



SERIES TEMPORALES PARA
UNA COMPAÑÍA DE LOGISTICA

BIG DATA, QUÉ ES Y PARA QUE SIRVE EL ALGORITMO DE TIME SERIES

PREDECIR EL N° DE REPARTOS A REALIZAR
POR UNA EMPRESA DE TRANSPORTES



ANTONIO JUSTICIA
www.go-insurances.com

Una serie temporal es un conjunto de muestras tomadas en diferentes intervalos de tiempo, que nos van a permitir analizar su comportamiento a corto y medio plazo.

Para ello hay que realizar un estudio en el que vamos a detectar patrones y hacer diferentes pronósticos de cómo será su comportamiento en el futuro, esta información se puede comportar de diferentes maneras a lo largo del tiempo, ya que puede presentar tendencias, estacionalidad, etc.

El resultado obtenido nos va a permitir estimar lo que va a suceder en el futuro, por ejemplo:

- ✓ El precio o la demanda de un producto o servicio.
- ✓ Morosidad de los clientes.
- ✓ Volumen de mercancías a repartir.
- ✓ Funcionamiento de una maquina antes de necesitar un mantenimiento.



La importancia del dato en una empresa.

Para entenderlo mejor vamos a verlo con un ejemplo que hemos implementado recientemente para una empresa de logística que trabaja con diferentes empresas de reparto que les hacen llegar las mercancías a los clientes finales.

Nuestro cliente tiene la necesidad de predecir el número de repartos a realizar por las empresas de transporte con las que trabaja a una semana vista, es decir los repartos que se realizarán en la siguiente semana.

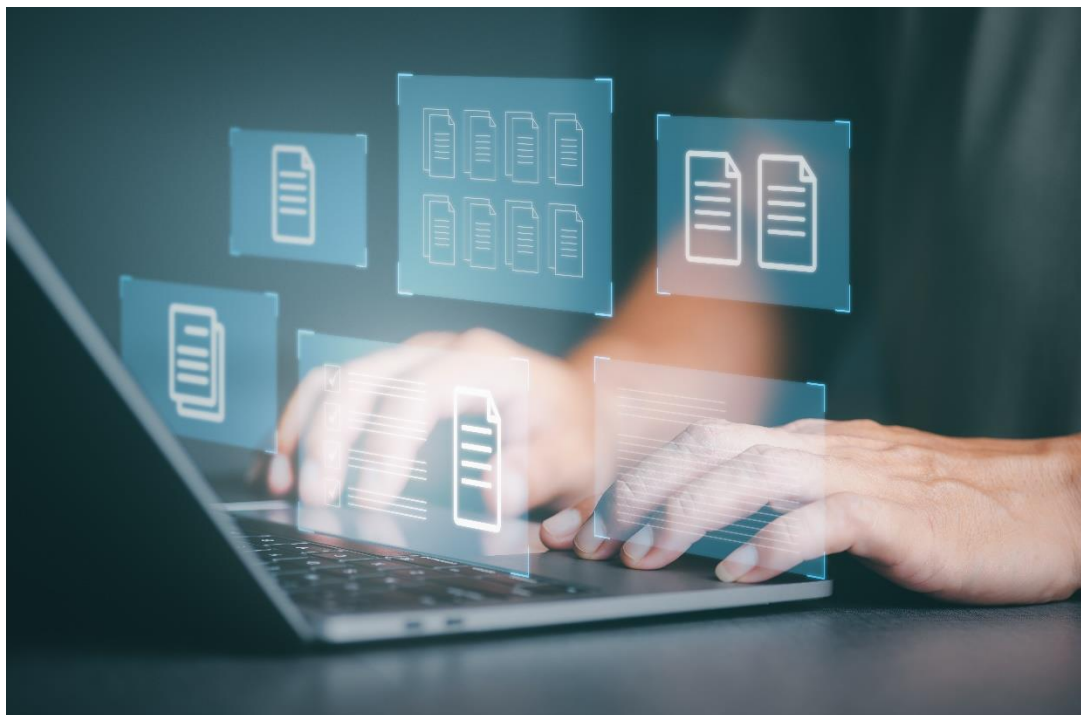


Para ello dispones de una base de datos con información asociada a:

- Información del cliente.
- Fecha de reparto.
- Estado del reparto.
- Productos.
- Categoría de productos.
- Centro de carga.
- Información de las empresas de transporte.
- Distancia recorrida en el reparto.
- Etc.

