Reporte de selección y parametrización de modelos:

Grupo 7 – Claudia Baquero, Daniel Hoyos y Johan Medina

El presente informe detalla el proceso de selección y parametrización de modelos para la detección de anomalías en series de tiempo de consumo de energía activa por parte de clientes no regulados de Electro Dunas; para lo cual se incorporan diferentes pasos para la selección de mejor modelo como se detalla a continuación:

1. Recopilación y Exploración de Datos:

Se trabaja con los datos históricos detallados sobre el consumo de energía activa por parte de clientes no regulados; los cuales fueron explorados para comprender las tendencias históricas, identificar patrones estacionales y detectar posibles anomalías previas.

Tabla 1 Distribución de datos por cliente

Cliente	Fecha_min	Fecha_max	Active_energy_min	Active_energy_max	Datos
1	2021-01-01	01-04-2023	0.001715	6.526612	19681
2	2021-01-01	01-04-2023	0.027246	8.283679	19681
3	2021-01-01	01-04-2023	0.001658	8.563269	19681
4	2021-01-01	01-04-2023	0.000240	7.204313	19681
5	2021-01-01	01-04-2023	0.004968	14.622644	19681
6	2021-01-01	01-04-2023	0.000092	6.365392	19681
7	2021-01-01	01-04-2023	0.000037	8.574905	19681
8	2021-01-01	01-04-2023	0.000250	7.322165	19681
9	2021-01-01	01-04-2023	0.000014	2.546763	19681
10	2021-01-01	01-04-2023	0.011014	8.343834	19681
11	2021-01-01	07-08-2022	0.000620	4.352005	14000
12	2021-01-01	21-04-2022	0.000000	3.591291	11415
13	2021-01-01	07-08-2022	0.001417	5.157014	14000
14	2021-01-01	07-08-2022	0.000011	0.781830	14000
15	2021-01-01	07-08-2022	0.000106	2.810176	14000
16	2021-01-01	24-03-2023	0.000052	6.266824	19500
17	2021-01-01	24-03-2023	0.000000	5.827472	19500
18	2021-01-01	24-03-2023	0.000000	9.943944	19500
19	2021-01-01	24-03-2023	0.127157	7.534833	19500
20	2021-01-01	24-03-2023	0.000000	5.503240	19500
21	2021-01-01	07-01-2022	0.000000	0.358380	8925
22	2021-01-01	07-01-2022	0.000000	0.542551	8925
23	2021-01-01	07-01-2022	0.000000	0.814156	8925
24	2021-01-01	07-01-2022	0.000000	0.445964	8925
25	2021-01-01	07-01-2022	0.000000	0.489633	8925
26	2021-01-01	21-04-2022	0.000000	3.237979	11415
27	2021-01-01	21-04-2022	0.000000	4.328475	11415
28	2021-01-01	21-04-2022	0.000000	3.970402	11415
29	2021-01-01	21-04-2022	0.000000	2.731480	11415
30	2021-01-01	21-04-2022	0.000000	3.591291	11415

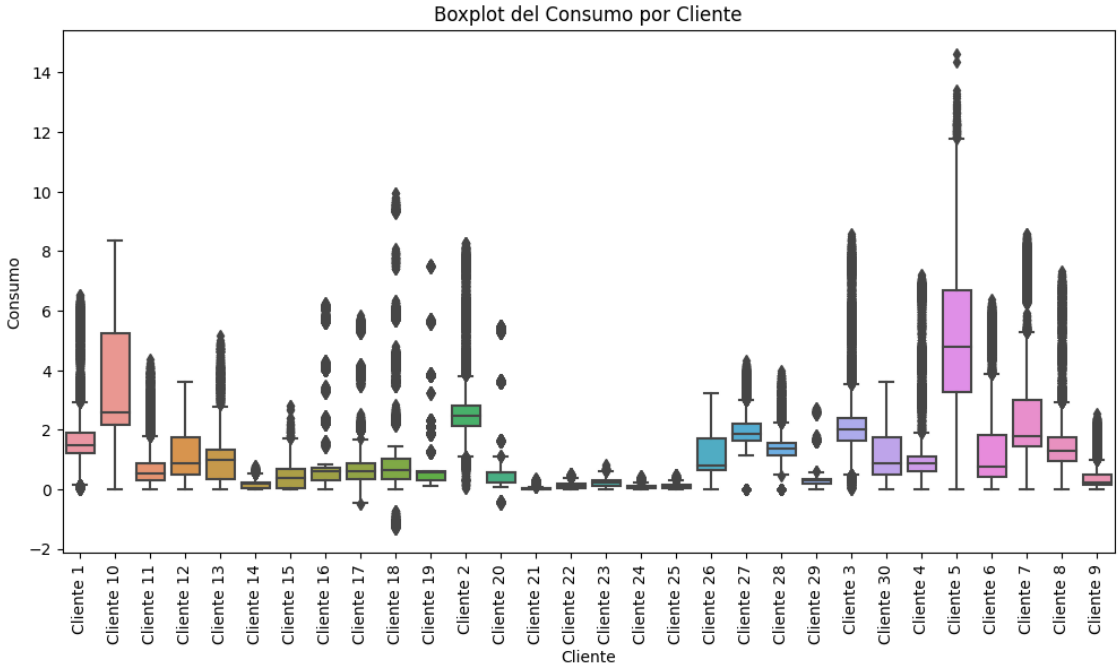
Los datos muestran un período de tiempo que va desde el 1 de enero de 2021 hasta el 1 de abril de 2023, lo que sugiere una recopilación de datos a lo largo de más de dos años para la mayoría de los clientes; durante este tiempo, se observa una variación en la energía activa mínima y máxima consumida por los clientes, con valores mínimos que oscilan entre 0.00 y 0.127157, y valores máximos que van desde 2.54hasta 14.62.

