PRÁCTICA 3
INFORME

Análisis de las consultas de Matlab en StackOverflow

Marsá Martín, Isabel Fátima

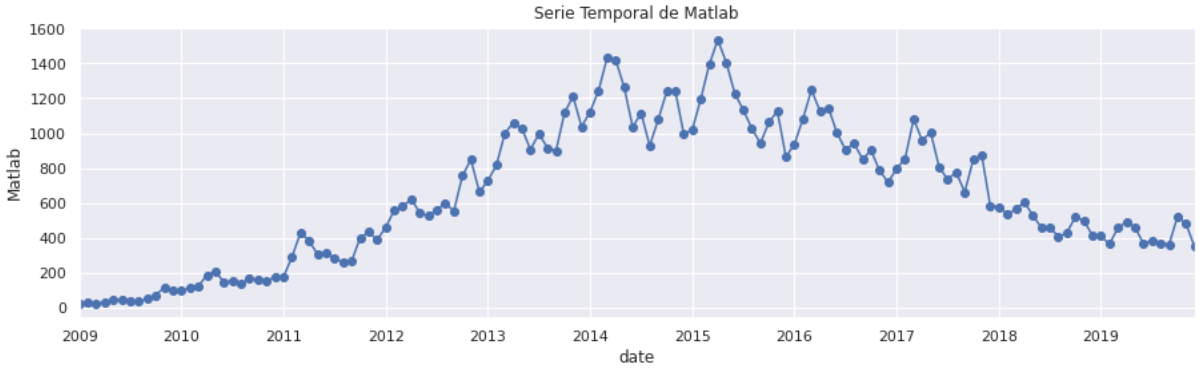
Análisis Predictivo | Noviembre 2022

Introducción

StackOverflow es una página web que contiene preguntas y respuestas sobre diversos temas de programación. Fue creada en 2008 y actualmente, es la mayor comunidad online para que los programadores aprendan, compartan sus conocimientos y avancen en sus carreras.

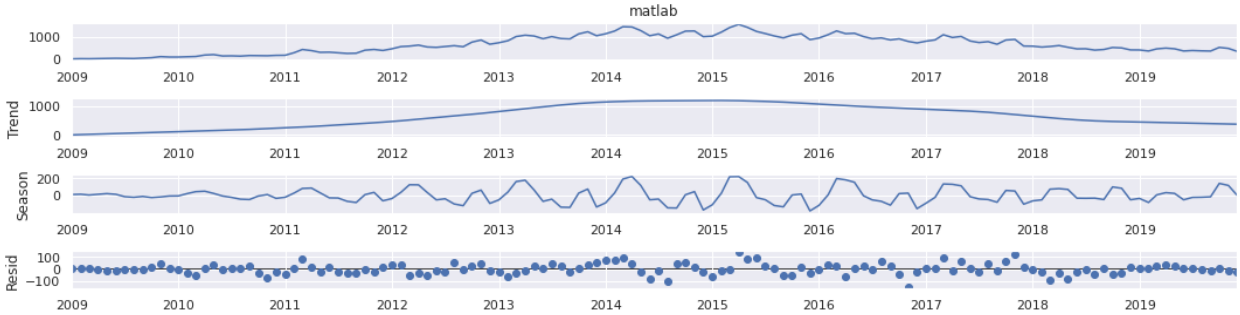
En este análisis trabajaremos sobre un dataset donde se encuentran datos sobre el número de nuevas consultas que se realizan mensualmente para los diferentes lenguajes de programación. Contiene datos desde enero de 2009 hasta diciembre de 2019. El fichero (StackOverflow.csv) contiene 82 variables, entre las que se encuentran la fecha cuando se registraron los datos y 81 lenguajes de programación distintos.

Nos centraremos en la evolución de las preguntas sobre **Matlab** y construiremos un modelo predictivo de series temporales.