En total disponemos de casi 6 millones de registros, por lo tanto, necesitamos seleccionar la información sobre la cual se desea realizar el estudio, los criterios que nos ha indicado nuestro cliente se corresponden con:

- Empresa, nos indica la empresa de transportes qué se encarga de realiza el reparto en el domicilio del cliente.
- Código postal del cliente.



El momento de la verdad: como sacamos partido a los datos

Ahora te estarás preguntando cómo podemos hacerlo, estos son los pasos:

1.- **Limpieza de los datos**, lo primero de todo es extraer los datos que tenemos y hay que hacer una limpieza de los mismos, este proceso involucra detectar, eliminar, corregir o transformar cualquier anomalía, perturbación o irrelevancia de los datos.



El objetivo de la limpieza de datos es validar que estos son correctos y se alinean con los objetivos del negocio, si los datos desde el inicio están mal los siguientes pasos se verán afectados por la calidad de la información que tenemos y nos llevará a tener resultado que no estén alineados con lo que estamos buscando.

2.- **Obtener la media móvil**, esta técnica se suele utilizar para suavizar los datos de una serie de tiempo, de esta forma conseguimos reducir el ruido en los datos e identificar más fácilmente patrones y tendencias. En nuestro ejemplo, hemos obtenido la media móvil de los últimos 7 días.

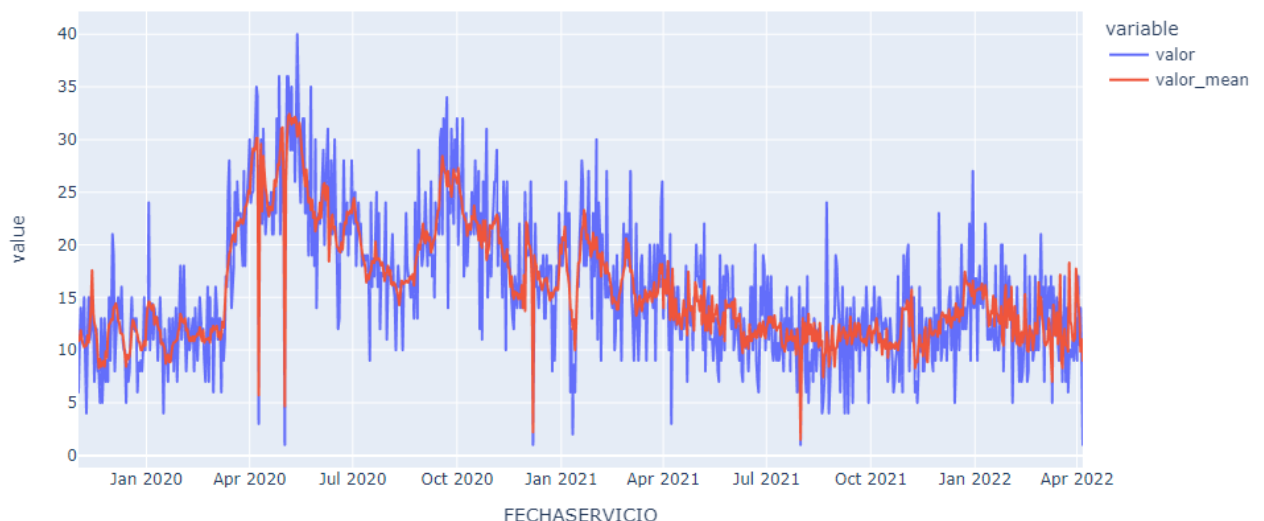
3.- **Exploración previa**, este paso nos va a permitir estudiar el significado y comportamiento de las variables sobre las cuales vamos a realizar el estudio.

| | valor | FECHASERVICIO | empresa | cpostal | valor_mean |
|---|-------|---------------|----------------|---------|------------|
| 0 | 11 | 2020-01-02 | 10010597797876 | 28025 | 14.000000 |
| 1 | 24 | 2020-01-03 | 10010597797876 | 28025 | 14.600000 |
| 2 | 10 | 2020-01-04 | 10010597797876 | 28025 | 14.500000 |
| 3 | 11 | 2020-01-07 | 10010597797876 | 28025 | 14.285714 |
| 4 | 17 | 2022-04-02 | 10010597797876 | 28025 | 14.571429 |

La información del estudio se corresponde con:

- Fecha servicio, nos indica la fecha en la cual se ha realizado el reparto.
- Valor, nos indica el número de repartos realizados.
- Empresa, nos indica la empresa de transportes encargada de realizar el reparto.
- Cpostal, nos indica el código postal del cliente.
- Valor_mean, nos indica la media móvil que hemos calculado a 7 días asociada al número de repartos a realizar.

