



CUNEF

MÁSTER UNIVERSITARIO EN CIENCIA DE DATOS

TÉCNICAS DE PREDICCIÓN: SERIES TEMPORALES

Práctica 1

Aitor Larrinoa Rementería

Noviembre 2021

Índice

1. Introducción	3
2. Datos	4
2.1. Limpieza de los datos	4
3. Preparación previa a los modelos	4
3.1. Gráfica de los ingresos	5
3.2. Comportamiento estacional	5
3.3. Elección de training y testing	6
3.4. Apple	6
4. Modelo ETS	8
4.1. Cálculo y análisis de residuos	10
4.2. Predicciones para los años 2021 y 2022	10
5. Modelo ARIMA	11
5.1. Cálculo y análisis de residuos	13
5.2. Predicción para los años 2021 y 2022	13
6. Efecto covid	14
7. Comparación de modelos	15
8. Comparación de las predicciones de Barclays	15
8.1. Comparación mediante ETS	16
8.2. Comparación mediante ARIMA	16

1. Introducción

En este informe, se podrá ver la realización de la primera práctica de la asignatura *Técnicas de predicción: series temporales*.

El principal objetivo de esta práctica se centrará en predecir las ventas de la empresa Apple. Esta predicción se dará mediante dos tipos de modelos. Por un lado hablaremos de los modelos ETS y, por otro lado, hablaremos de los modelos ARIMA. Obtendremos las predicciones de ambos modelos y, después, trataremos de comparar qué modelo se ajusta mejor. Esto será posible analizando los residuos de cada uno de los modelos.

Durante el informe iremos atendiendo cuestiones interesantes como por ejemplo las ventas anuales de Apple, analizaremos el efecto covid y también compararemos nuestras predicciones con las de Barclays Research. Los datos del informe realizado por Barclays Research se pueden encontrar en un archivo en la plataforma canvas.

El trabajo constará de 7 secciones. En la primera trataremos con los datos, veremos de qué forma nos vienen dados y cómo podemos transformarlos para trabajar adecuadamente. Seguiremos con una preparación previa al modelo. En esta sección veremos el gráfico de los ingresos de Apple, analizaremos el comportamiento estacional y haremos la elección del training y el testing. Los dos siguientes capítulos tratan de los modelos ETS y ARIMA respectivamente. En ellos veremos cuales son las predicciones para cada uno de los modelos. Seguiremos analizando el efecto covid y, después compararemos los dos modelos obtenidos atendiendo a los datos obtenidos en los capítulos 4, 5 y 6. Y, por último, finalizaremos con una comparativa de las predicciones obtenidas y las de Barclays.

2. Datos

Los datos con los que trabajaremos son las ventas de la empresa Apple. Los datos se encuentran en el archivo `aapl.csv`, visible en la plataforma canvas. Estos datos contiene ingresos trimestrales desde el último trimestre de 1989 hasta el segundo trimestre de 2021.

Este archivo csv tiene tres columnas y 127 filas. Las columnas son las siguientes:

- **Trimestre:** Trimestre y año de cada dato
- **Fecha:** Fecha de cada dato
- **Ingresos:** La cantidad de ingresos recibida por Apple

La columna de los ingresos está en formato español, es decir, los números decimales aparecen con comas. Esto es incompatible con el programa Python. Luego, deberemos cambiar los decimales con comas a decimales con puntos. Este cambio se ha realizado en el propio archivo csv. Por otro lado, las filas indicarán el ingreso que obtuvo la empresa Apple en una fecha y trimestre determinadas.

2.1. Limpieza de los datos

En primer lugar, se debe realizar una limpieza de los datos. Para el trabajo únicamente nos interesan las columnas segunda y tercera. Es por ello que eliminaremos la columna “trimestre”. De esta forma, nos quedamos únicamente con las columnas “fecha” e “Ingresos”. Además, indicaremos que la columna “fecha” deberá ser el índice del data frame. También hemos comprobado si hay algún *NA*. El resultado ha sido negativo, lo que es satisfactorio.

Así, con el fin de obtener una visión general del data frame, hacemos un `.head()` sobre el mismo y el resultado que obtenemos es el siguiente:

Ingresos	
fecha	
1989-12-29	1493.4
1990-03-30	1346.2
1990-06-29	1364.8
1990-09-28	1354.1
1990-12-28	1675.5

Figura 1: Primeras 5 filas del data frame “apple”

3. Preparación previa a los modelos

Antes de poder introducir los modelos, es conveniente graficar los ingresos, atender al comportamiento estacional y elegir los datos de training y de testing.

