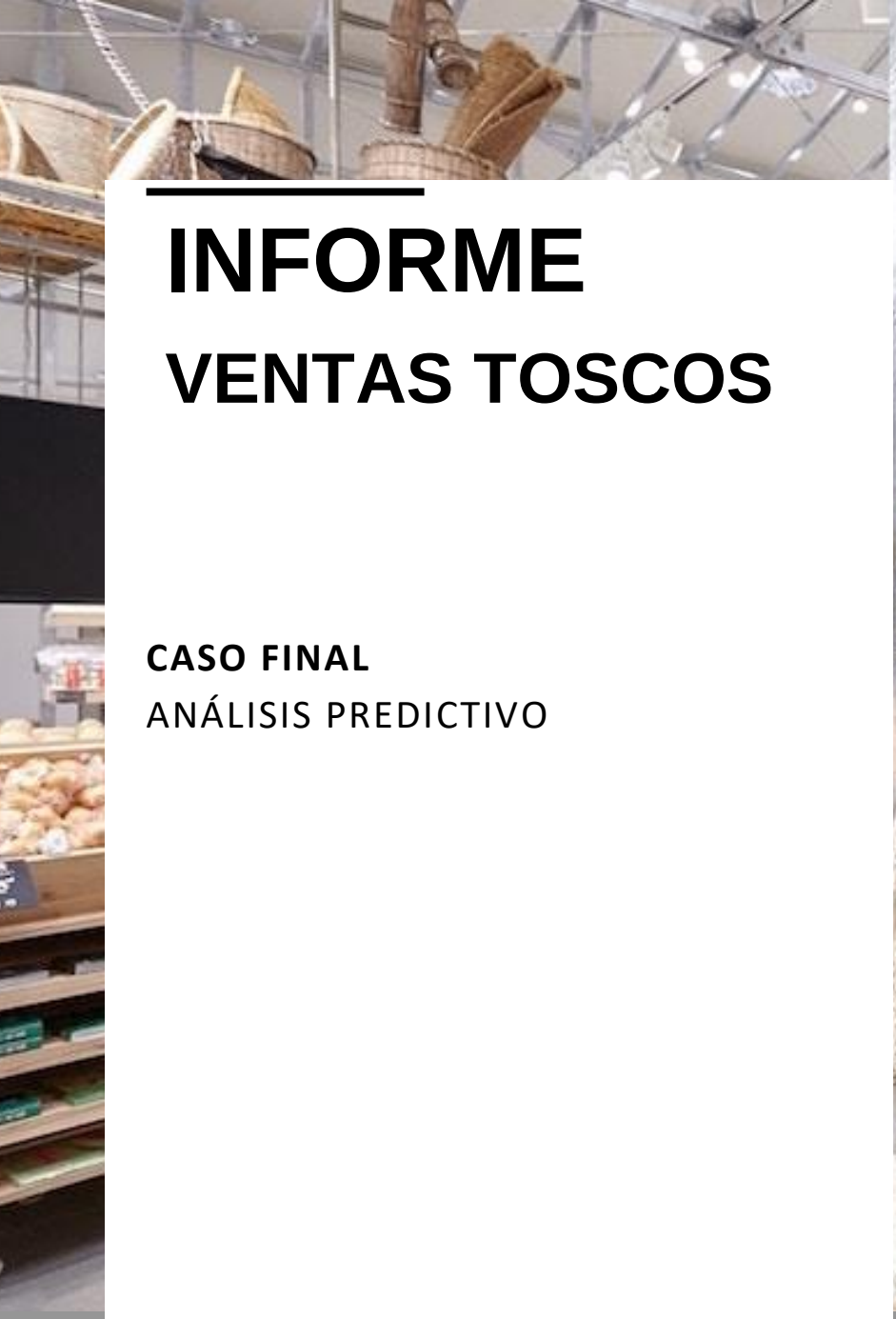


INFORME

VENTAS TOSCOS

CASO FINAL

ANÁLISIS PREDICTIVO

ENERO 2023

Isabel Fátima Marsá Martín

Profesor: Ricardo Queralt

CUNEF
COLEGIO UNIVERSITARIO DE ESTUDIOS FINANCIEROS

Introducción

TOSCOS es una cadena de supermercados con diez tiendas divididas en tres zonas (1,2 y 3) y cada zona tiene 3, 3 y 4 tiendas respectivamente. El periodo que hemos utilizado para el estudio es del 01 de enero de 2013 al 31 de julio de 2015.

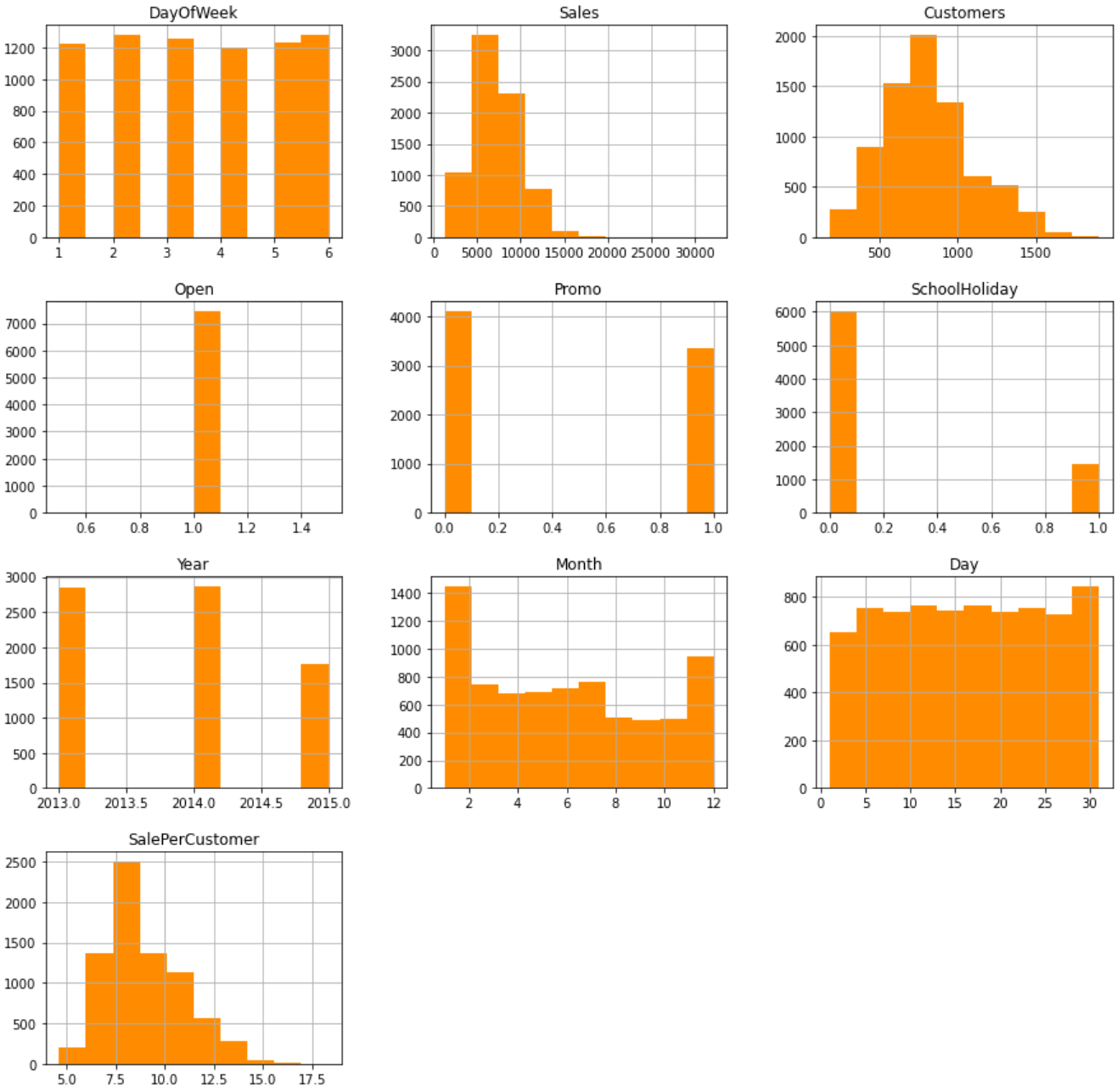
El objetivo es predecir las ventas de diez tiendas de la cadena de supermercados desde el 1 de agosto de 2015 al 10 de septiembre de 2015. Hemos realizado una predicción de las ventas totales diarias y las ventas totales semanales. Para ello, hemos utilizado los dos modelos de predicción que mejores resultados han dado, siendo estos Prophet Facebook y ETS (Error Trend Seasonality). Con estas predicciones, la empresa TOSCOS espera poder identificar los productos y las tiendas que desempeñan un papel clave en sus ventas y utilizar esa información para tomar las medidas correctas que garanticen el éxito de su negocio.