La mayoría de los clientes parecen estar distribuidos uniformemente a lo largo de este período, con una cuenta constante de 19,681 filas por cliente; sin embargo, hay algunas excepciones, como los clientes 11, 12, 13, 14 y 15, que tienen un recuento de filas de 14,000 en lugar de 19,681, y los clientes 21 al 30, que tienen un recuento de filas de 8,925 en lugar de 19,681.

Lo anterior se debe a que hay una variabilidad en las fechas mínimas y máximas de los datos para algunos clientes; por ejemplo, los clientes 11, 13, 14 y 15 tienen una fecha máxima de 7 de agosto de 2022, mientras que los clientes 16 al 20 tienen una fecha máxima de 24 de marzo de 2023 lo que podría indicar diferentes patrones de consumo de energía o diferentes períodos de tiempo de observación para estos grupos de clientes.

En resumen, los datos proporcionan una visión detallada del consumo de energía activa de varios clientes a lo largo de un período de tiempo significativo, con algunas variaciones en los patrones de consumo y la disponibilidad de datos para ciertos grupos de clientes; por lo que se realiza una exploración de las variaciones de estos mediante el siguiente box plot:

Tabla 2 Energía Activa por cliente



Se observa una amplia gama de consumos de energía activa entre los diferentes clientes; por ejemplo, el cliente 5 tiene un consumo promedio de aproximadamente 4.99, mientras que el cliente 21 tiene un consumo promedio significativamente más bajo de aproximadamente 0.03; lo que refleja una gran variabilidad en los patrones de consumo entre los clientes.

La desviación estándar también varía considerablemente entre los clientes, lo que indica la dispersión de los datos en torno a la media; por ejemplo el cliente 5 tiene una desviación estándar de aproximadamente 2.44, lo que indica una dispersión considerable en los datos, mientras que el cliente 21 tiene una desviación estándar mucho más baja de aproximadamente 0.04, lo que sugiere una menor variabilidad en los patrones de consumo; pero que también es explicado por los datos históricos presentes.

Al observar los percentiles 75% y max, podemos identificar los clientes que tienen un consumo de energía activa más alto, en este caso es el cliente 5 tiene con un máximo de 14.62, lo que indica que este cliente tiene un consumo de energía activa notablemente más alto en comparación con otros clientes.

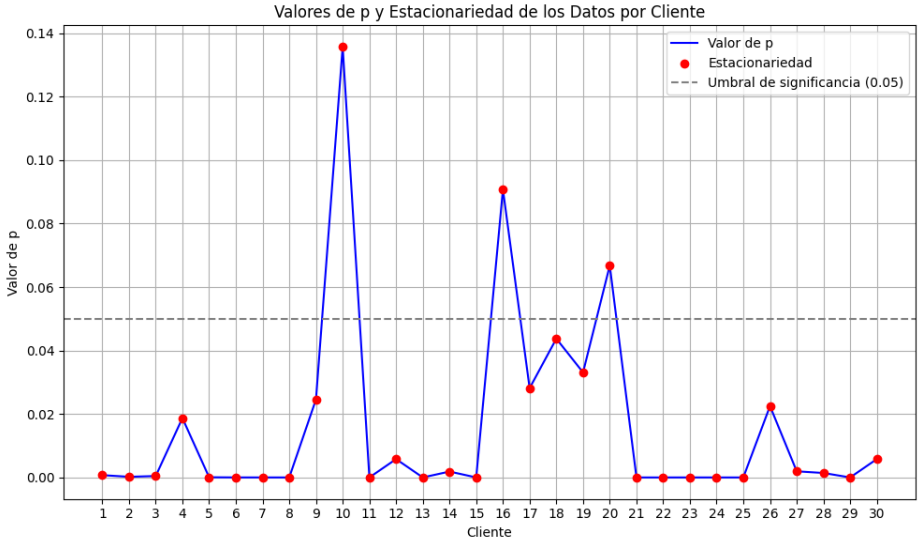
Del mismo modo, los percentiles 25% y min nos permiten identificar los clientes con un consumo de energía activa más bajo; el cual corresponde al cliente 21 que tiene un consumo mínimo de aproximadamente 0.00.