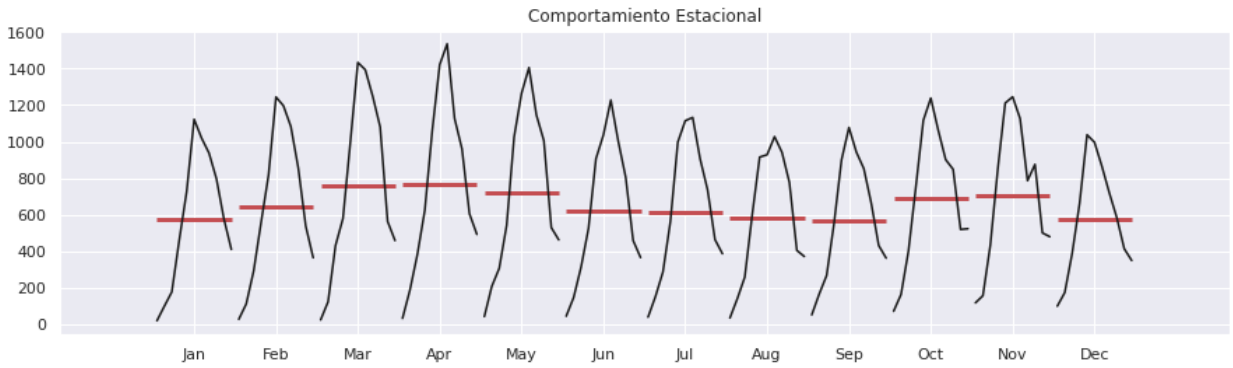


Observamos la evolución de las preguntas sobre Matlab en StackOverflow a lo largo de los años. Podemos ver que, desde 2009 hasta abril de 2015, donde observamos un pico que alcanza las 1535 consultas, la tendencia es creciente. A partir de esa fecha, las consultas nuevas bajan gradualmente hasta diciembre de 2019 que las consultas son tan solo 349.

Estacionariedad y estacionalidad



El grafico superior nos muestra que la serie temporal de los datos es **no estacionaria** porque la tendencia y la variabilidad cambian a lo largo de la serie. Los cambios en la media determinan una tendencia a crecer o decrecer a largo plazo, por lo que la serie no oscila alrededor de un valor constante. Además, podemos comprobar en el tercer gráfico (season) que la serie tiene **estacionalidad** porque podemos observar que existe un patrón de variación periódica predecible cada año. Estos patrones indican que los meses de junio, julio y agosto las nuevas consultas de Matlab decrecen, lo cual se debe al efecto del verano cuando menos personas utilizan la herramienta de programación. En cambio, los meses de marzo, abril y noviembre es cuando más consultas se realizan. A continuación, podemos observar en el gráfico de comportamiento estacional la media de consultas por mes.



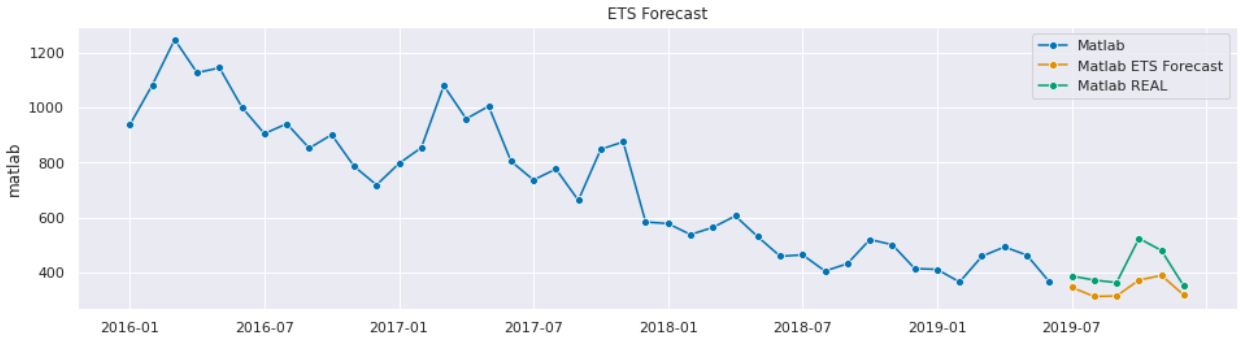
Modelos de Predicción

Para obtener el accuracy de las predicciones de los modelos utilizaremos las métricas de MAPE (mean average percentage error) y RMSE (root mean square error).

Modelo ETS

El algoritmo ETS es especialmente útil para conjuntos de datos con estacionalidad y otras suposiciones previas sobre los datos. ETS calcula un promedio ponderado sobre todas las observaciones en el conjunto de datos de las series temporales de entrada como su predicción. Las ponderaciones disminuyen exponencialmente con el tiempo, en lugar de las ponderaciones constantes en los métodos de promedio móvil simple. Las ponderaciones dependen de un parámetro constante, conocido como parámetro de suavizamiento.

Hemos utilizado AutoETS de la librería Sktime para construir este modelo. Los resultados que hemos obtenido incluyen un AIC de 1417.907. Se ha establecido un horizonte de predicción igual al tamaño de nuestra muestra de test, que es igual a 6 meses.



Con esta predicción hemos obtenido un MAPE de 0.16278756408015796 y un RMSE de 81.69177527717292.