Hemos tenido a nuestra disposición variables de interés como la cantidad de ventas desde el año 2013 hasta 2015 de las diez tiendas, el número de clientes que ha entrado en cada tienda, si estaba abierta, si tuvo promociones, si fue festivo (estatal o escolar). tiendas A la hora de realizar las predicciones hemos tenido en cuenta variables como los festivos, si estaba abierta o no y si tenía promociones o no, así como el nombre de cada tienda. El dataframe inicial sobre el que hemos trabajado contiene 9,236 datos y las siguientes variables:

Date	Store	DayOfWeek	Sales	Customers	Open	Promo	StateHoliday	SchoolHoliday
2013-01-01	T2c	2	0	0	0	0	a	1
2013-01-01	T3a	2	0	0	0	0	a	1
2013-01-01	T3b	2	0	0	0	0	a	1
2013-01-01	T3c	2	0	0	0	0	a	1
2013-01-01	T3d	2	0	0	0	0	a	1

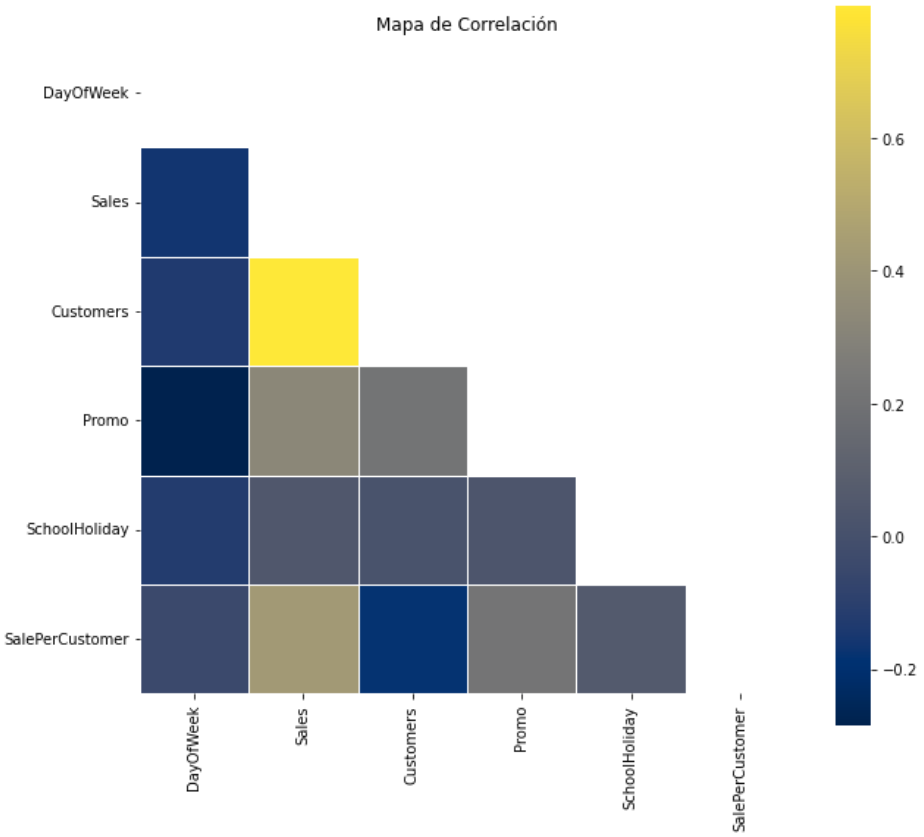
Análisis Exploratorio de Datos

Lo primero que hemos realizado es un análisis exploratorio de datos. Hemos analizado cada variable y nos hemos hecho una idea general así. Podemos observar la distribución de las variables:

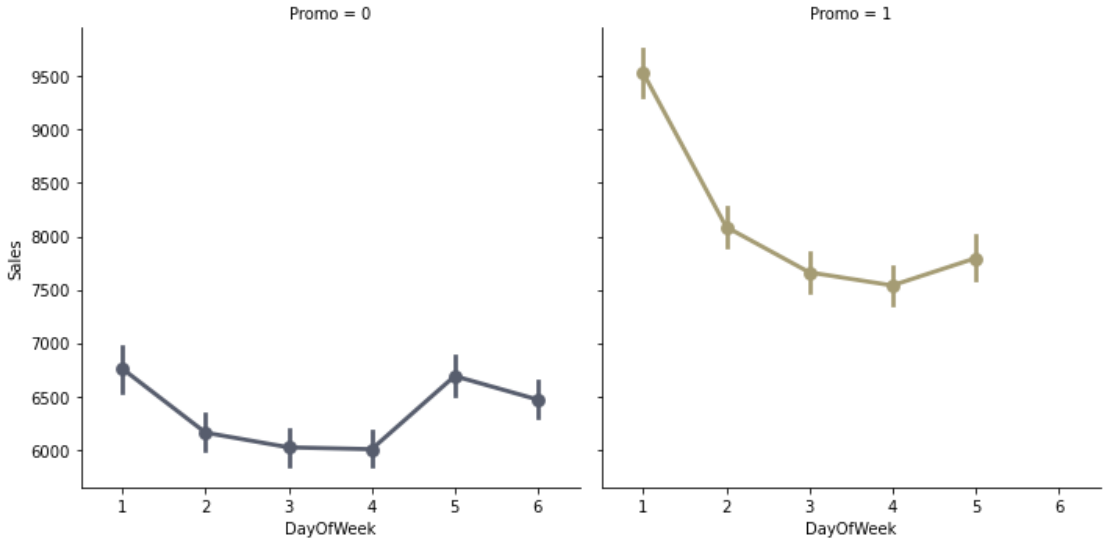


Por un lado, hemos observado que el dataset contiene datos de días de ventas nulas, como por ejemplo los domingos. Además, en total 2 días alguna tienda abrió un día laboral y tuvo cero ventas y hemos asumido que existe una razón externa detrás de este suceso, como por ejemplo una huelga o un problema interno de la tienda. Para el análisis exploratorio, hemos procedido a eliminar las filas que incluyan fechas en las que la tienda estaba cerrada y las ventas fueron 0. Hemos añadido una variable nueva (SalePerCustomer) a partir de la variable ventas y clientes para obtener la cantidad de ventas por cliente ya que hemos considerado que es una variable de interés para comprender mejor los datos. A continuación, podemos observar la distribución de cada una de las variables. Observamos que el día de la semana más frecuente es el martes (2) y el sábado (6) y las ventas se concentran entre los valores 5,000 y 10,000. Además, el número de clientes medio es cercano a 800. Las variables dicotómicas que toman los valores 0 y 1 son las variables promoción, festivo escolar, festivo estatal y si estuvo abierta (1) o no (0). Cabe mencionar que la variable StateHoliday era inicialmente una variable con los valores 0, a, b y c, siendo los tres últimos tres tipos de festivo. Lo que hemos hecho es remplazar estos tres caracteres por 1, ya que indica que ese día fue festivo, por tanto, hemos convertido esta variable tipo objeto en una variable dicotómica. También nos hacemos una idea de que los meses más frecuentes son enero y diciembre, que corresponden a los meses de Navidad. También es interesante señalar los clientes se gastan al día una media de 8.95.

Un mapa de correlación nos ha servido para comprender la relación entre las variables del dataset. Observamos que existe una correlacion positiva muy elevada entre ventas (Sales) y clientes (Customers) de una tienda. Hay correlacion positiva también entre las variables de ventas y venta por cliente (SalePerCustomer) y entre el hecho de que una tienda tenga una promocion ese día (Promo=1) y la cantidad de clientes que acudieron a esa tienda ese día. No hay casi relacion entre promocion y día de la semana, lo cual quiere decir que las promociones probablemente no tengan nada que ver con el día de la semana en el que se dan.