Dado lo anterior y debido a la falta de etiquetas que nos indiquen anomalías históricas, la exploración de datos nos sugiere que lo ideal es ajustar un modelo no supervisado, en este caso de pronóstico para cada cliente de manera individual dadas sus notorias diferencias tanto en periodos de tiempo registrados como en patrones de consumo de energía activa. Estos modelos de pronósticos nos podrán servir para que, dado un umbral de confianza, podamos determinar posibles consumos anómalos que es el objetivo final de esta iniciativa.

2. Evaluación de Estacionariedad

Dado que el objetivo principal del modelo es la detección de anomalías por medio de pronósticos, se realizaron pruebas de estacionariedad para garantizar que las series de tiempo de cada cliente sean estacionarias.

Gráfico 1 Estacionariedad por cliente según energía activa



Los resultados de las pruebas de estacionariedad (ADF) indican que la mayoría de los clientes exhiben datos estacionarios en sus series temporales de energía activa. Específicamente, de los 30 clientes analizados, 27 muestran datos estacionarios, como se evidencia por los valores de p significativamente bajos (menores que 0.05) y las estadísticas ADF negativas. Esto sugiere que las series temporales de energía activa para la mayoría de los clientes no regulados se pueden considerar estacionarias, lo que es fundamental para modelar y predecir anomalías.

Los clientes (10, 16 y 20) muestran resultados de prueba que indican la no estacionariedad de sus datos. En estos casos, se observan valores de p más altos (mayores que 0.05) junto con estadísticas ADF menos negativas o incluso positivas. Este hallazgo sugiere que estos clientes pueden requerir un enfoque diferente en el modelado y la detección de anomalías debido a la naturaleza no estacionaria de sus series temporales de energía activa.

Tabla 3 Conclusión supuesto estacionariedad cliente

Cliente	Estadística ADF	Valor p	Conclusión
1	-4,174790014	0,000727	Estacionarios
2	-4,491405477	0,000204	Estacionarios
3	-4,301114208	0,000442	Estacionarios
4	-3,222778636	0,018703	Estacionarios
5	-4,787967122	5,73E-05	Estacionarios
6	-5,212424738	8,29E-06	Estacionarios
7	-5,275952647	6,14E-06	Estacionarios
8	-6,830970661	1,89E-09	Estacionarios
9	-3,127750712	0,024567	Estacionarios
10	-2,421804532	0,135677	No estacionarios
11	-8,956256027	8,46E-15	Estacionarios
12	-3,598753015	0,005785	Estacionarios
13	-8,31282128	3,75E-13	Estacionarios
14	-3,925987828	0,001852	Estacionarios
15	-5,368412922	3,94E-06	Estacionarios
16	-2,611001049	0,090745	No estacionarios
17	-3,079091819	0,02814	Estacionarios
18	-2,914094771	0,043733	Estacionarios
19	-3,019758197	0,03309	Estacionarios
20	-2,743794673	0,066762	No estacionarios
21	-11,37756491	8,67E-21	Estacionarios
22	-6,787238266	2,41E-09	Estacionarios
23	-5,277779306	6,09E-06	Estacionarios
24	-8,395426809	2,31E-13	Estacionarios
25	-8,024335897	2,04E-12	Estacionarios
26	-3,158978351	0,022485	Estacionarios
27	-3,914660986	0,00193	Estacionarios
28	-4,00099343	0,001405	Estacionarios
29	-6,829599263	1,91E-09	Estacionarios
30	-3,598753015	0,005785	Estacionarios

Teniendo en cuenta lo anterior, se realizará la selección y parametrización de modelos como Random Forest, XGBoost y LSTM principalmente por su flexibilidad y su capacidad para manejar relaciones no lineales y complejas en los datos (data-based); caso contrario a modelos tradicionales como el ARIMA (model-based) en el que se asume que la serie temporal es estacionaria y lineal, lo que puede limitar su capacidad para capturar patrones complejos o no lineales presentes en los datos; esto considerando que se ha evidenciado que tres clientes no cumplen con estos supuestos. Esta decisión también se sustenta en la premisa de que nos importa más una buena precisión en la detección de anomalías que en la interpretabilidad del modelo utilizado.
Con esto se procede a realizar la selección y prueba de los modelos de aprendizaje automático referenciados anteriormente, el proceso implementado se relaciona en las siguientes líneas; adicionalmente por su capacidad para trabajar con datos de múltiples características se podrán implementar con mayor facilidad.

3. Partición de datos y medida de desempeño:

La partición de datos se llevó a cabo siguiendo una estrategia de 70-20-5-5, donde el 70% de los datos se asignaron para entrenar los modelos, el 20% se reservó para realizar pruebas y evaluar el rendimiento de los modelos, y un 5% adicional se destinó específicamente para realizar pronósticos iniciales que nos permitieran definir umbrales de confianza a partir de los cuales se marcan las anomalías de los clientes según el mejor modelo definido.