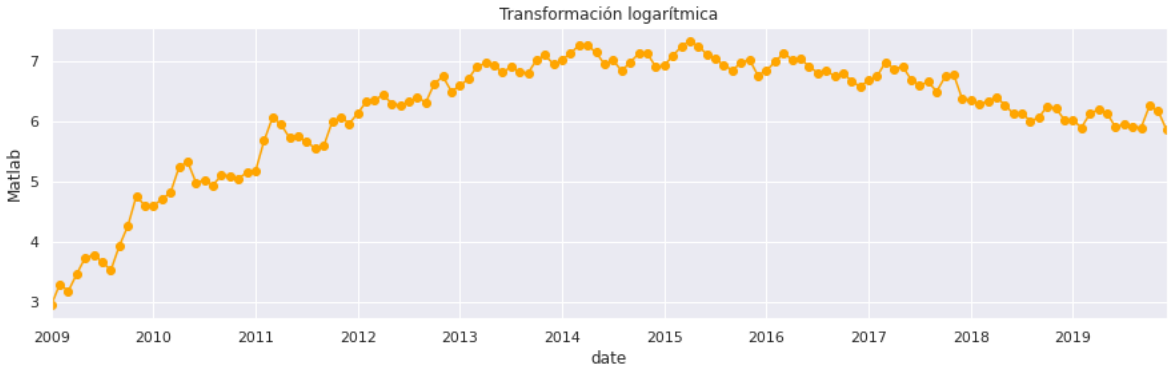
Modelo ARIMA

Este modelo es útil para predecir series temporales en los que la media y varianza no son constantes a lo largo del tiempo (t), es decir, series temporales no estacionarias que no presenten estacionalidad tampoco. ARIMA significa auto regressive integrated moving average y es igual que un modelo ARMA pero en vez de predecir la serie temporal en sí, predice la serie temporal

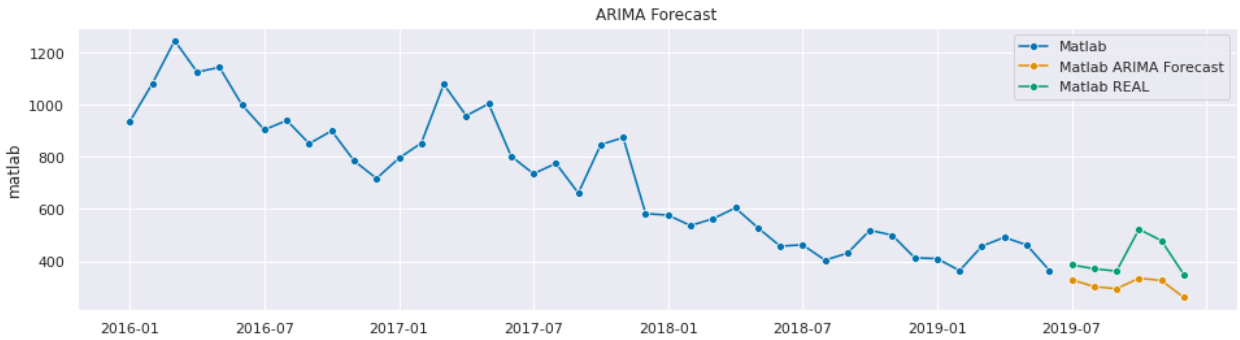
de las diferencias. Esa es la razón por la que se hace una transformación, en este caso logarítmica en el que se crea un nuevo time series.

En un modelo ARMA predeciríamos a_t pero en un modelo ARIMA predeciremos z_t, que es la diferencia de a, es decir, a_{t+1}-a_t. En este caso utilizaremos ARIMA (1,1,1) ya que es el más simple y utiliza la primera diferencia.

La transformación Box-Cox es una transformación potencial que corrige la asimetría de las varianzas. Cuando el parámetro de transformación lambda es igual a 0 se realiza una transformación logarítmica y cuando es distinto de 0, se utilizaría otro tipo de transformación. En este caso hemos realizado una transformación logarítmica con la función LogTransformer() de Sktime y hemos obtenido la siguiente serie temporal.

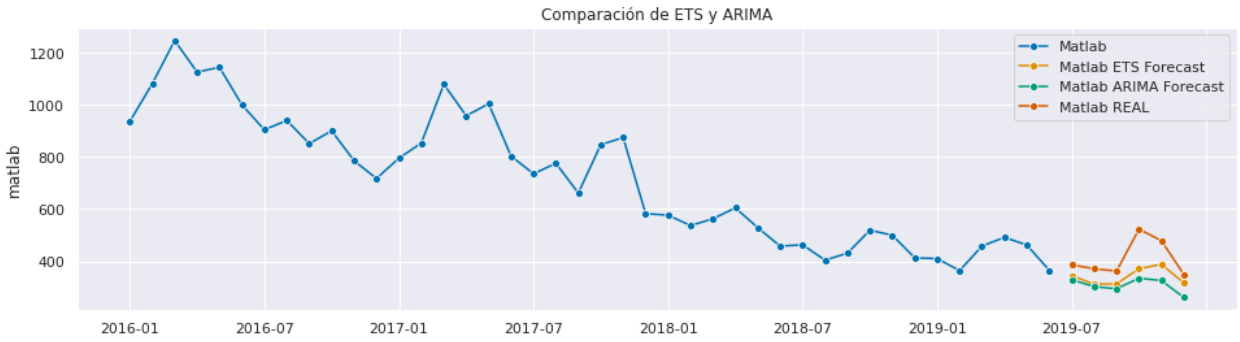


Los resultados del modelo incluyen un AIC de - 127.332, lo cual es un valor muy bueno. La predicción que obtenemos es la siguiente:



Con esta predicción hemos obtenido un MAPE de 0.24206452411181997 y un RMSE de 114.9161009047849.