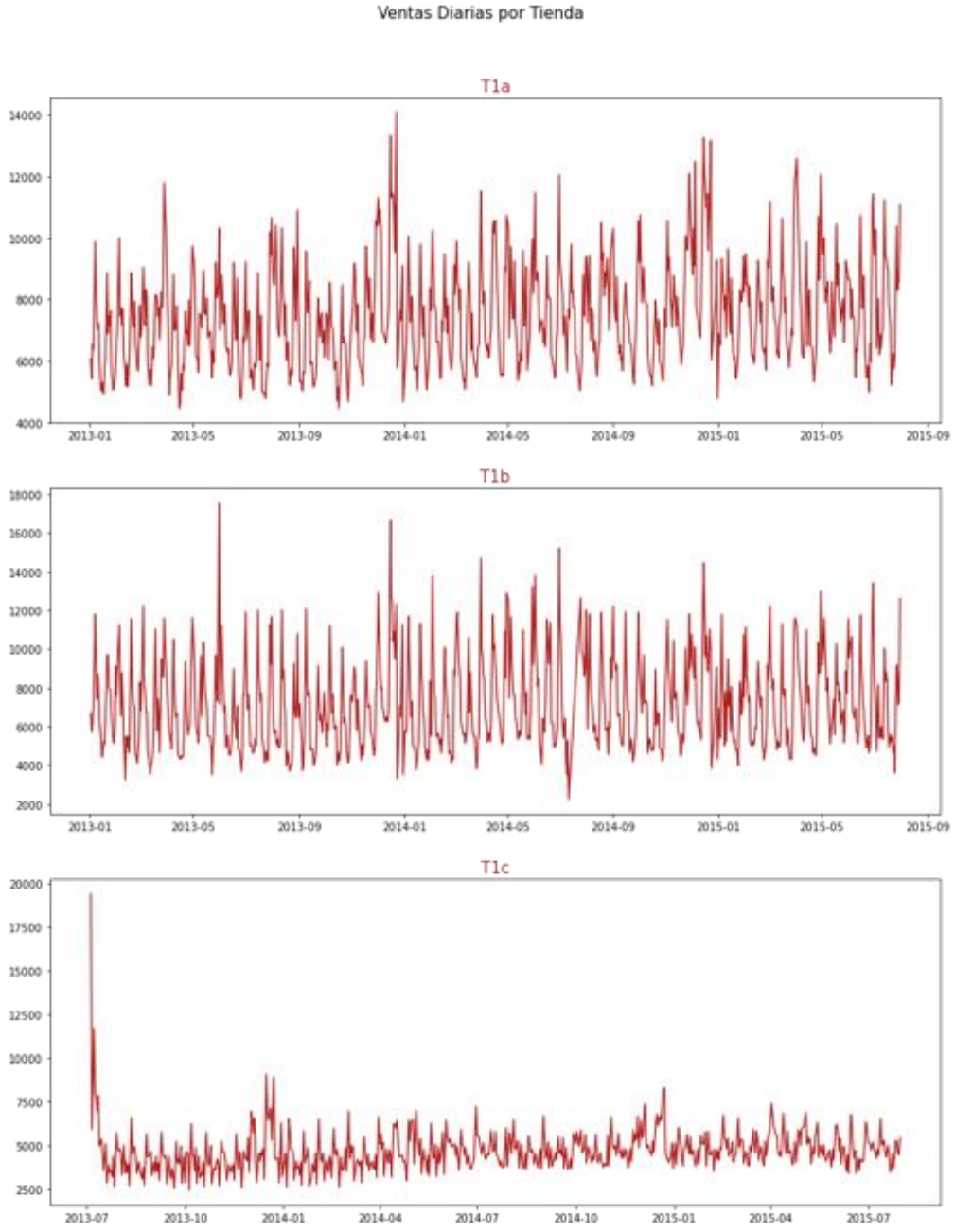


Hemos considerado interesante analizar el efecto de las promociones en las ventas, por lo que hemos realizado los siguientes gráficos. Nos hacemos una idea sobre la influencia de las promociones en las ventas, independientemente del día de la semana. Los lunes es claramente el día de la semana en que más ventas hay. Sin embargo, el jueves es el día en que menos ventas se registran. Por tanto, podemos concluir que las promociones tienen un enorme efecto sobre la cantidad de ventas.