3.1. Gráfica de los ingresos

Primeramente, una vez se tienen los datos limpios, haremos el gráfico de la serie temporal de los ingresos de Apple. El resultado es el siguiente:

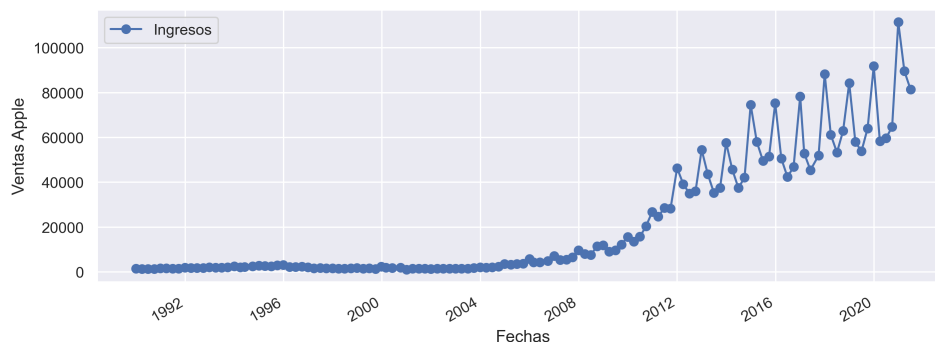


Figura 2: Serie temporal de los ingresos de Apple

Como se puede observar, el incremento comienza a ser visible a partir del año 2010 más o menos. Se puede observar que el rango de los ingresos va desde 0 hasta 25000 durante los primeros 20 años. A partir del año 2010, el rango de los ingresos va desde 20000 hasta 100000 en tan solo 11 años. Luego, como se puede apreciar, el incremento es considerable a partir del año 2010.

Dejando de lado el incremento de Apple a partir del año 2010, la figura 2 nos proporciona todavía más información. Introduzcamos el concepto de serie estacionaria.

Definición. Se dice que una serie es estacionaria cuando la media y la variabilidad de la serie se mantienen constantes a lo largo del tiempo.

En nuestro caso, vemos claramente que la serie no es estacionaria, pues por un lado, la variabilidad de la serie temporal aumenta con el paso del tiempo y, por otro, la serie tiene una clara tendencia positiva. Por ende, la serie temporal que estamos analizando no es estacionaria en varianza ni tampoco en media. Es por ello que podemos asegurar que, la serie temporal de los ingresos de Apple no es estacionaria.

3.2. Comportamiento estacional

Analizar la componente estacional de un gráfico es comprobar si los ingresos de Apple dependen, en este caso, de algún trimestre. Con el fin de analizar la componente estacional, haremos el gráfico estacional de los datos trimestrales. El resultado es el siguiente:

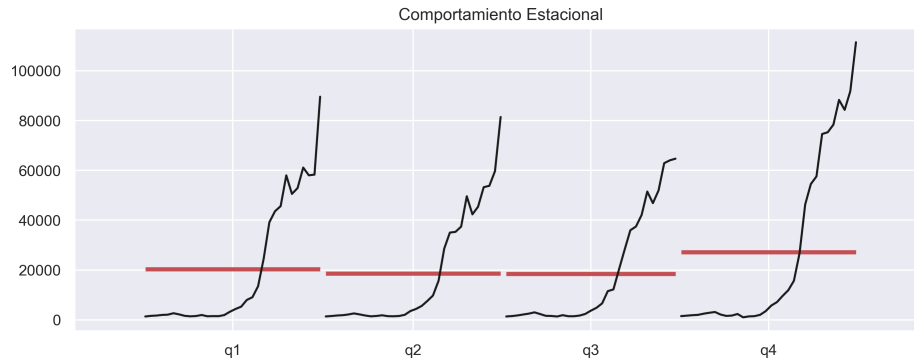


Figura 3: Gráfico estacional de los datos trimestrales

Claramente observamos que el mayor impacto se da en el cuarto trimestre. Recordemos que el cuarto trimestre es el periodo comprendido entre octubre y diciembre. Recordemos que Apple lanza sus nuevos artículos en el mes de octubre. Luego, es intuitivo pensar que los meses más fuertes para la compañía son octubre y noviembre que es cuando la novedad de los artículos de Apple llega a las calles.

3.3. Elección de training y testing

Para comprobar que un modelo funciona correctamente, deberemos separar nuestra base de datos en dos grupos. Estos son el *training* y el *testing*.