A continuación, comparamos los modelos ETS y ARIMA.



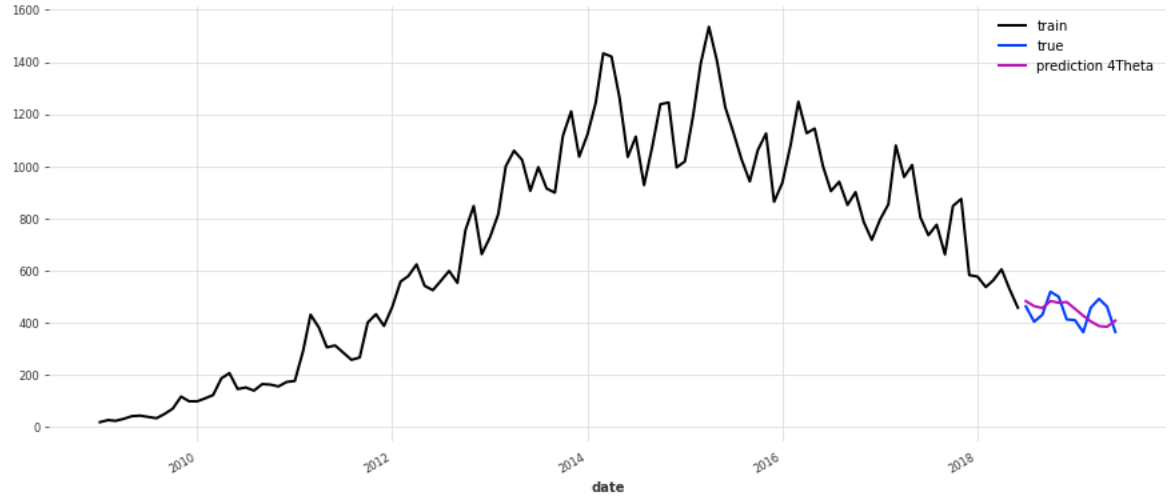
Metric	ETS	ARIMA
MAPE	0.162788	0.242065
RMSE	81.691775	114.916101

Entre ETS y ARIMA escogeríamos ETS ya que los errores de predicción son mucho menores.

Modelo 4Theta

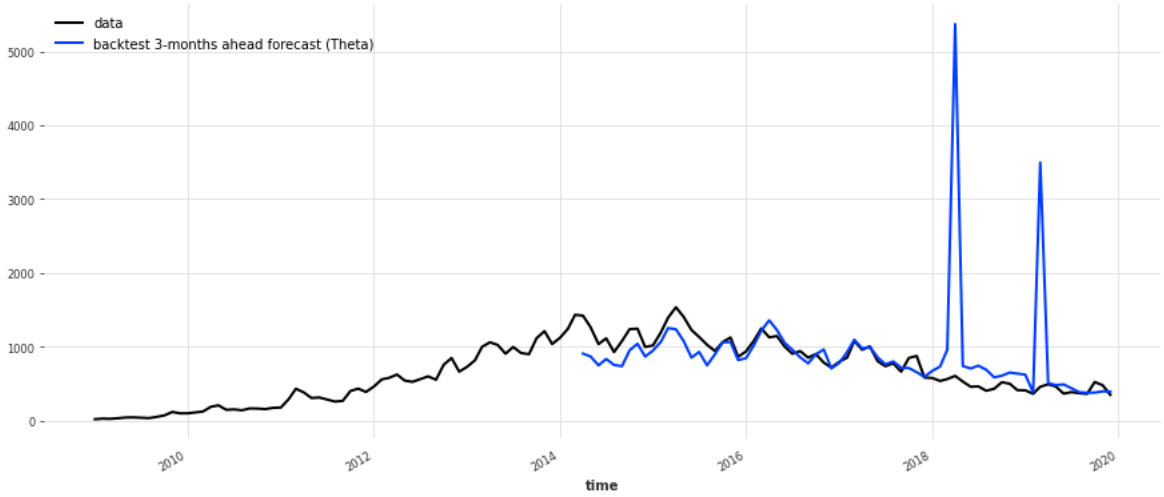
Para construir este modelo será necesario utilizar la librería Darts. El modelo Theta se basa en la descomposición de las series temporales: la tendencia, los componentes estacionales y los residuos. Esta descomposición se realiza para combinar las predicciones de cada componente u asi obtener las predicciones finales. Por lanto el modelo Theta se basa en el concepto de modificar la curvatura local de la serie temporal. Esta modificación es manejada por un parámetro al que conocemos por *Theta*. Esta modificación es aplicada sobre la segunda diferencia de la serie, es decir, que es diferenciada dos veces. Cuando theta se encuentra entre 0 y 1, la serie es deflactada, es decir, las fluctuaciones a corto plazo son menores y nos centramos en efectos a largo plazo. Cuando theta llega a 0, la serie es transformado en una regresión linear. Cuando theta es mayor que 1, las fluctuaciones a corto plazo son agrandadas, por lo que nos centramos en los efectos a largo plazo y por ello decimos que la serie es inflada. En la practica se utilizan dos thetas, un theta igual a 0 y otro igual a 2. El modelo theta es utilizado sobre datos no estacionales. Por ello, en este caso hacemos una transformación sobre los datos ya que son estacionales como hemos podido comprobar. Primero, eliminamos la estacionalidad con el