Esta última porción del 5% se utiliza para detectar y validar posibles anomalías en los datos de los clientes según el modelo escogido y los umbrales definidos en la sección anterior;
al dedicar un porcentaje específico de los datos para esta validación, se pudo identificar y abordar de manera más efectiva cualquier desviación o comportamiento inesperado en las series temporales de energía activa de los clientes.

Con esta estrategia de partición de datos se logra entrenar modelos robustos, realizar pruebas exhaustivas y validar la precisión de los pronósticos, garantizando así una evaluación completa y confiable del desempeño de los modelos de predicción.

Con respecto a la medida de desempeño utilizada en la evaluación de modelos de pronósticos y considerando la pregunta de negocio de “poder generar alertas ante anomalías en el consumo de energía de clientes no regulados” decidimos utilizar el error absoluto porcentual medio MAPE que nos permite evaluar que tan bien se ajusta el pronóstico a la historia, definir umbrales porcentuales considerando el error porcentual y además, nos permite comparar entre clientes con diferentes escalas o proporciones de consumo de energía ya que se encuentra en términos porcentuales.

4. Selección de Modelos y Calibración:

Se inicia con la evaluación de un único cliente con el fin de establecer las variables que deben tenerse en cuenta para los diferentes modelos a implementar; se trabaja entonces

con el Cliente 5; que presenta una serie completa desde la fecha mínima y la fecha máxima de los datos; y variabilidad alta en la energía activa.

Tabla 4 Resultado Iteración Random Forest Cliente 5

	Frecuencia	Rezago	MSE	MAE	RMSE	MAPE	R2	Len_train	Len_test	Len_total
0	12H	15	0.144926	0.215042	0.380691	10.198100	-0.454763	524	131	655
1	12H	30	0.091259	0.187130	0.302091	8.072581	0.024116	500	125	625
2	1H	15	0.002083	0.023493	0.045635	19.484270	-0.022603	6413	1604	8017
3	1H	30	0.001661	0.020705	0.040752	12.249518	0.116328	6389	1598	7987
4	8H	15	0.086101	0.144589	0.293431	10.261707	0.036536	791	198	989
5	8H	30	0.040104	0.122307	0.200261	9.865950	0.010210	767	192	959

Teniendo en cuenta los resultados obtenidos en el ejercicio anterior se determinaron las variables requeridas para las diferentes iteraciones y la búsqueda del mejor modelo para cada cliente según los datos históricos de energía entregados por ElectroDunas.

Luego se procedió a iterar en la información de energía sobre cada cliente, donde se aplicaron una serie de pasos que se detallan a continuación; se ordenaron los datos por fecha de manera ascendente y se seleccionó el 90% de los datos disponibles.

Posteriormente se realizaron iteraciones sobre diferentes frecuencias y rezagos para el remuestreo de los datos y la creación de columnas de rezagos; se dividieron los datos en conjuntos de entrenamiento y prueba, y se aplicaron los tres modelos de aprendizaje automático definidos *(Random Forest, XGBoost y LSTM)*.

Adicionalmente se exploraron varios hiperparámetros para cada modelo XGB y RF mediante una búsqueda en rejilla (GridSearch) que consideró todas las combinaciones posibles; para la red neuronal, se implementaron generadores de series temporales para el conjunto de entrenamiento y prueba y, además, se definieron múltiples combinaciones de hiperparámetros, como la cantidad de neuronas en la capa LSTM, la función de activación y el número de épocas de entrenamiento.

Basados en todas las combinaciones posibles de estos hiperparámetros, se iteró sobre cada una, entrenando y evaluando el modelo correspondiente utilizando métricas de rendimiento, como el error absoluto porcentual medio (MAPE), para evaluar el rendimiento de cada modelo en el conjunto de prueba; con lo anterior se generaron más de 6.200 modelos por Cliente teniendo en cuenta las iteraciones del cuadro resumen siguiente y las consideraciones señaladas posteriormente:

Tabla 5 Combinaciones Probadas para Modelos de Clientes

Modelo	Frecuencia	Rezagos	Parámetros
Random Forest	1H, 8H, 12H	10, 15, 20, 25, 30	n_estimators_values = [100, 300, 500]
XGBoost			max_depth_values = [10, 20, 30, None]

Modelo	Frecuencia	Rezagos	Parámetros
			min_samples_split_values = [2, 5, 10] min_samples_leaf_values = [1, 2, 4] bootstrap_values = [True, False]
LSTM Inicial	1H, 8H, 12H	10, 30	neuronas = [100] f_activacion = ['tanh','relu'] num_epochs = [25]
LSTM 2nda versión	8H, 12H, 24H, 48H	15,30	neuronas = [100] f_activacion = ['tanh'] num_epochs = [25]

Cómo se observa en la tabla, se evaluaron los modelos mencionados con diferentes frecuencias (formas de agrupar el consumo de energía activa) incluyendo los datos iniciales entregados por hora, diferentes rezagos e hiperparámetros asociados a cada algoritmo.