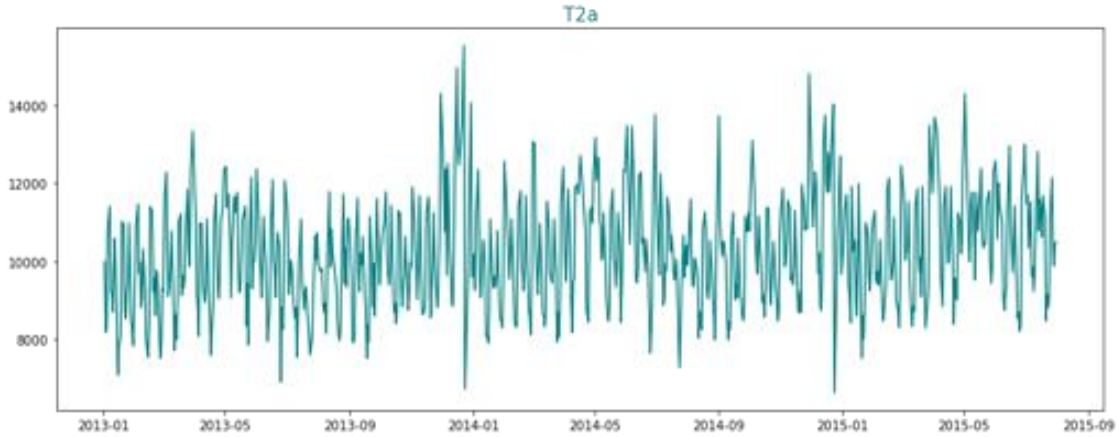


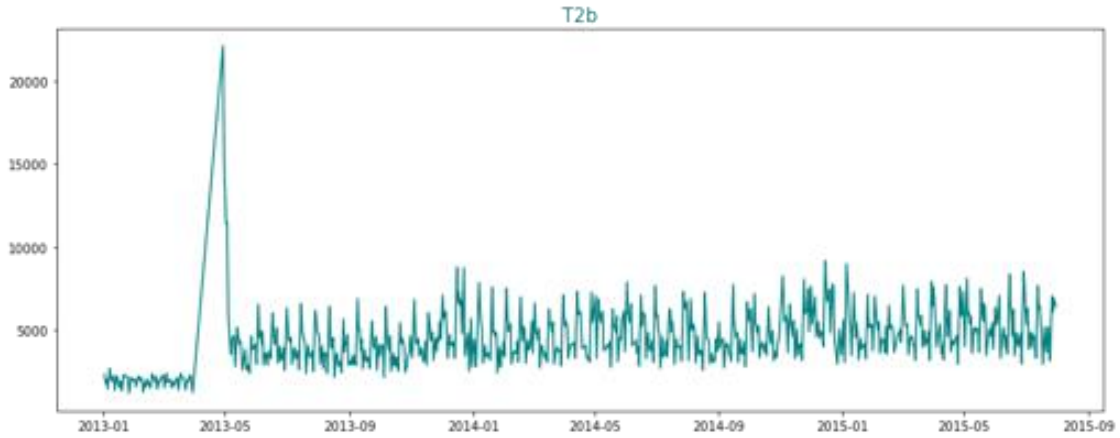
Ventas diarias por tienda

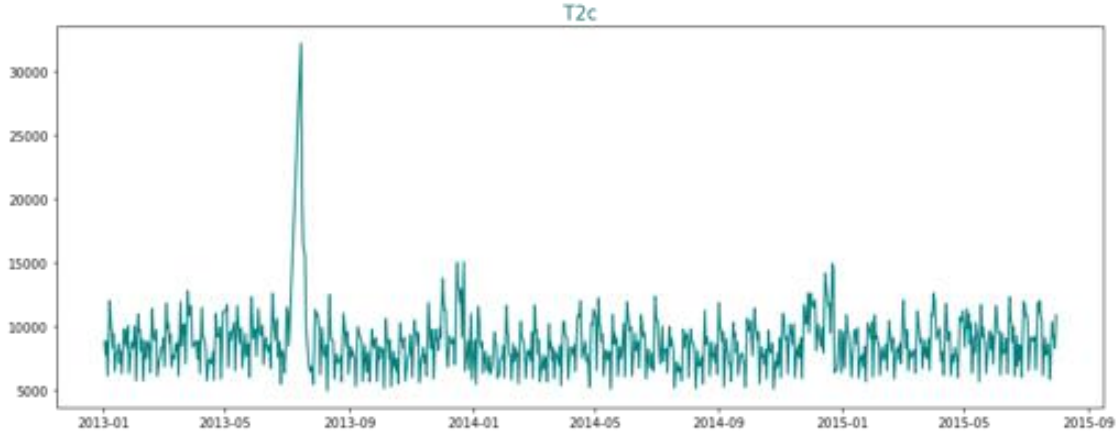
Hemos realizado un filtro de ventas por tienda para observar la diferencia entre las ventas según la tienda.
Por un lado, observamos los gráficos de las tiendas de la zona 1:



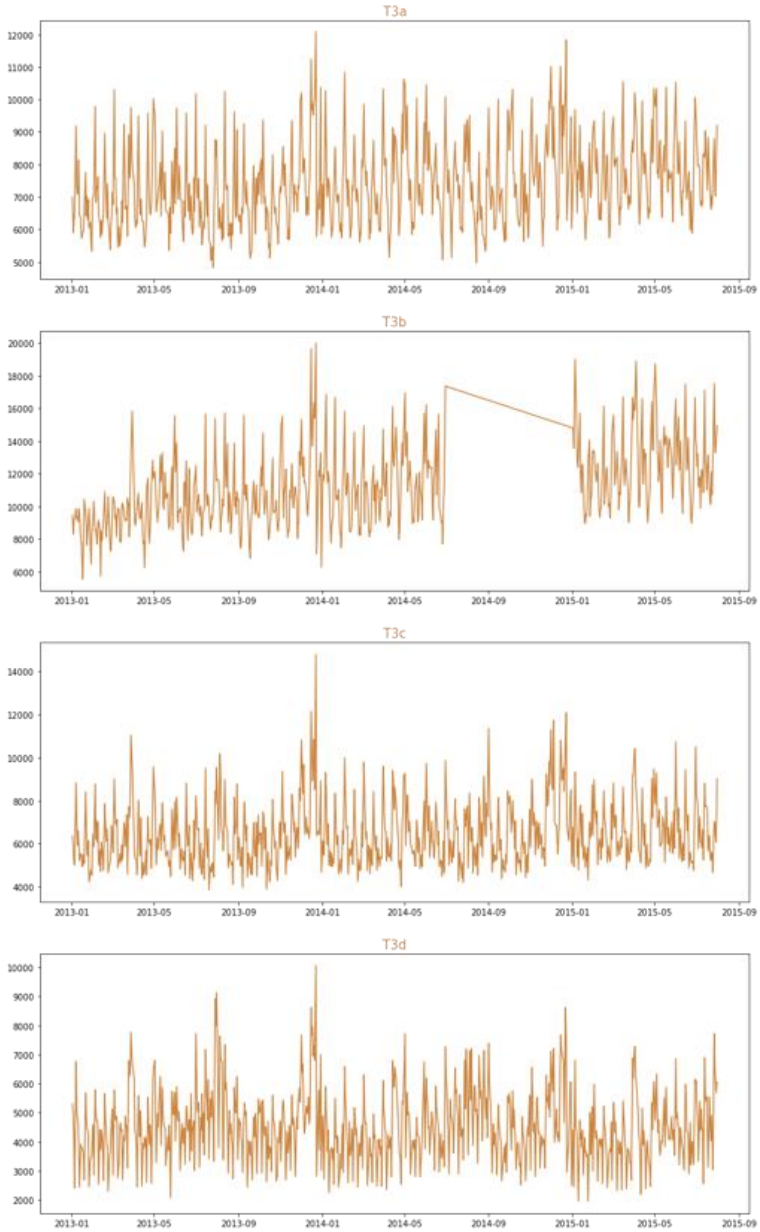
También observamos los gráficos de las tiendas de la zona 2







Y por último las tiendas de la zona 3:



La tienda con máximo número de ventas acumuladas es la tienda a situada en la zona 2 (T2a), con un total de 7,939,205. Además, hay que tener en cuenta que ha abierto 784 días en el periodo estudiado, siendo de las que más días ha abierto. Tiene unas ventas por cliente igual a 6,031.54 y ha tenido un total de 1,028,904 clientes durante este periodo. La tienda que más ventas por cliente acumuladas tiene es la T2c con un total de 8,095.36 clientes. Por otro lado, la tienda T1c tiene el menor número de ventas, 2,959,073 pero es la segunda que menos días ha abierto, liderada por la t3c que abrió un total de 622 días. Hemos observado que existe algún valor atípico en estos datos,

pero no hemos considerado oportuno eliminarlos ya que no aporta demasiado valor a la hora de realizar las predicciones.

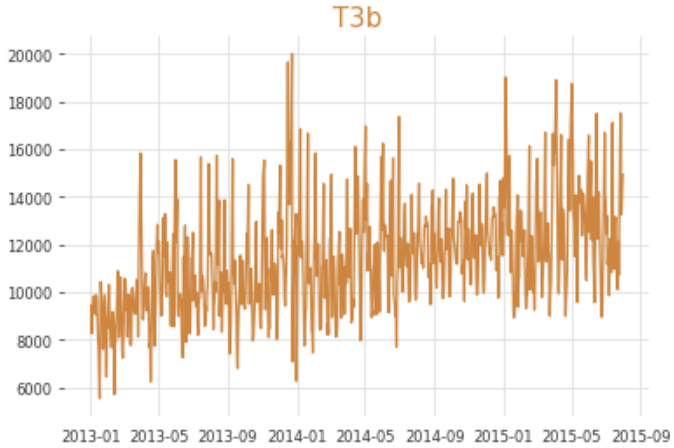
Tienda 3B

Un hecho interesante que hemos detectado es que la tienda T3b no tiene valores desde julio de 2014 hasta el 1 de enero de 2015.
por ello, para poder realizar la predicción correctamente, deberemos arreglar esto remplazando esos valores *missing*. Hemos considerado dos opciones para tratar estos valores nulos:

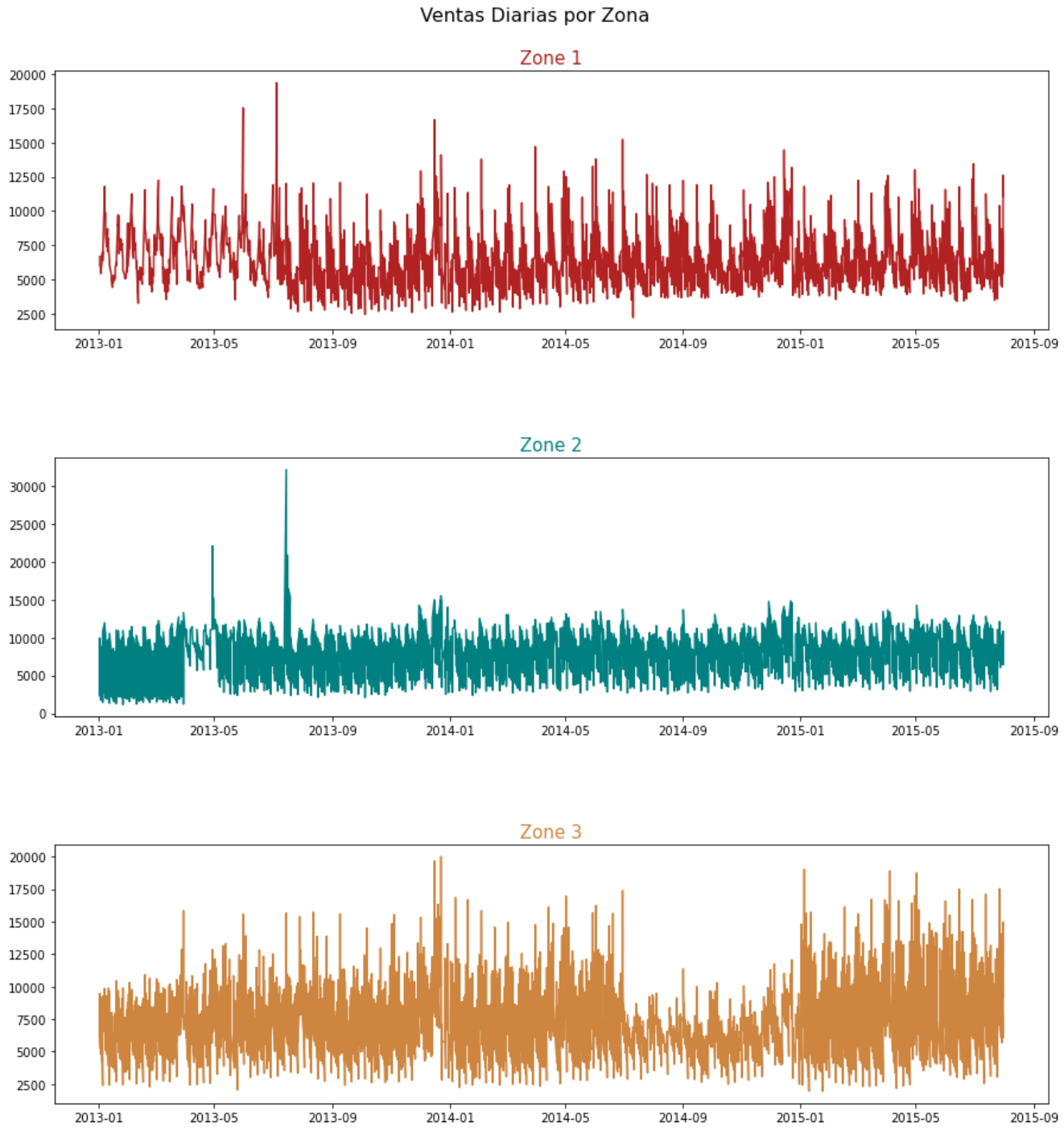
- Sustituir estos valores por valores medios
- Remplazar estos valores con sus predicciones

Finalmente, lo que hemos realizado para solucionar esto es aplicar un modelo de predicción sobre los valores desde enero de 2013 hasta junio de 2014 para poder rellenar estos datos y que no afecte a la predicción. Hemos utilizado un modelo Theta que es una herramienta que pertenece a la librería de Darts. Hemos optado por escoger el modelo Theta ya que es el que nos daba los resultados más coherentes y el que menor error tenía de todos. Hemos probado también con ARIMA y ETS, pero finalmente escogimos Theta por estas razones. Tras obtener las predicciones del rango de fechas faltante, hemos añadido las filas al dataset original y el nuevo dataset es el que utilizaremos para realizar las predicciones. Las cinco primeras predicciones que hemos obtenido se muestran a continuación. El gráfico de ventas diarias de la tienda T3b ha quedado así después de esta modificación:

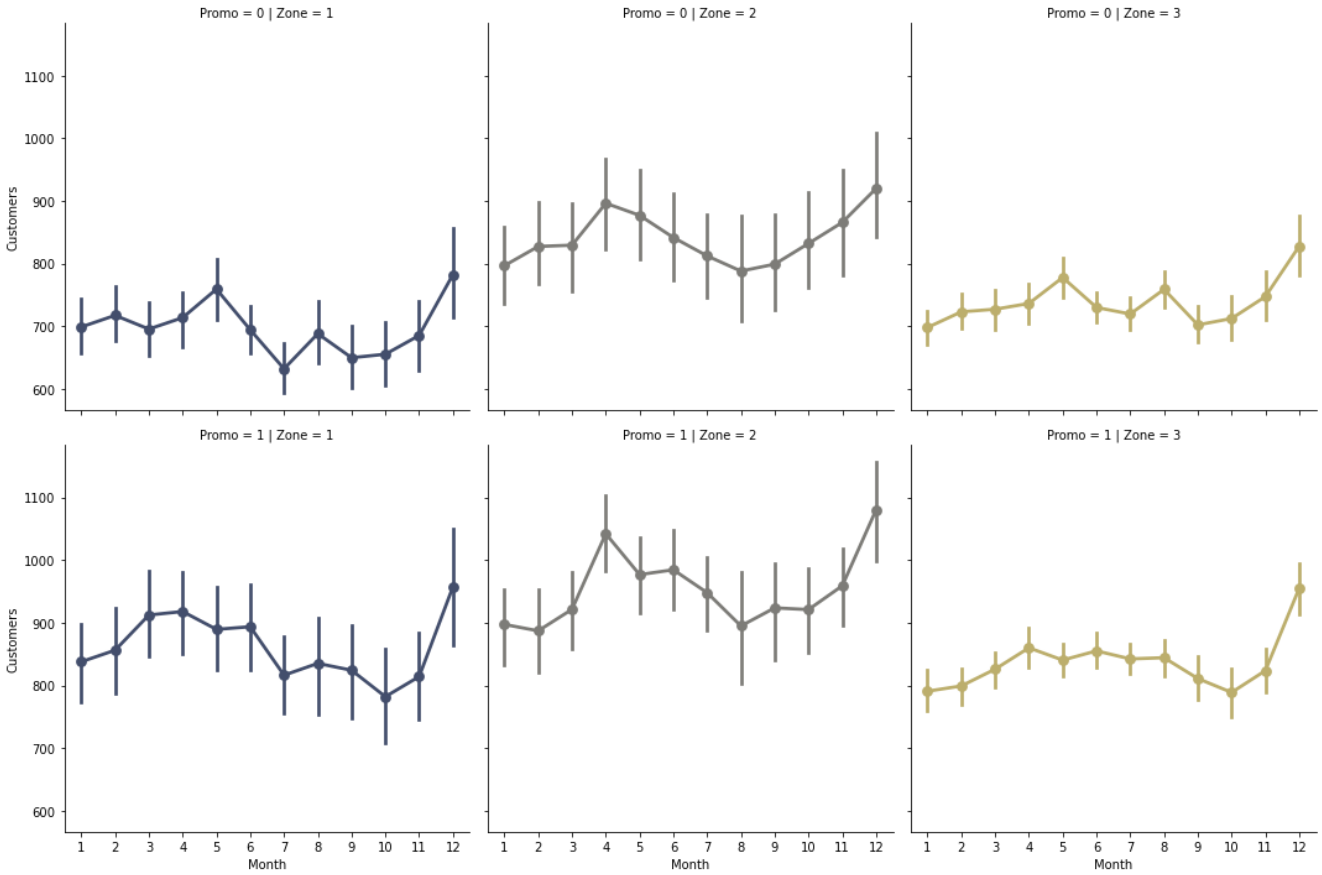
	Sales
2014-07-01	11884.686743
2014-07-02	11042.370674
2014-07-03	12290.894176
2014-07-04	11046.254549
2014-07-05	11092.507988