parámetro seasonality_period, al que le asignamos un valor de 12, ya que nuestros datos están en meses y por tanto hay 12 periodos al año. Descomponemos en dos líneas theta, por defecto. Con la muestra de validación (val) probamos con distintos valores theta y escoge el mejor de todos (best_theta). Sobre la muestra de test, establecemos el valor de theta que ha escogido y realizamos la predicción. Con la función de gridsearch de FourTheta obtenemos los modos optimos, en este caso un modelo ADITIVO, una estacionalidad ADITIVA, y una tendencia EXPONENCIAL. Establezco estos parámetros y los inserto en mi nuevo modelo. Obtenemos la predicción que se muestra a continuación.



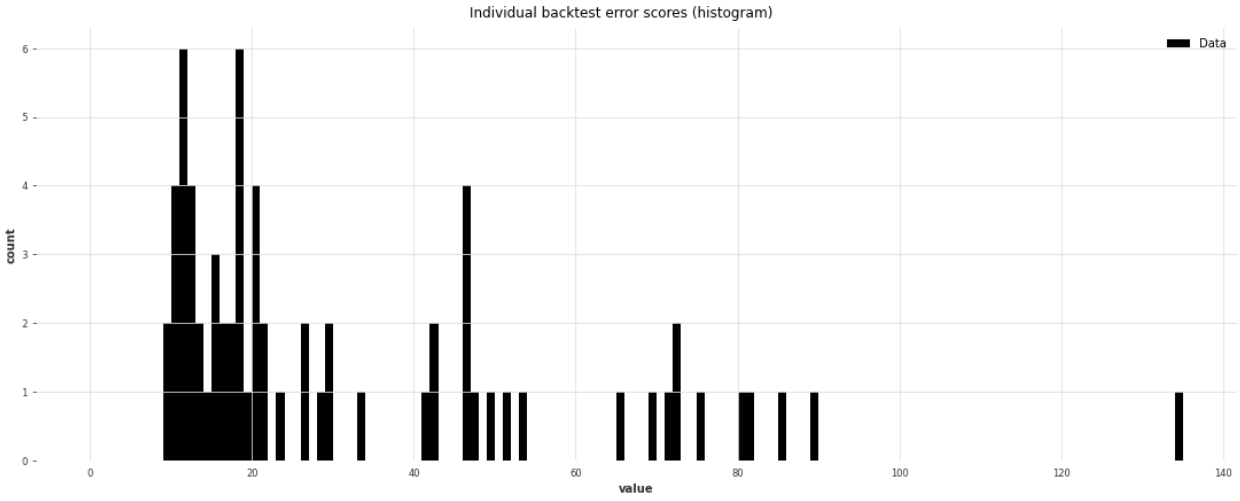
Para evaluar el modelo que hemos construido utilizaremos las herramientas de Historical Forecast y Backtest, con los que predeciremos el 60% estableciendo un parámetro de start igual a 0.4. Cuanta menos dispersión tenga el MAPE más robusta será la predicción.

Historical Forecast



Con historical forecast obtenemos un MAPE igual a 19.65%.