Además, se tuvo tratamientos adicionales durante el entrenamiento de los modelos cómo:

- **Validación inicial:** Los modelos fueron entrenados utilizando únicamente los datos históricos disponibles, sin aplicar consideraciones adicionales. Este enfoque generó desafíos en la estimación del Error Porcentual Absoluto Medio (MAPE) para algunos clientes, debido a posibles errores en los datos o características específicas de ciertos clientes.
- **Imputación de valores cero**: Para abordar la presencia de valores de energía cero en los datos, se llevó a cabo un proceso de imputación donde estos valores fueron reemplazados por un número muy cercano a cero (0.001); esta estrategia permitió calcular el MAPE y facilitó la comparación entre modelos. Se tomó esta decisión basándose en el hecho de que los valores con energía cero representan menos del 1% del total de datos tanto a nivel general como en cada cliente individualmente.
- **Inclusión de la energía reactiva:** En algunas iteraciones, ciertos clientes no lograron alcanzar un rendimiento óptimo incluso después de la imputación de valores cero; en respuesta a esto, se decidió incorporar la energía reactiva como una variable adicional en el modelo; generando así una mejora significativa en el rendimiento de algunos modelos, lo que sugiere que la energía reactiva contiene información relevante para la predicción de la serie temporal.
- **Adaptación para clientes específicos:** Durante la exploración de modelos anteriores, se identificó que algunos clientes, específicamente los números 6, 7, 8, 11, 13 y 15, enfrentaban dificultades para lograr un MAPE relevante; teniendo en cuenta valores hasta de 300 en esta medida de desempeño; ante esta situación, se implementó una nueva arquitectura de red neuronal con una frecuencia de agrupamiento mayor en un intento por mejorar la precisión de la predicción; además, se optó por utilizar exclusivamente la función de activación “Tanh” con el fin de aprovechar el procesamiento paralelo ofrecido por una unidad de procesamiento gráfico (GPU), lo que permitió acelerar el proceso de modelado y exploración de hiperparámetros.

El top 3 de los mejores modelos encontrados para cada cliente se encuentran al final de este documento en la *sección de anexos* debido a la gran cantidad de modelos (90; 3 por cliente) e incluyen los modelos ganadores y las variables empleadas para el entrenamiento y las medidas de desempeño obtenidas en cada caso.

En la tabla siguiente, se muestra el mejor modelo para cada uno de los treinta clientes; en consideración al valor mínimo de error (MAPE) obtenido tras la iteración y experimentación descrita anteriormente; adicionalmente se presentan las características utilizadas en su entrenamiento (rezagos, frecuencia, imputación, inclusión de energía reactiva, y otros a considerar.) para su posterior replicación:

Tabla 6 Mejor modelo por cliente y características.

Cliente	Frecuencia	Rezago	Modelo	Hiperparametros	Imputado	Incluye_reactiva	MSE	MAE	RMSE	MAPE	R2
1	12H	15	XGB	(100, 5, 0.1, 0)	NO	SI	2,71	0,88	1,65	5,96	0,99
2	12H	30	XGB	(100, 5, 0.1, 0.1)	SI	SI	1,22	0,87	1,11	2,91	1,00
3	12H	15	XGB	(500, 3, 0.01, 0.1)	SI	SI	2,84	1,19	1,68	4,56	0,99
4	24H	15	LSTM	(100, 'tanh', 25)	NO	NO	3,46	0,94	1,86	6,23	0,88
5	24H	15	LSTM	(100, 'tanh', 25)	NO	NO	145,56	9,44	12,06	10,04	-0,04
6	24H	30	LSTM	(100, 'tanh', 25)	NO	NO	2,19	1,22	1,48	9,60	0,76
7	48H	30	LSTM	(100, 'tanh', 25)	NO	NO	1181,26	13,83	34,37	10,32	-0,05
8	48H	30	LSTM	(100, 'tanh', 25)	NO	NO	33,92	4,51	5,82	8,38	-0,41
9	12H	30	XGB	(100, 5, 0.1, 0)	NO	SI	0,59	0,30	0,77	6,75	0,99
10	48H	15	LSTM	(100, 'tanh', 25)	NO	NO	22,70	3,74	4,76	3,54	-0,02
11	24H	30	LSTM	(100, 'tanh', 25)	NO	NO	141,07	8,16	11,88	27,95	0,13
12	12H	15	XGB	(100, 5, 0.05, 0)	SI	SI	3,09	1,13	1,76	7,73	0,93
13	24H	15	LSTM	(100, 'tanh', 25)	NO	NO	3,08	1,37	1,75	4,94	-0,05
14	24H	15	LSTM	(100, 'tanh', 25)	NO	NO	0,47	0,54	0,69	9,49	-0,60
15	24H	15	LSTM	(100, 'tanh', 25)	NO	NO	2,68	1,37	1,64	9,84	-0,05
16	48H	15	LSTM	(100, 'tanh', 25)	NO	NO	1,23	0,57	1,11	2,74	-0,04
17	48H	15	LSTM	(100, 'tanh', 25)	NO	NO	1,44	0,74	1,20	3,18	0,05
18	48H	15	LSTM	(100, 'tanh', 25)	NO	NO	2,48	1,01	1,58	3,96	0,03
19	48H	30	LSTM	(100, 'tanh', 25)	NO	NO	1,39	0,30	1,18	1,82	-0,04
20	12H	30	XGB	(300, 7, 0.05, 0.1)	SI	SI	17,48	1,81	4,18	9,99	0,91
21	12H	15	XGB	(300, 5, 0.01, 0)	SI	SI	0,05	0,09	0,22	18,07	0,35
22	12H	15	XGB	(500, 10, 0.01, 0.1)	SI	SI	0,11	0,20	0,33	11,62	0,38
23	12H	30	XGB	(300, 3, 0.01, 0.8, 0.8, 0.1)	SI	NO	0,28	0,34	0,53	11,53	0,04
24	12H	15	XGB	(100, 3, 0.05, 0)	SI	SI	0,07	0,14	0,27	13,73	0,52
25	12H	30	XGB	(100, 3, 0.05, 0)	SI	SI	0,07	0,17	0,26	11,94	0,39
26	12H	15	XGB	(100, 5, 0.1, 0.1)	SI	SI	1,11	0,55	1,05	4,26	0,97
27	12H	15	XGB	(500, 5, 0.01, 0.2)	SI	SI	0,98	0,62	0,99	2,53	0,98
28	12H	30	XGB	(100, 3, 0.05, 0.1)	SI	SI	0,63	0,60	0,79	3,45	0,98
29	12H	30	XGB	(300, 3, 0.01, 0)	SI	SI	1,39	0,61	1,18	13,17	0,46
30	12H	15	XGB	(100, 5, 0.05, 0)	SI	SI	3,09	1,13	1,76	7,73	0,93