El conjunto de training servirá para entrenar al modelo, mientras que el conjunto de testing será útil para ver cuan bueno es el modelo obtenido. Así, dado que el enunciado del ejercicio exige que el conjunto de testing esté compuesto por los últimos 8 trimestres. De esta forma, graficamos el conjunto de training y testing.

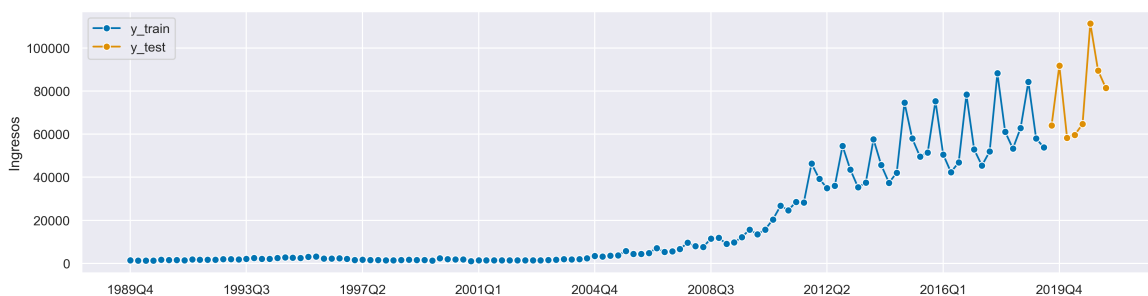


Figura 4: Training y testing

En total se tiene que el conjunto de testing está formado por 119 datos mientras que el conjunto de training está formado por 8, como se ha mencionado.

3.4. Apple

Los datos con los que estamos trabajando son datos cuatrimestrales, pero, ¿cuáles son las ventas anuales de Apple? Para responder a esta pregunta únicamente deberemos sumar los ingresos de la compañía durante los cuatro trimestres de cada año. Los resultados se muestran en la siguiente tabla.

Año	Ingreso	Año	Ingreso	Año	Ingreso
1989	1493,4	2000	6647,0	2011	127841,0
1990	5740,6	2001	5731,0	2012	164687,0
1991	6496,0	2002	5839,0	2013	173992,0
1992	7224,2	2003	6741,0	2014	199800,0
1993	8445,6	2004	9763,0	2015	234440,0
1994	9551,9	2005	16190,0	2016	218118,0
1995	11378,0	2006	20689,0	2017	238536,0
1996	8814,0	2007	27063,0	2018	261612,0
1997	6530,0	2008	38941,0	2019	267683,0
1998	6073,0	2009	46708,0	2020	294135,0
1999	6767,0	2010	76283,0	2021	171018,0

Cuadro 1: Ingresos anuales de la compañía Apple

Para visualizar mejor estos valores, dibujemos una serie temporal anual de estos datos obtenidos. Con el fin de hacer una gráfica consistente, eliminaremos los datos de 1989 y 2021 ya que no tienen sus trimestres completos.

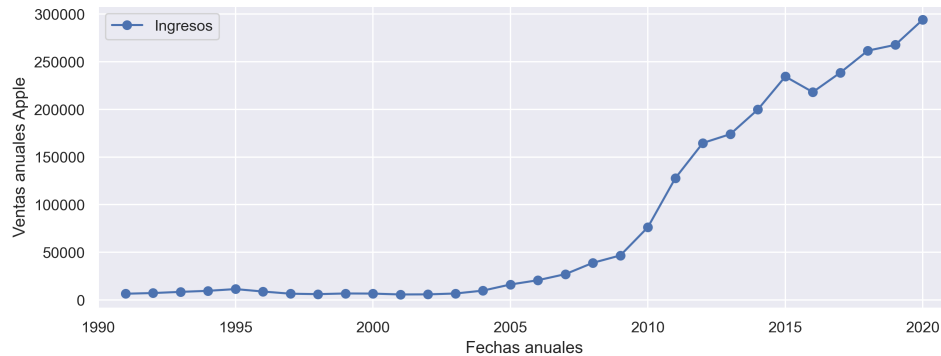


Figura 5: Serie temporal de los ingresos anuales de Apple

Otra forma de ver el incremento de la compañía es calcular los incrementos anuales de los ingresos. Estos incrementos se calculan de manera muy sencilla mediante el comando `pct_change()`. Los incrementos anuales han sido calculados sin tener en cuenta los años 1985 y 2021 ya que carecen de datos datos anuales. Por otro lado, en el gráfico de incrementos que mostraremos a continuación, el año 1990 no aparece ya que el cálculo se hace respecto al año anterior y, como hemos dicho, el dato de 1985 ha sido eliminado.