Backtest

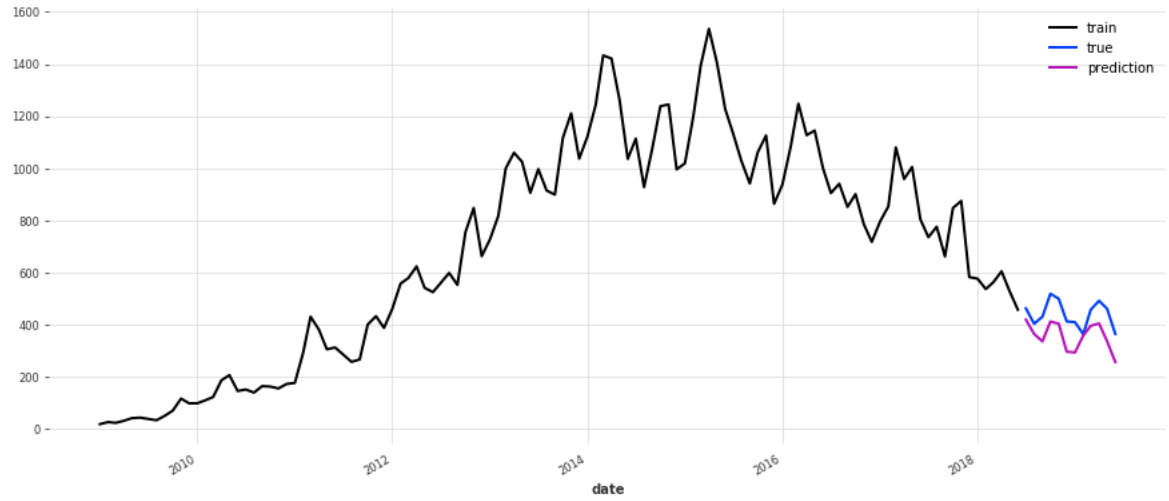


El MAPE sobre todas las predicciones históricas (historical forecasts) son igual a 33.85%

Modelo TBATS

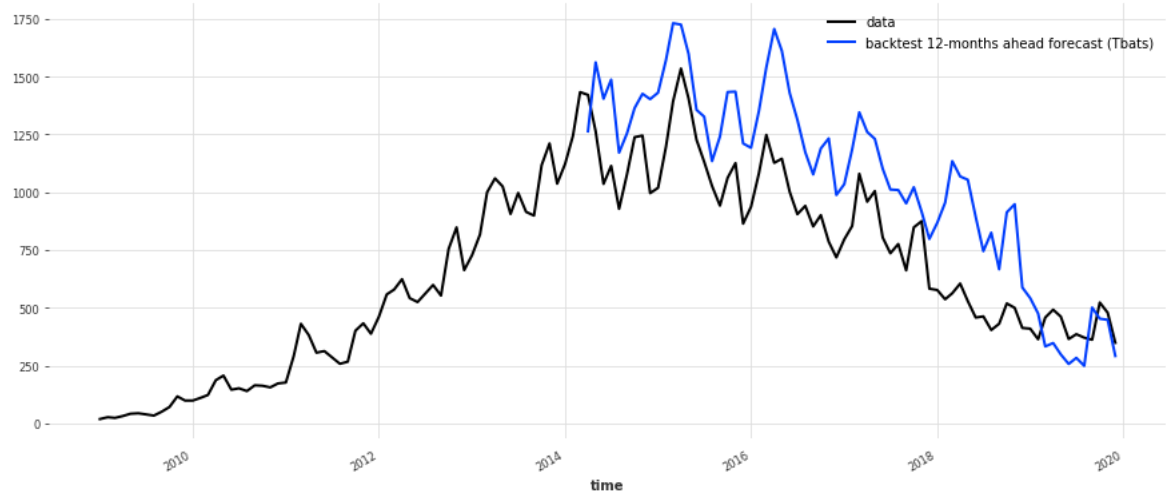
Es un modelo muy útil para datos con más de una estacionalidad, es decir, para datos diarios o semanales entre otros. En este caso, los datos son mensuales por lo que únicamente tienen una estacionalidad por año. El objetivo de TBATS es predecir series temporales con patrones estacionales complejos utilizando exponential smoothing. TBATS es el acrónimo de

Estacionalidad trigonométrica, Transformación Box-Cox, Errores ARMA, Tendencia y Componentes estacionales. Primero, se aplica una transformación Box-Cox a la serie temporal original, y luego ésta se modela como una combinación lineal de una tendencia suavizada exponencialmente, un componente estacional y un componente ARMA. Los componentes estacionales se modelan mediante funciones trigonométricas a través de series de Fourier. Dentro de la función TBATS insertamos los hiperparámetros para que seleccione el modelo final y este modelo final se escoge utilizando AIC siempre.