Grafico de notas por Valor y Valor Medio



4.- Aplicar los algoritmos de series temporales, existen infinidad de algoritmos para poder realizar un análisis de series temporales, para este proyecto se han utilizado todos estos ya que nos ha permitido hacer un estudio bastante completo, son los siguientes:

- Ridge.
- Lasso.
- DecisionTreeRegressor.
- AdaBoostRegressor.
- GradientBoostingRegressor.
- RandomForestRegressor.
- ExtraTreesRegressor.
- XGBRegressor.
- LGBMRegressor.
- CatBoostRegressor.

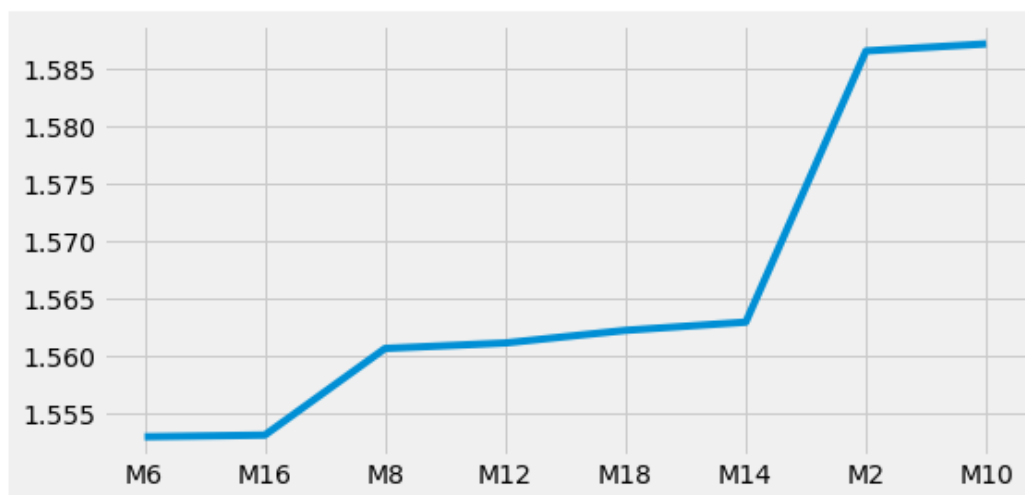
Además, hemos realizado otros cálculos adicionales sobre los algoritmos, estos se corresponden con:

- Hiperparámetros, son parámetros que se pueden ajustar en la creación de un modelo, los cuales deben de ser configurados correctamente para obtener un rendimiento óptimo, es una tarea crucial para aumentar el rendimiento del algoritmo seleccionado, en nuestro caso se han incluido:
 - Número de predicciones a realizar, al ser predicciones semanales hemos optado por 7, 14, 21 y 28 días.
 - Número de veces que se diferencian los datos, el cual nos indica el número de veces que a cada valor del número de predicciones se le resta el valor anterior, para ello hemos optado por los valores 0, 1, 2, 3, 4 y 5.
 - Así como otros parámetros adicionales que requiere el algoritmo a ejecutar.
- Además, se han incluido variables adicionales, éstas se denominan variables exógenas y hacen referencia a algo que se forma en el exterior y no depende de la evolución del modelo, en nuestro ejemplo hemos utilizado:
 - El día del mes, para poder estudiar el comportamiento del modelo al incluir el día del mes.
 - El día de la semana, para poder estudiar el comportamiento del modelo al incluir el día de la semana.
 - El día del mes y el día de la semana, para estudiar el comportamiento al incluir las 2 variables de forma conjunta.