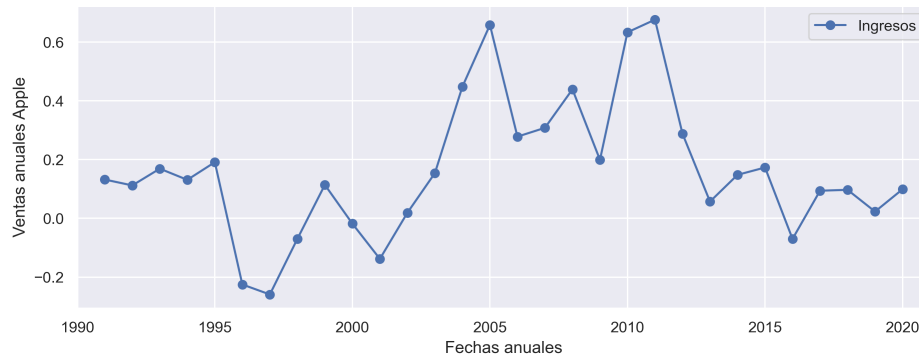


Figura 6: Serie temporal de los rendimientos anuales de Apple

Vemos que tenemos tanto incrementos negativos como positivos pero, predominan los incrementos positivos. También hemos calculado la media de los rendimientos anuales y el valor obtenido es 0.162. Esto certifica el crecimiento que ha tenido la empresa Apple desde el año 1990.

4. Modelo ETS

El modelo ETS es un algoritmo estadístico de uso común para la predicción de series temporales. Cuando un conjunto de datos se dice que tiene estacionalidad, estos modelos son especialmente útiles. En función de la tendencia y la estacionalidad de la serie, tenemos un total de 15 tipos de modelos ETS.

Dado que como hemos podido observar en ??, nuestro conjunto de datos tiene estacionalidad, es correcto hacer uso de este tipo de modelos. Para hallar el mejor de los modelos ETS, haremos uso del paquete *sktime*. Mediante el comando `AutoETS`, el resultado obtenido es el que sigue:

ETS Results

Dep. Variable:	Ingresos	No. Observations:	119			
Model:	ETS(MAM)	Log Likelihood	-968.025			
Date:	Sat, 30 Oct 2021	AIC	1956.050			
Time:	18:06:33	BIC	1983.841			
Sample:	12-31-1989	HQIC	1967.335			
	- 06-30-2019	Scale	0.018			
Covariance Type:	approx					
	coef	std err	z	P> z 	[0.025	0.975]
smoothing_level	0.4225	0.055	7.739	0.000	0.316	0.530
smoothing_trend	0.1685	0.034	4.890	0.000	0.101	0.236
smoothing_seasonal	0.3568	0.077	4.646	0.000	0.206	0.507
initial_level	1370.9326	598.929	2.289	0.022	197.053	2544.813
initial_trend	46.0330	96.115	0.479	0.632	-142.349	234.415
initial_seasonal.0	0.8649	0.393	2.199	0.028	0.094	1.636
initial_seasonal.1	0.8891	0.402	2.210	0.027	0.101	1.678
initial_seasonal.2	0.9181	0.414	2.216	0.027	0.106	1.730
initial_seasonal.3	1.0000	0.441	2.265	0.024	0.135	1.865
Ljung-Box (Q):	5.15	Jarque-Bera (JB):	112.84			
Prob(Q):	0.74	Prob(JB):	0.00			
Heteroskedasticity (H):	1.20	Skew:	-0.39			
Prob(H) (two-sided):	0.57	Kurtosis:	7.71			

Figura 7: Resultado del modelo ETS

El modelo elegido es MAM, tal y como se puede observar en la tabla. Por otro lado, si mostramos las predicciones para el horizonte de predicción elegido, los resultados son los siguientes:

Trimestre	Predicción
2019 Q3	58581,329864
2019 Q4	85963,046083
2020 Q1	59751,803939
2020 Q2	53079,531957
2020 Q3	58323,474571
2020 Q4	85584,248858
2021 Q1	59488,216501
2021 Q2	52845,119874

Cuadro 2: Valores de la predicción de ETS

Lo que haremos ahora será mostrar el gráfico de las predicciones que acabamos de mostrar en la tabla, junto con los valores reales. Consideraremos los datos a partir del año 2014 para así poder ver de una manera más clara las predicciones.