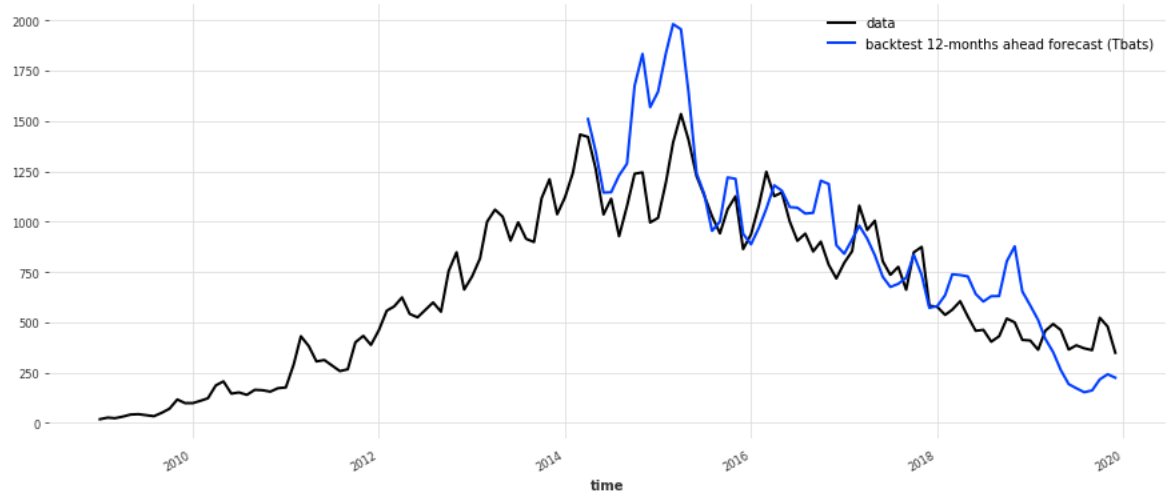


Cross Validation: Historical

Calcula las previsiones históricas que habría obtenido este modelo sobre la serie. Este método construye repetidamente un conjunto de entrenamiento desde el principio de la serie. Entrena el modelo en el conjunto de entrenamiento, emite una previsión de longitud igual al forecast_horizon (fh), y luego mueve el final del conjunto de entrenamiento hacia adelante por pasos de tiempo. Por defecto, este método siempre reentrena los modelos en todo el historial disponible, lo que corresponde a una estrategia de ventana expansiva. Tras definir el modelo inicial obtenemos la predicción siguiente, con un MAPE de 27.12%.



A continuación se muestra la predicción del modelo que TBATS ha escogido como el mejor (model_best) que tiene un MAPE de 27.46%. Podemos comprobar que el MAPE es un poco más alto que el anterior. Por esto, creemos que el TBATS no es el modelo adecuado para predecir en este caso.

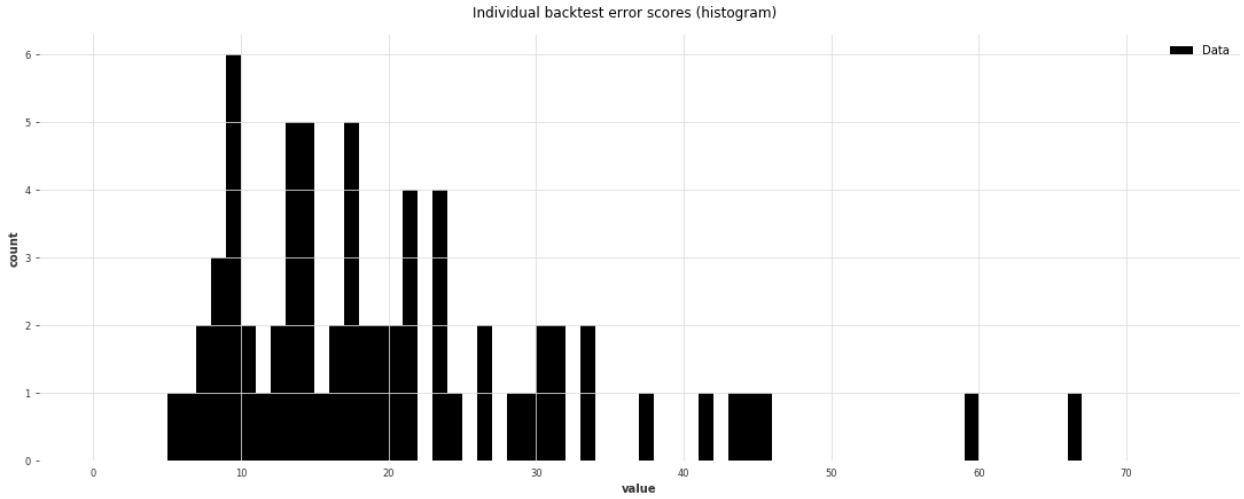


Backtest

Esta herramienta calcula los valores de error que el modelo habría producido al ser utilizado en las series.

Construye repetidamente un conjunto de entrenamiento desde el principio de la serie. Entrena el modelo actual en el conjunto de entrenamiento, emite una previsión de longitud igual al forecast_horizon (fh) y, a continuación, adelanta el final del conjunto de entrenamiento en pasos

de tiempo. A continuación, se evalúa una métrica (dada por la función métrica) sobre la previsión y los valores reales. Finalmente, el método devuelve una reducción (la media por defecto) de todas estas puntuaciones métricas. Por defecto, este método siempre reentrena los modelos en todo el historial disponible, lo que corresponde a una estrategia de ventana expansiva. Nos devolverá el error score.