En general, hemos encontrado que para el 46.6% de los 30 clientes de Electro Dunas, el mejor modelo de predicción es XGBoost, mientras que para el resto se recomienda utilizar LSTM. Para ambos tipos de modelos, se debe tener en cuenta una variedad de factores, incluidos diferentes hiperparámetros, como la cantidad de rezagos y variables predictoras, así como diversos preprocesamientos, como la frecuencia de agrupación y la imputación de valores en cero según se especifica en la tabla adjunta.

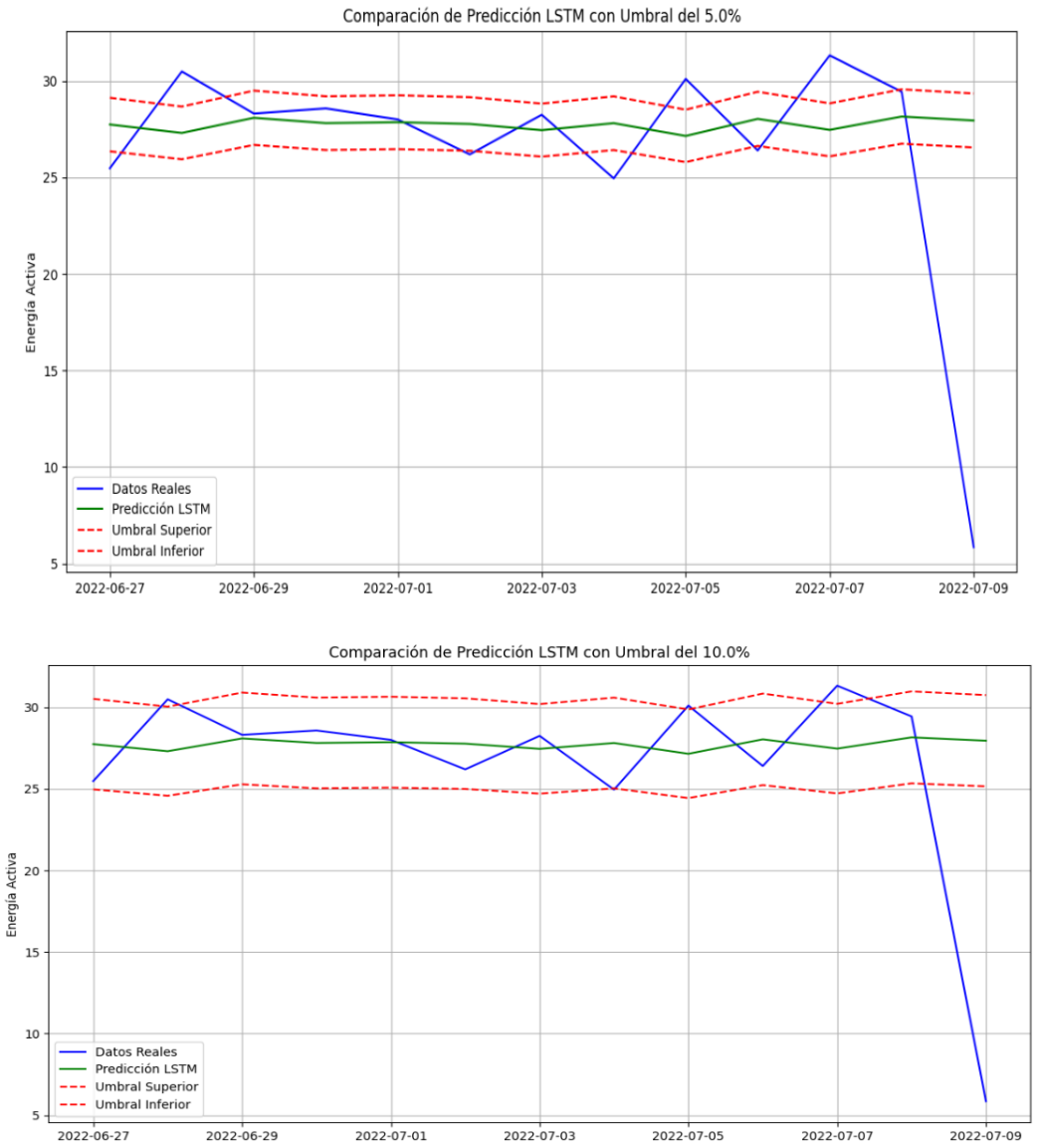
Adicionalmente, encontramos que 21 de los clientes tienen un MAPE menor al 10% y 28 de los 30 clientes tienen un MAPE inferior al 15%; con lo cual el promedio de los modelos a implementar tiene un MAPE de 8,27% que cumple con el requisito de desempeño especificado para el prototipo (MAPE < 10%).