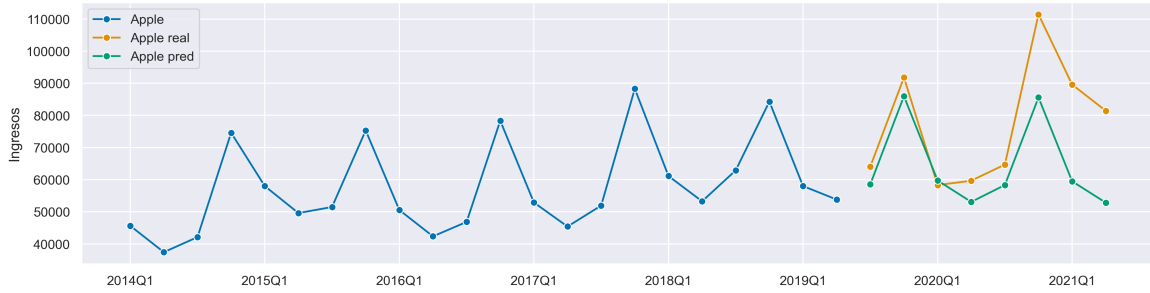


Figura 8: Comparación de la realidad y la predicción de ETS

En los tres primeros trimestres de la predicción el modelo parece que se comporta bien pero, a partir de entonces la diferencia entre la realidad y la predicción comienza a ser notable. Es sorprendente como la intrusión del covid tuvo repercusión positiva en los ingresos de Apple.

4.1. Cálculo y análisis de residuos

Para poder ver la precisión del modelo, deberemos atender al error. El error se define como la diferencia entre el valor observado y el valor esperado y es más comúnmente conocido por *residuo*.

En nuestro caso, para determinar la precisión de las predicciones, haremos uso de 3 medidas, todas ellas indican cifras absolutas. Denotaremos $e_t = y_t - \hat{y}_t$ donde y_t indica el valor real, \hat{y}_t el valor esperado, y e_t el error, donde t varía entre 1 y h . Las medidas de las que haremos uso son:

- Mean absolute error (MAE) = $\frac{\sum_{t=1}^h e_t}{h}$
- Mean squared error (MSE) = $\frac{\sum_{t=1}^h e_t^2}{h}$
- Root mean squared error (RMSE) = \sqrt{MSE}

Aunque en el caso MAE, lo que consideraremos será la medida relativa de ésta (MAPE). Los valores obtenidos en los 3 casos se muestran a continuación.

Medida de error	Valor
MAPE	0,1865
MSE	317746827,889
RMSE	17825,4545

Cuadro 3: Valores de las medidas de error para ETS

Posteriormente, a la hora de calcular los errores del modelo ARIMA, compararemos los errores obtenidos en esta tabla con los errores que obtengamos con el modelo ARIMA.

4.2. Predicciones para los años 2021 y 2022

Una vez tenemos el modelo ETS, realizaremos las predicciones de lo que resta de año y del año 2022. Es decir, prediciremos un total de 6 trimestres. A continuación mostramos los

valores de las predicciones.

Trimestre	Predicción
2021 Q3	94928,198
2021 Q4	153907,757
2022 Q1	112068,412
2022 Q2	104446,064
2022 Q3	120404,572
2022 Q4	192615,717

Cuadro 4: Resultados de la predicción para 2021 y 2022 mediante ARIMA

Con el fin de poder ver estas predicciones de manera gráfica, vamos a realizar el gráfico de las mismas para ver su tendencia. Lo haremos a partir del año 2014 para verlo de manera más clara.

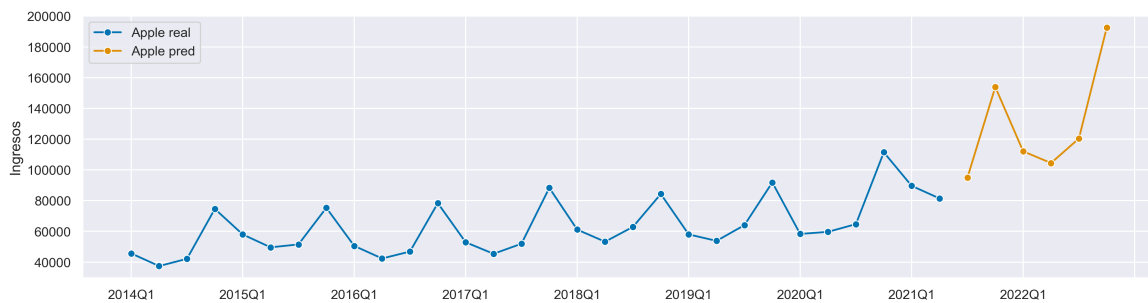


Figura 9: Predicciones para los años 2021 y 2022 mediante ETS

Vemos que las predicciones son muy positivas para la empresa.