Ventas diarias de las tres zonas

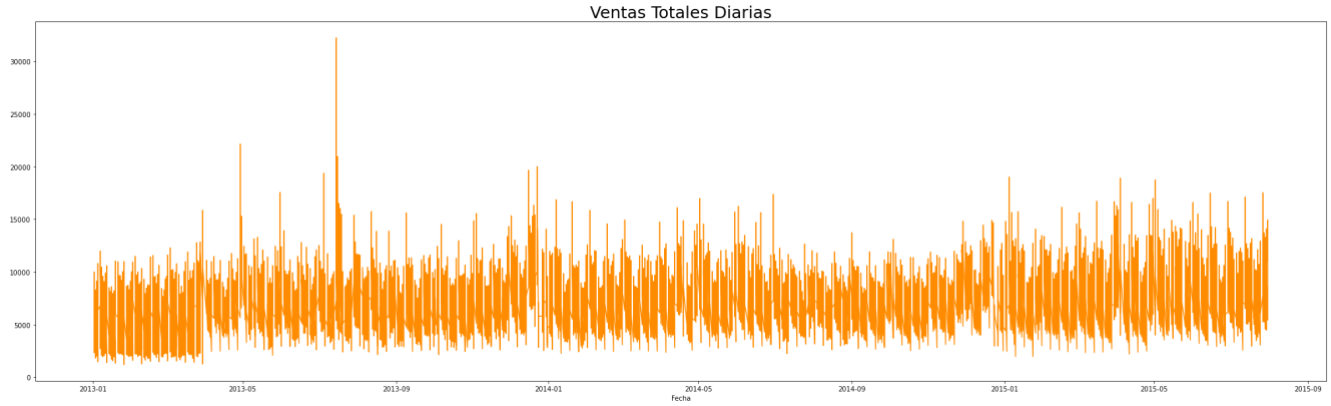


Efecto de las promociones en las ventas por zona

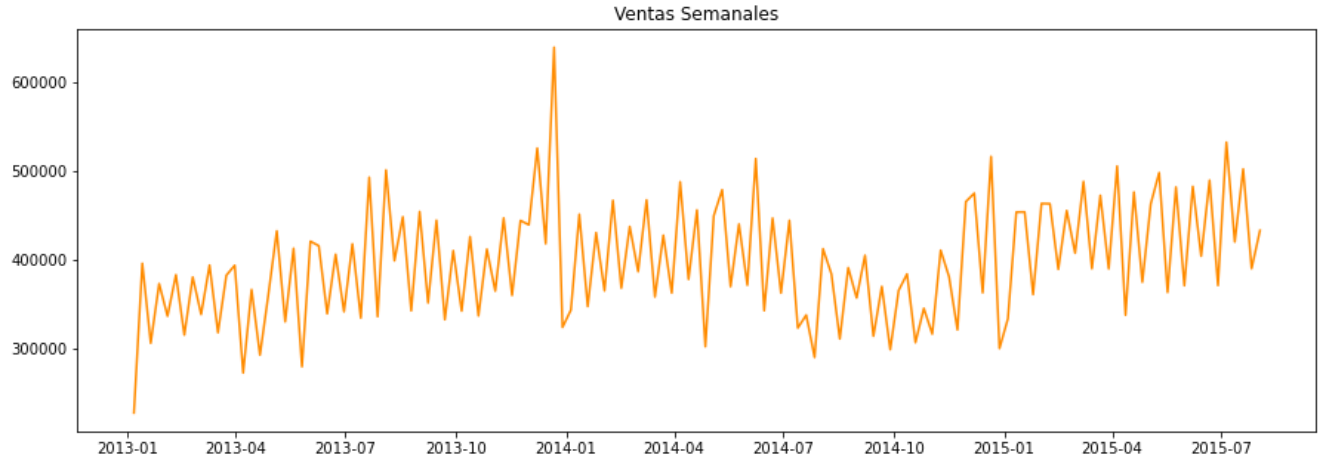


Observamos que la gente tiende a comprar más los lunes cuando hay promociones y cuando no las hay.

Ventas totales diarias



Ventas totales semanales



Análisis Predictivo

Finalmente realizaremos la predicción de las ventas de TOSCOS desde el 1 de agosto de 2015 al 10 de septiembre de 2015, es decir, las seis semanas próximas. Por un lado, hemos realizado la predicción de las ventas totales diarias utilizando Prophet Facebook. La predicción de ventas semanales totales la hemos realizado con un modelo ETS. Primero, consideramos realizar las predicciones por tienda, pero nos dieron unos resultados muy malos probando con diferentes modelos. Por tanto, finalmente decidimos realizar una predicción de las ventas de todas las tiendas.

Para realizar las predicciones hemos utilizado el dataset modificado, con los datos faltantes de la tienda T3b rellenos con las predicciones que hemos realizado previamente con el modelo Theta. Utilizar este dataset en vez del original ha hecho que nuestras predicciones sean de mayor calidad y hemos obtenido resultados de métricas de error mucho más bajos.

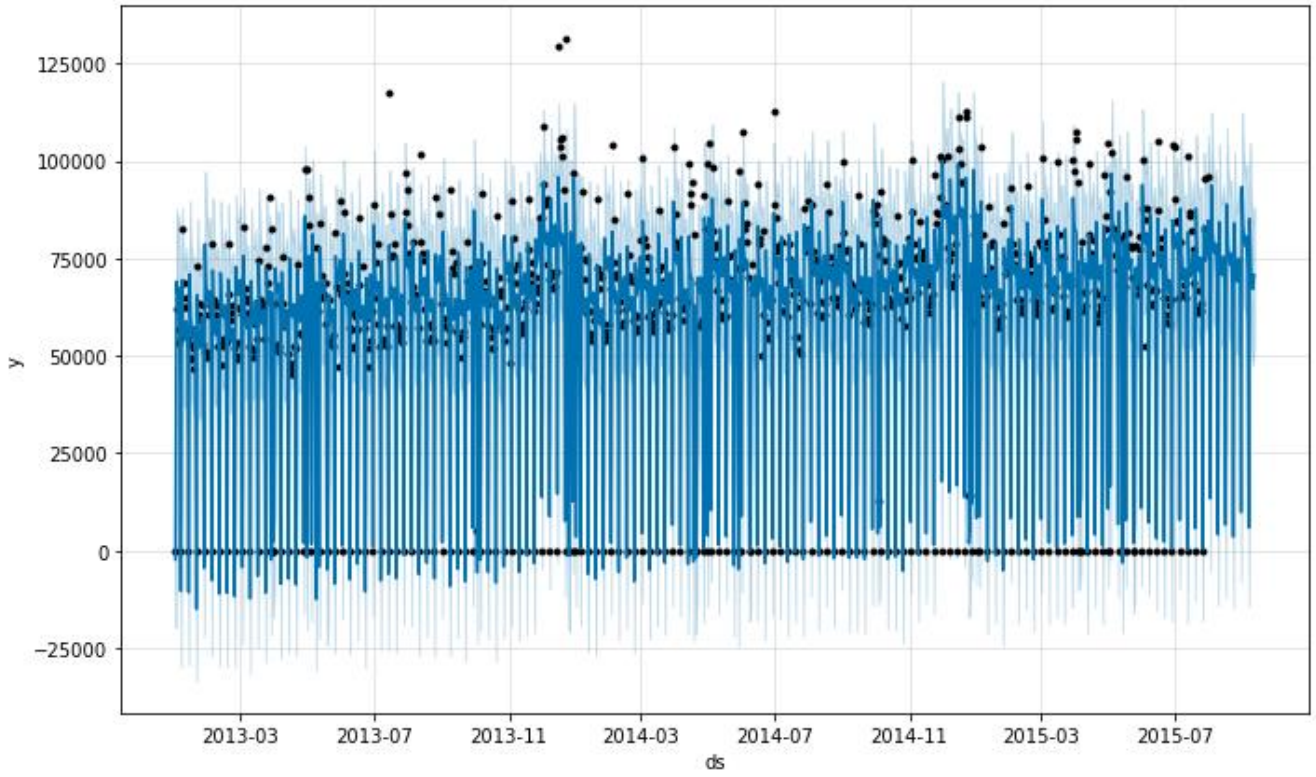
Predicción de ventas diarias totales

Prophet es una herramienta para realizar predicciones de calidad que ha sido desarrollada por Facebook. Utiliza un modelo estadístico basado en series de tiempo para hacer predicciones a futuro. Es fácil de utilizar y puede manejar tanto datos diarios como semanales. Funciona mejor con series temporales que tienen fuertes efectos estacionales y varias temporadas de datos históricos. Adicionalmente, cabe destacar que es resistente a los datos faltantes y los cambios en la tendencia y, por lo general, maneja bien los valores atípicos.