5.- Resultados finales, se han ejecutado 18 modelos diferentes y estos son los resultados con el número de repartos a realizar:

| | Real | M1 | M2 | M5 | M6 | M7 | M8 | M9 | M10 | M11 | M12 | M13 | M14 | M15 |
|------------|------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| 2021-10-12 | 12.6 | 13.165376 | 10.545322 | 12.579713 | 10.379757 | 11.970500 | 10.518868 | 12.992500 | 10.563543 | 12.177167 | 13.427340 | 12.471831 | 10.365910 | 13.141263 |
| 2021-10-13 | 12.0 | 12.284349 | 10.821424 | 12.148674 | 10.587788 | 10.488333 | 10.785046 | 9.584167 | 10.838624 | 9.926667 | 12.689302 | 12.024891 | 10.564352 | 12.564368 |
| 2021-10-14 | 13.0 | 12.695282 | 10.860521 | 12.027882 | 10.588700 | 11.126667 | 10.821565 | 10.860333 | 10.871943 | 10.943000 | 12.904596 | 11.885826 | 10.547189 | 12.070349 |
| 2021-10-15 | 12.0 | 13.053902 | 10.901622 | 12.623741 | 10.632612 | 12.078000 | 10.862185 | 12.399667 | 10.904902 | 11.774000 | 12.999785 | 12.457259 | 10.568076 | 12.375207 |
| 2021-10-16 | 7.0 | 12.913507 | 11.035706 | 12.577981 | 10.705937 | 12.403333 | 10.999620 | 11.973667 | 11.024682 | 12.136000 | 12.536724 | 12.382306 | 10.616566 | 11.993719 |

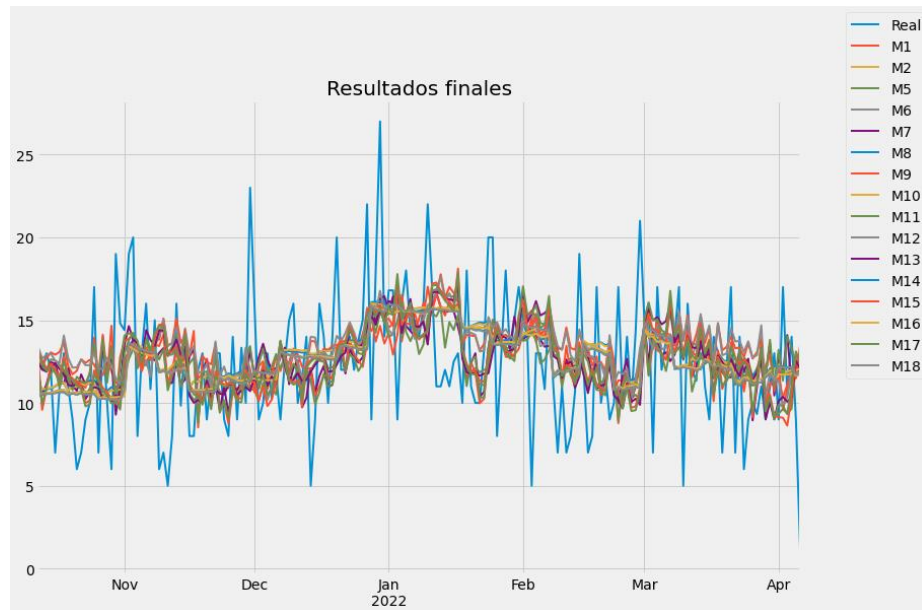
El mejor algoritmo ejecutado nos indica un error medio (MAE) de 1,552931, es decir nos podemos equivocar en 1,55 repartos hacia arriba o hacia abajo.