5. Modelo ARIMA

Autoregressive Integrated Moving Average, al igual que ETS, es un algoritmo estadístico que sirve para predecir series temporales.

AAAAAAAAAa

De esta forma, tenemos el training y el testing elegidos para la transformación logarítmica de los datos y, aplicando a los datos la función *AutoARIMA*, el resultado obtenido es:

```

=====
SARIMAX Results
=====
Dep. Variable:          y          No. Observations:          119
Model:                 SARIMAX(0, 1, 1)x(1, 0, 1, 4)          Log Likelihood          58.362
Date:                  Sun, 31 Oct 2021          AIC          -108.724
Time:                  18:06:20          BIC          -97.642
Sample:                0          HQIC          -104.224
Covariance Type:      opg
=====
              coef      std err          z      P>|z|      [0.025      0.975]
-----
ma.L1          -0.3363      0.089      -3.795      0.000      -0.510      -0.163
ar.S.L4         0.9156      0.062     14.822      0.000       0.795       1.037
ma.S.L4        -0.4810      0.106     -4.547      0.000      -0.688      -0.274
sigma2         0.0211      0.002     12.465      0.000       0.018       0.024
=====
Ljung-Box (L1) (Q):          0.00      Jarque-Bera (JB):          552.07
Prob(Q):                    0.98      Prob(JB):              0.00
Heteroskedasticity (H):      1.49      Skew:              -1.49
Prob(H) (two-sided):         0.22      Kurtosis:             13.17
=====

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).

```

Figura 10: Resultados del modelo ARIMA

El modelo seleccionado es $(p, d, q)x(P, D, Q) = (0, 1, 1)x(1, 0, 1, 4)$. El último 4 del segundo paréntesis indica que los datos analizados son trimestrales. (p, d, q) hace referencia a si los datos son regulares, es decir, si la variable aleatoria analizada depende del trimestre anterior. Por otro lado, (P, D, Q) indica si la variable aleatoria depende del año anterior.

Para obtener las predicciones se debe deshacer el cambio logarítmico aplicando la función inversa del logaritmo neperiano a los datos, la exponencial. De esta manera, los valores de las predicciones obtenidas son:

Trimestre	Predicción
2019 Q3	60183,689
2019 Q4	83757,702
2020 Q1	61526,150
2020 Q2	56094,944
2020 Q3	62284,130
2020 Q4	84295,514
2021 Q1	63554,963
2021 Q2	58398,533

Cuadro 5: Valores de la predicción de ETS

Dibujemos estos datos en un gráfico para ver cuan buenos son frente a la realidad. Los dibujaremos a partir de 2014.

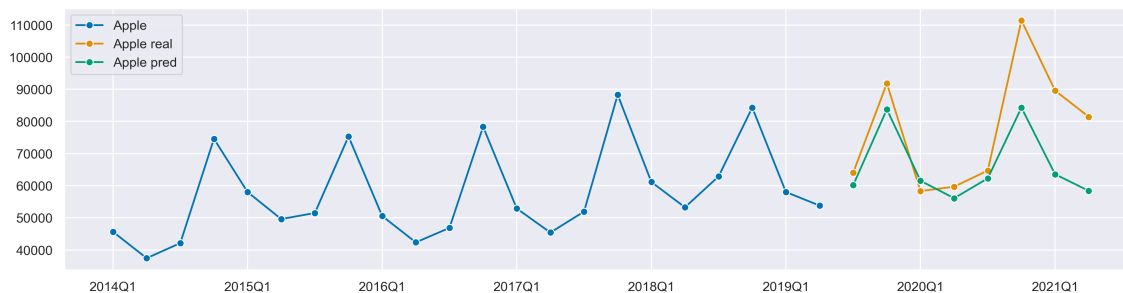


Figura 11: Comparación de la realidad y la predicción de ARIMA

De igual forma que con el modelo ETS, la realidad se aleja de la predicción a partir del quinto trimestre del horizonte de predicción. Esto se puede deber al sorprendente incremento de la empresa en tiempos de covid.

5.1. Cálculo y análisis de residuos

Lo más importante para los modelos ARIMA es que los residuos sean ruido blanco ¹.