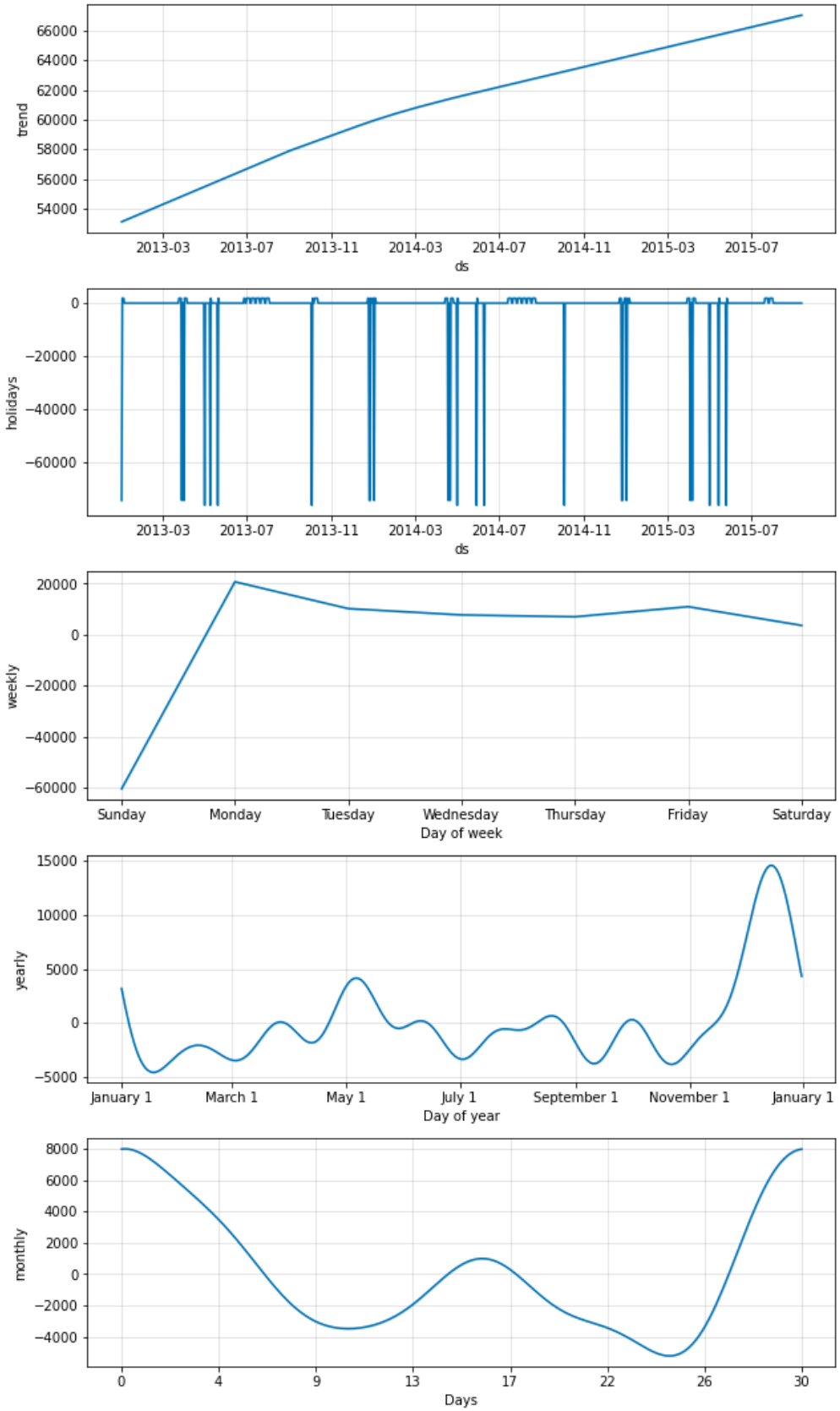
Por estas razones, hemos utilizado Prophet para realizar la predicción de las ventas diarias de la tienda TOSCOS. Para evaluar estos modelos hemos hecho uso del coeficiente de determinación, R^2. Este indica qué porcentaje de la variación en los datos observados puede ser explicado por el modelo. Un R^2 cercano a 1 indica un ajuste excelente del modelo a los datos, mientras que un R^2 cercano a 0 indica un ajuste pobre.
Primero realizamos un modelo únicamente teniendo en cuenta las ventas y hemos obtuvimos un R^2 igual a 0.71. Después, añadimos la estacionalidad y los festivos al modelo y obtuvimos un R^2 igual a 0.89, lo cual quiere decir que mejoró. Para añadir los festivos de las diez tiendas hicimos lo siguiente. Consideraremos que un dia es festivo si la variable StateHoliday o SchoolHoliday es mayor de 6 para cada día porque al hacer el agrupamiento anteriormente por fecha se suman estas variables. Por lo que si por ejemplo tenemos 10 en StateHoliday quiere decir que hay fiesta estatal en las diez tiendas de TOSCOS. Por lo tanto, hemos considerado un numero aceptable 6 ya que consideraremos que fue fiesta si lo fue en más del 60% de las tiendas.
A continuación se muestran las predicciones de los primeros diez días que hemos obtenido:

	ds	yhat
942	2015-08-01	77279.198862
943	2015-08-02	13391.151974
944	2015-08-03	93830.611157
945	2015-08-04	82153.668155
946	2015-08-05	78454.851671
947	2015-08-06	76398.871669
948	2015-08-07	78788.100391
949	2015-08-08	69755.084664
950	2015-08-09	4150.319974
951	2015-08-10	84395.051231

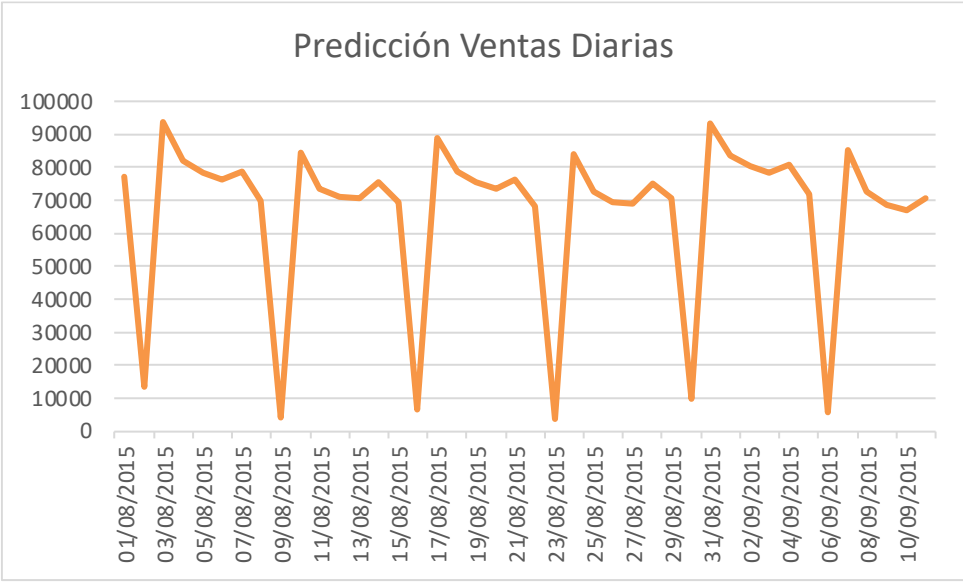
Prophet grafica los valores observados de la serie temporal (los puntos negros) , los valores de la predicción (línea azul) y los intervalos de las predicciones (las regiones azules sombreadas). En el grafico que se muestra a continuación podemos visualizar la predicción:



Además, observamos a continuación un gráfico de descomposición de Prophet:



En el primer gráfico observamos que el total de las ventas de las tiendas han ido incrementando linealmente a lo largo de los meses. El segundo gráfico muestra los festivos incluidos en el modelo. El tercer gráfico muestra que el pico de ventas se da los lunes, y que los domingos no hay ventas y el cuarto gráfico muestra que la época de más ventas es Navidad, es decir, diciembre y enero. Por último, el último gráfico señala que los días del mes que más ventas hay es entre los días 1 y 4 y los últimos del mes.