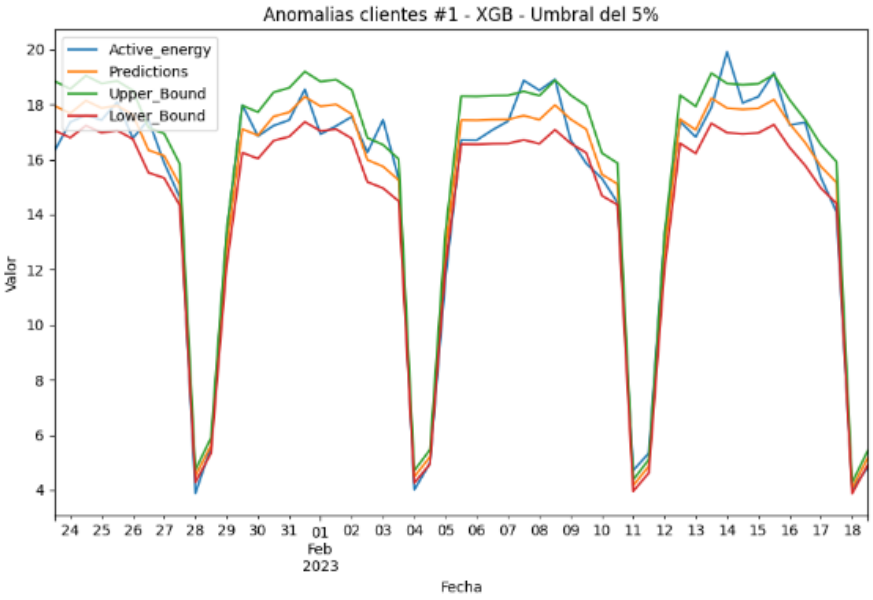
5. Evaluación y selección del rango del umbral para anomalías:

Utilizando el modelo de pronóstico entrenado para cada cliente según los mejores parámetros y variables determinadas por el menor MAPE se procede a utilizar un 5% de los datos para pronosticar y obtener valores futuros con precisión.

Con la predicción de la energía activa; se itera sobre algunos multiplicadores (umbrales) para establecer los límites de anomalías y así determinar la sensibilidad de la detección de estas.

Gráfico 2 Pruebas de umbrales diferentes modelos





Se utiliza una función para de manera iterativa establecer diferentes umbrales con algunos clientes y partir de esto se define el límite superior e inferior de anomalías y definir la detección de consumos atípicos según el umbral establecido.

Tabla 7 Umbrales y cantidad de anomalías para algunos clientes

LSTM

cliente	umbral	anomalia	no_anomalia
Cliente 13	0.05	5	8
Cliente 13	0.10	5	8
Cliente 13	0.15	1	12
Cliente 13	0.20	1	12
Cliente 14	0.05	2	11
Cliente 14	0.10	6	7
Cliente 14	0.15	4	9
Cliente 14	0.20	2	11
Cliente 16	0.05	4	20
Cliente 16	0.10	1	23
Cliente 16	0.15	0	24
Cliente 16	0.20	0	24

XGBoost

cliente	umbral	anomalia	no_anomalia
Cliente 1	0.20	0	53
Cliente 1	0.15	0	53
Cliente 1	0.10	5	48
Cliente 1	0.05	18	35
Cliente 2	0.20	22	31
Cliente 2	0.15	26	27
Cliente 2	0.10	30	23
Cliente 2	0.05	46	7
Cliente 3	0.20	20	33
Cliente 3	0.15	26	27
Cliente 3	0.10	35	18
Cliente 3	0.05	45	8