Medida de error	Valor
MAPE	0,157
MSE	256725946,507
RMSE	16022,67

Cuadro 6: Valores de las medidas de error para ARIMA

Si comparamos los valores obtenidos con los obtenidos en 3, podemos comprobar que los errores obtenidos para ARIMA son menores, en todas las medidas de error, que los obtenidos para ETS.

5.2. Predicción para los años 2021 y 2022

Al igual que se ha hecho para el modelo ETS, se deberán calcular las predicciones para los años 2021 y 2022. Para ello, lo primero que haremos será tomar todos los datos como conjunto de training. Como hemos visto, se debe trabajar con el logaritmo para conseguir una serie estacionaria en varianza, es por ello que consideraremos como conjunto de entrenamiento todos los datos disponibles en formato logaritmo. De esta forma, y considerando un horizonte de predicción de 6 trimestres, el resultado de las predicciones que se obtiene es el siguiente:

¹El ruido blanco, tiene la característica que su acf y pacf son cero para todos los retardos, por lo tanto su historia no explica el comportamiento futuro de la serie

Trimestre	Predicción
2021 Q3	89906,424
2021 Q4	133552,661
2022 Q1	101656,988
2022 Q2	95662,618
2022 Q3	105102,498
2022 Q4	152053,519

Cuadro 7: Resultados de la predicción para 2021 y 2022 mediante ARIMA

Veamos los resultados gráficamente.

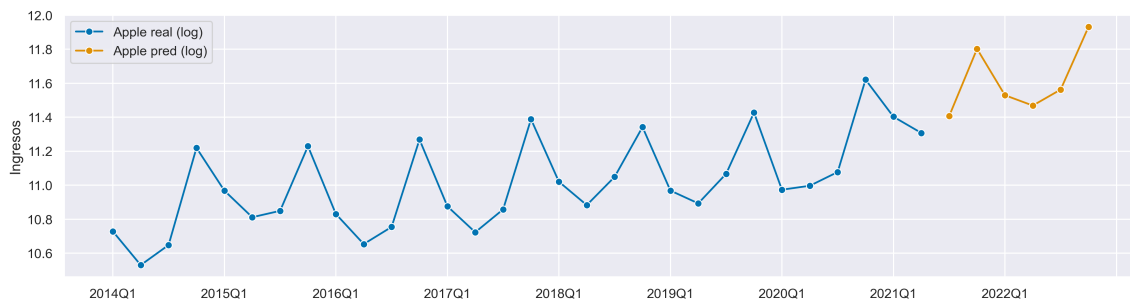


Figura 12: Gráfico de la predicción para 2021 y 2022 mediante ARIMA

6. Efecto covid

Hemos visto en la figura 2 que el covid ha irrumpido de manera positiva en los ingresos de la empresa Apple, sorprendentemente. También hemos visto en la figura 8 que el modelo se aleja un poco de la realidad. Pero, ¿cuánto se aleja? Cuantifiquemos el efecto covid con los modelos estudiados y comparemos los resultados.

Lo que haremos será calcular la diferencia entre el valor real y el valor predicho desde el primer trimestre del año 2020 hasta el cuarto trimestre de ese mismo año. A continuación mostramos una tabla comparativa de los resultados.

Trim fiscal	Valor real	Pred ETS	Pred ARIMA	Res ETS	Res ARIMA
2020 Q1	91819,0	85963,046	83757,702	5855,954	8061,298
2020 Q2	58313,0	59751,8	61526,150	-1438,8	-3213,15
2020 Q3	59685,0	53079,532	56094,944	6605,468	3590,056
2020 Q4	64698,0	58323,475	62284,130	6374,525	2413,87

Cuadro 8: Resultados de la predicción para 2021 y 2022 mediante ARIMA

Vemos que los residuos son menores con el modelo ETs para los trimestres 1 y 2, mientras que los residuos son menores con el modelo ARIMA para los trimestres 3 y 4.

7. Comparación de modelos

Hemos podido comprobar que el error obtenido en el modelo ARIMA es inferior al obtenido en el modelo ETS. Véase los cuadros 3 y 10. Dado que la elección de un modelo frente a otro se realiza en función del modelo con menor error, en este caso, podremos asegurar que el mejor de los dos modelos es ARIMA. De hecho, veamos en el siguiente gráfico como se ajusta mejor el modelo ARIMA.