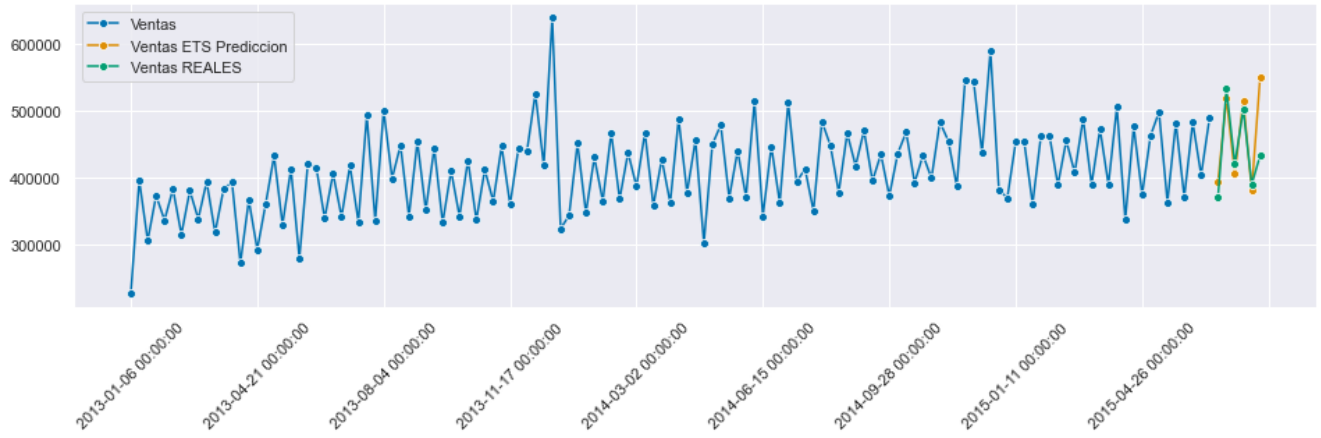


Predicción de ventas semanales totales

Realizamos la predicción de las seis semanas posteriores al periodo estudiado. Antes de realizar las predicciones semanales totales, hemos evaluado distintos modelos para obtener el óptimo. Hemos probado los modelos ETS, ARIMA, VAR, Prophet y TBATS. El modelo ETS ha sido el que finalmente nos ha dado los mejores resultados a la hora de ajustarse a nuestros datos, por lo que hemos optado por escoger este modelo para realizar nuestra predicción de ventas semanal. Tras agrupar los datos por semanas, hemos detectado dos factores que podrían afectar negativamente a nuestras predicciones. Por un lado, los valores nulos y por otro los valores atípicos. Hemos remplazado todos los valores de 0 en la variable de ventas por valores faltantes ya que los 0 nos pueden dar errores para la predicción. Después hemos aplicado la función forward fill, que reemplaza los valores faltantes con el valor inmediatamente anterior en la serie de tiempo. Esto se hace para evitar tener valores faltantes en la serie de tiempo, ya que esto podría afectar negativamente a las predicciones realizadas.

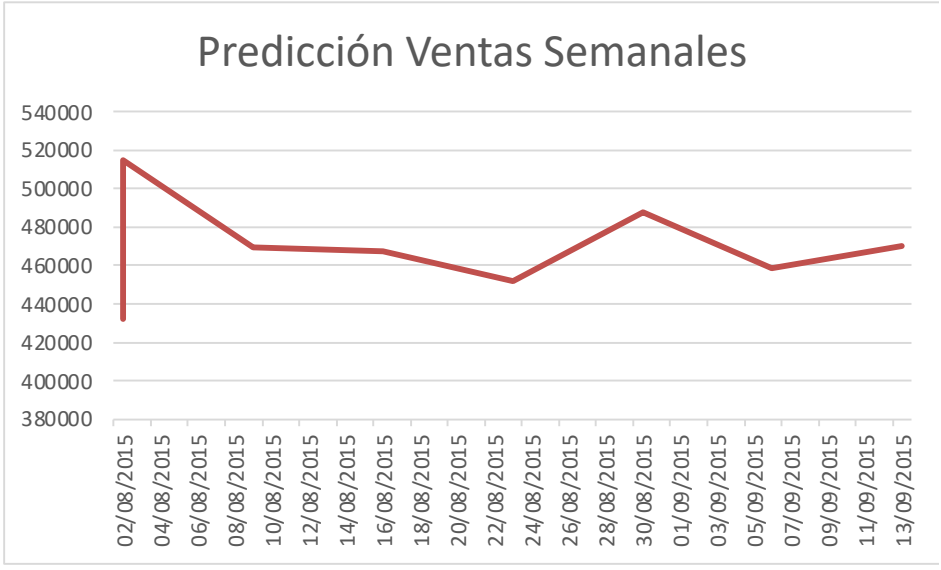
Hemos detectado valores atípicos las semanas de los días 2013-12-22 y 2014-12-21, con unas ventas totales excesivamente altas siendo estas 638654.00 y 589003.08 respectivamente. Finalmente hemos optado por no tratar estos valores atípicos ya que no mejoraban nada nuestro modelo.
Hemos realizado el test de Augmented Dickey Fuller para evaluar si la serie temporal semanal es estacionaria o no, obteniendo un Test Statistic de -3.390 y un p-valor de 0.011. Por tanto, hemos llegado a la conclusión de que esta serie es estacionaria y un modelo ETS debería funcionar bien sobre ella. Si la serie temporal no fuese estacionaria, sería necesario realizar alguna técnica de suavizado previamente como la diferenciación para poder aplicar un modelo ETS correctamente.

Los modelos ETS se basan en la suposición de que una serie temporal es una combinación de una tendencia a largo plazo, un patrón estacional y un error aleatorio y se dividen en tres tipos según el tipo de componentes incluidos en el modelo: Modelos ETS (A) solo incluyen la componente de tendencia, modelos ETS (M) que incluyen tanto la componente de tendencia como la componente estacional y los modelos ETS (A,A,N) que incluyen tendencia, estacionalidad y un factor no estacional. Una de las ventajas de los ETS es que son capaces de capturar tanto tendencia como patrones estacionales, así como adaptarse a cambios en estos factores a lo largo del tiempo. El modelo que hemos utilizado en este caso es el AutoETS, con un periodo estacional igual a 52, por que tenemos 52 semanas por año.
Cuando he evaluado el modelo estimado, he obtenido un AIC (Criterio de Información Akaike) igual a 3205.22. El AIC es una medida de la bondad del ajuste del modelo, pero también tiene en cuenta la complejidad del modelo y cuanto menor sea este valor indica que el modelo es mejor. A continuación, se muestra el grafico que incluye las ventas desde 2013 y podemos observar en verde las ventas reales y en amarillo la predicción que realiza el modelo sobre los datos de entrenamiento. Se puede decir que el modelo ha predicho muy bien los datos ya que las líneas verde y amarilla se comportan de manera similar. El MAPE que hemos obtenido es 0.07, que es un valor muy bueno.



El AIC que hemos obtenido al ajustar el modelo a nuestros datos ha sido de 3348.42 y las predicciones semanales son las siguientes:

	Sales
2015-08-02	514890.348611
2015-08-09	469485.234731
2015-08-16	467301.609035
2015-08-23	451909.714195
2015-08-30	487352.206034
2015-09-06	458223.493008
2015-09-13	470023.116022



Conclusiones

Hemos realizado las predicciones de las ventas de diez tiendas de TOSCOS desde el 1 de agosto de 2015 al 10 de septiembre de 2015. Para ello hemos utilizado el dataset que nos ha aportado la empresa, que incluyen datos desde 01 de enero de 2013 al 31 de julio de 2015.

Tras realizar un análisis exploratorio de datos hemos detectado que la tienda que más ventas ha obtenido ha sido la T2a, con un total de 7,939,205 y T1c tiene el menor número de ventas, 2,959,073 pero es la segunda que menos días ha abierto. Además, hemos descubierto que la tienda 3b no tenía datos desde julio de 2014 hasta enero de 2015 por lo que hemos optado por realizar una predicción de esos datos con el modelo Theta y así hemos rellenado los datos faltantes. Una vez hemos hecho el análisis exploratorio y la modificación de los datos, hemos procedido a realizar las predicciones diarias y semanales. Tras realizar varios modelos, hemos llegado a la conclusión de que el modelo que mejor se ajusta a nuestros datos diarios es el modelo Prophet. El modelo que mejor se ajusta a nuestros datos semanales es el ETS.

Con esta información sobre las posibles ventas futuras, la empresa podrá identificar los productos y las tiendas que desempeñan un papel clave en sus ventas y utilizar esa información para tomar las medidas correctas que garanticen el éxito de su negocio.