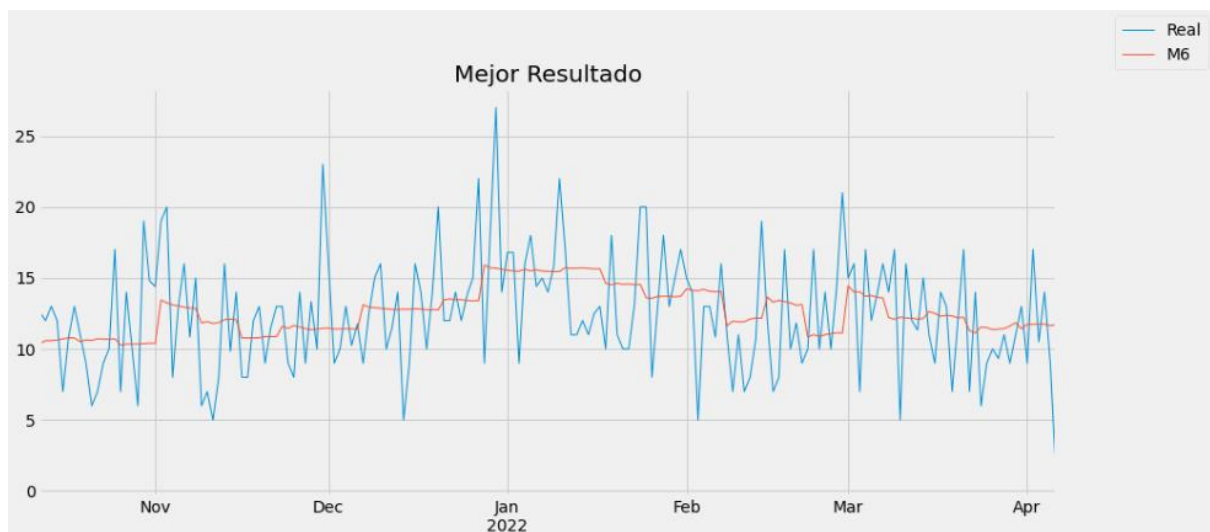
Además, hay que tener en cuenta el modelo M16 que da un error casi similar al modelo M6.

El modelo M16 tiene en cuenta la variable exógena **día de la semana**, comentándolo con el cliente nos indica que los lunes y los viernes suelen ser los días de más repartos, así como la primera semana del mes y la última semana del mes, siempre que no existan días de fiestas entre semana donde las predicciones no son normales, por lo tanto, en una siguiente versión hay que incluir cómo variables exógenas el número de la semana del mes y los días festivos.

Este es el gráfico de todos los modelos ejecutados.



Este es el gráfico del mejor resultado, **modelo M6** comparado con el valor real.



En resumen, el resultado nos indica que el número de repartos diarios para la empresa de transportes y los códigos postales seleccionados al inicio, está entre 11,7 aproximadamente, con un error de 1,5, aproximadamente un 12% de error.

| | Real | M1 | M2 | M3 | M6 | M7 | M8 | M9 | M10 | M11 | M12 | M13 | M14 | M15 |
|------------|------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| 2022-04-02 | 17.0 | 10.329273 | 11.778147 | 9.632718 | 11.734258 | 11.782667 | 11.765361 | 9.578333 | 11.765129 | 11.348333 | 9.864890 | 10.369153 | 12.177086 | 9.102435 |
| 2022-04-03 | 10.5 | 11.101923 | 11.918598 | 9.366301 | 11.749624 | 14.097500 | 11.870791 | 10.826667 | 11.892793 | 14.006333 | 10.298840 | 10.072097 | 12.163880 | 8.633288 |
| 2022-04-04 | 14.0 | 11.828161 | 11.977090 | 9.646106 | 11.743215 | 11.067833 | 11.884814 | 12.188333 | 11.987597 | 13.344000 | 11.914255 | 10.326644 | 12.135572 | 10.415464 |
| 2022-04-05 | 9.0 | 12.357288 | 11.951201 | 12.372940 | 11.642685 | 11.720500 | 11.889308 | 11.486333 | 11.970932 | 12.261667 | 12.529177 | 12.572450 | 11.693193 | 12.934593 |
| 2022-04-06 | 1.0 | 12.610134 | 12.046284 | 11.925123 | 11.714048 | 13.065500 | 11.950797 | 13.552500 | 12.065529 | 12.570000 | 12.656059 | 12.069085 | 11.803754 | 12.288661 |

Resultado con la predicción

👍 En **Go Insurance Software** sabemos que las herramientas están y las posibilidades también. Ahora depende de que bien las organizaciones o sus directivos decidan subirse a esta ola. ¿Te subes con nosotros?



¿Está tu organización lista para descubrir el poder de los datos y de las últimas tecnologías?

► Si quieres saber cómo desde mi empresa te podemos ayudar, concierta conmigo una reunión haciendo clic en este enlace:

<https://calendly.com/ajusticialopez/15min>