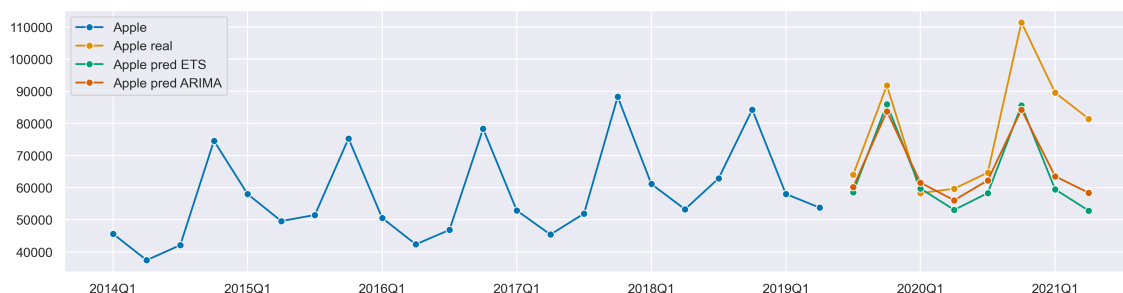


Figura 13: Gráfico de los dos modelos junto con la realidad

Por otro lado, también se ha cuantificado el efecto covid en ambos modelos. Si atendemos a esto, nos quedamos con el modelo

8. Comparación de las predicciones de Barclays

En la plataforma canvas se encuentra un documento pdf que habla de las predicciones de Apple realizadas por Barclays. Veamos si sus predicciones se parecen a las nuestras o si por el contrario son muy diferentes.

Lo primero que ha de tenerse en cuenta es que las predicciones de Barclays están hechas en años fiscales mientras que las predicciones nuestras han sido realizada teniendo en cuenta los años naturales. Sin embargo, esto no es un problema ya que se pueden transformar nuestros datos en años naturales a años fiscales. Una vez transformados, podemos observar que los datos ya no acaban en el segundo trimestre de 2021 sino que acaban en el tercero, véase la siguiente tabla en la que se muestra el data frame en años fiscales.

Ingresos	
fecha	
2020Q1	91819.0
2020Q2	58313.0
2020Q3	59685.0
2020Q4	64698.0
2021Q1	111439.0
2021Q2	89584.0
2021Q3	81434.0

Figura 14: Últimas 7 filas del data frame trimestral en años fiscales

Dado que Barclays nos proporciona los ingresos anuales, lo que debemos hacer es mostrar nuestros datos también en formato anual. Deberemos añadir, a la tabla anterior, nuestras

predicciones para 2021 y 2022 para de esa manera poder hallar los ingresos anuales y comparar finalmente con las predicciones de Barclays.

8.1. Comparación mediante ETS

Veamos cuan bien se ajustan las predicciones del modelo ETS con las de Barclays. La siguiente tabla muestra los ingresos anuales de Apple en el año 2020 y las predicciones de los años 2021, 2022 y 2023.

Año fiscal	Ingreso
2020	274515,0
2021	377385,198
2022	490826,805
2023	458982,216

Cuadro 9: Ingresos anuales de Apple mediante modelo ETS

El valor esperado del ingreso anual de 2021 se separa en dos sumandos. Por un lado el valor conocido de los tres primeros trimestres (282457,0) y el valor esperado del último trimestre (94928,198).

Comparando estos resultados con los de Barclays, podemos sacar las siguientes conclusiones:

- La predicción obtenida para 2021 es muy parecida, pero esto era de esperar ya que conocemos los datos de los 3 primeros trimestres y únicamente se ha hecho la predicción del último trimestre.
- Por otro lado, se puede observar que las predicciones que obtenemos para los años 2022 y 2023 son muy diferentes a las obtenidas por Barclays, hay un gran salto entre ambas.

Veamos si el modelo ARIMA se ajusta mejor que el ETS a las predicciones de Barclays.

8.2. Comparación mediante ARIMA

En esta ocasión, haremos uso de las predicciones halladas con el modelo ARIMA. En la sección anterior hemos visto que el modelo ARIMA era mejor que el modelo ETS pero, ¿serán las predicciones ARIMA más precisas que las predicciones ETS? Veámoslo. A continuación mostramos los ingresos anuales de Apple en el año 2020 y las predicciones de los ingresos anuales de Apple en los años 2021, 2022 y 2023.

Año fiscal	Ingreso
2020	274515,0
2021	372363,424
2022	435974,765
2023	502882,931

Cuadro 10: Ingresos anuales de Apple mediante modelo ETS

El valor esperado del ingreso anual de 2021, al igual que antes, se separa en dos sumandos. Por un lado el valor conocido de los tres primeros trimestres (282457,0) y el valor esperado del último trimestre ().