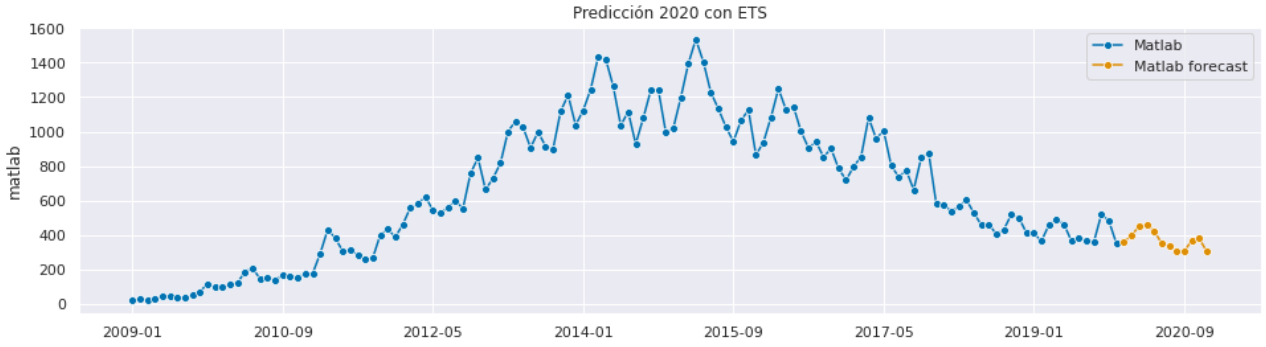
El MAPE sobre todas las predicciones históricas (historical forecasts) son igual a 21.49%

Predicción para 2020 con ETS

Tras comparar los cuatro modelos, hemos escogido realizar la predicción para 2020 con ETS. A continuación, se muestran los resultados obtenidos y los gráficos.

```python
# ETS Model
fh = np.arange(1, 13) #12 meses
matlab_ETS_model = AutoETS(auto=True, sp=12, n_jobs=-1)
matlab_ETS_model.fit(sof_matlab_m)
print(matlab_ETS_model.summary())
```

```
                                ETS Results
==============================================================================
Dep. Variable:                  matlab   No. Observations:                  132
Model:                      ETS(MAM)   Log Likelihood                -721.978
Date:                Mon, 21 Nov 2022   AIC                           1479.956
Time:                        11:02:03   BIC                           1531.846
Sample:                    01-31-2009   HQIC                          1501.042
                         - 12-31-2019   Scale                            0.015
Covariance Type:                approx
==============================================================================
                         coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
smoothing_level         0.8776      0.104      8.426      0.000       0.673       1.082
smoothing_trend         0.0179      0.016      1.098      0.272      -0.014       0.050
smoothing_seasonal   1.224e-05        nan        nan        nan         nan         nan
initial_level          11.6855        nan        nan        nan         nan         nan
initial_trend           6.8762        nan        nan        nan         nan         nan
initial_seasonal.0      0.9735        nan        nan        nan         nan         nan
initial_seasonal.1      1.1984        nan        nan        nan         nan         nan
initial_seasonal.2      1.1435        nan        nan        nan         nan         nan
initial_seasonal.3      0.9351        nan        nan        nan         nan         nan
initial_seasonal.4      0.9262        nan        nan        nan         nan         nan
initial_seasonal.5      1.0068        nan        nan        nan         nan         nan
initial_seasonal.6      1.0436        nan        nan        nan         nan         nan
initial_seasonal.7      1.2360        nan        nan        nan         nan         nan
initial_seasonal.8      1.3257        nan        nan        nan         nan         nan
initial_seasonal.9      1.2800        nan        nan        nan         nan         nan
initial_seasonal.10     1.1215        nan        nan        nan         nan         nan
initial_seasonal.11     1.0000        nan        nan        nan         nan         nan
===================================================================================
Ljung-Box (Q):                       27.53   Jarque-Bera (JB):                 32.06
Prob(Q):                              0.28   Prob(JB):                         0.00
Heteroskedasticity (H):               0.31   Skew:                             0.58
Prob(H) (two-sided):                  0.00   Kurtosis:                         5.12
===================================================================================
```



2020-01	359.566425
2020-02	398.479076
2020-03	449.348217
2020-04	459.754661
2020-05	423.397142
2020-06	353.065656
2020-07	336.321953
2020-08	305.485289
2020-09	304.429659
2020-10	367.417529
2020-11	379.957712
2020-12	304.512173

Conclusión

El número de nuevas consultas sobre Matlab en la plataforma de StackOverflow incrementó exponencialmente desde 2009 hasta 2015 y desde 2015 comenzó a decrecer. Se trata por tanto de una serie no estacionaria y que presenta estacionalidad, ya que hay un patrón en los datos que se repite cada año. Hemos comparado cuatro modelos de predicción distintos: ETS, ARIMA, 4THETA y TBATS. Comparando la métricas de evaluación MAPE, hemos obtenido que el mejor modelo para predecir esta serie temporal es ETS, ya que tiene el MAPE más bajo.