Tabla 8 Resumen de anomalías para algunos clientes utilizando diferentes modelos

LSTM

umbral	anomalia	no_anomalia
0.05	11	39
0.10	12	38
0.15	5	45
0.20	3	47

XGBoost

umbral	total_anomalia	total_no_anomalia
0.05	109	50
0.10	70	89
0.15	52	107
0.20	42	117

Basados en los anteriores datos, se establece el umbral óptimo del 20% que permitirá identificar con menor sensibilidad los posibles consumos anómalos de energía activa por cliente; este será desplegado en el tablero final para cada uno de los 30 clientes y a utilización para la definición de las anomalías en los datos de validación a emplear.

6. Siguientes pasos para lograr el prototipo final:

Teniendo entonces ya un modelo adecuado para pronosticar el consumo cada uno de los 30 clientes en conjunto con el umbral para marcar las anomalías definido. Los siguientes pasos para lograr el prototipo final son:

- Crear un notebook o script que sea capaz de: entrenar los modelos definidos para cada cliente tomando los datos de manera directa del repositorio de GitHub, generar un pronóstico del último 5% la información presente para cada cliente y generar un CSV por cliente que contenga las columnas: fecha, energía activa, predicción, límite superior y límite inferior.
- Crear un tablero o dashboard en Dash que podamos desplegar en una máquina EC2 y consultar por medio de una dirección web.
- Este tablero debe contar con: filtros para seleccionar los clientes, filtros para seleccionar sectores, filtros para fecha y hora inicial y final, una gráfica descriptiva del total del consumo de energía activa por clientes del sector filtrado, un grafica del total de voltaje FA y FC por cliente y/o sector filtrado, una gráfica de series de tiempo parecida a la presentada en la sección de anomalías en donde se pueda identificar cuando el pronóstico se aleja de la realidad y aparecen datos anómalos.
- A la gráfica de series de tiempo se le incorporarán los datos generados por el script que contiene fecha, energía activa, predicción, límite superior y límite inferior del cliente que se encuentre filtrado y se marcarán aquellas franjas en las que el consumo real supere los limites previstos. Si no hay un cliente filtrado esta gráfica sale vacía.

- Una vez se tenga desarrollado el tablero, procederemos con el despliegue de este en una máquina EC2 en la nube como se especificó en la primera entrega del prototipo fachada y los requisitos del proyecto.
- Por último, se documentará tanto los archivos y artefactos de la solución como la forma de desplegar y utilizar el prototipo final.

Con lo anterior, esperamos dar solución a la problemática de ElectroDunas alertando de manera visual consumos anómalos de clientes no regulados a los operarios encargados del monitoreo y la gestión de anomalías.

Anexos:

Anexo 1 Top tres mejores modelos por cliente

Cliente	Frecuencia	Rezago	Modelo	Hiperparametros	MSE	MAE	RMSE	MAPE	R2
Cliente 1	12H	15	XGB	(100, 5, 0.1, 0)	2,71	0,88	1,65	5,96	0,99
Cliente 1	12H	15	XGB	(100, 5, 0.05, 0.2)	2,42	0,86	1,56	6,01	0,99
Cliente 1	12H	15	XGB	(100, 5, 0.05, 0)	2,44	0,86	1,56	6,01	0,99
Cliente 10	24H	30	LSTM	(100, 'tanh', 25)	11,68	2,33	3,42	4,57	-0,08
Cliente 10	48H	15	LSTM	(100, 'tanh', 25)	22,7	3,74	4,76	3,54	-0,02
Cliente 10	48H	30	LSTM	(100, 'tanh', 25)	25,49	4,08	5,05	3,85	-0,02
Cliente 11	24H	15	LSTM	(100, 'tanh', 25)	146,28	8,26	12,09	31,56	0,27
Cliente 11	24H	30	LSTM	(100, 'tanh', 25)	141,07	8,16	11,88	27,95	0,13
Cliente 11	48H	30	LSTM	(100, 'tanh', 25)	1199,52	21,86	34,63	36,19	-0,4
Cliente 12	12H	15	XGB	(100, 5, 0.05, 0)	3,09	1,13	1,76	7,73	0,93
Cliente 12	12H	15	XGB	(500, 5, 0.01, 0.1)	3,09	1,13	1,76	7,73	0,93
Cliente 12	12H	15	XGB	(500, 5, 0.01, 0)	3,1	1,13	1,76	7,74	0,93
Cliente 13	8H	30	LSTM	(100, 'tanh', 25)	0,91	0,73	0,95	8,08	0,61
Cliente 13	24H	15	LSTM	(100, 'tanh', 25)	3,08	1,37	1,75	4,94	-0,05
Cliente 13	24H	30	LSTM	(100, 'tanh', 25)	3,16	1,37	1,78	4,95	-0,08
Cliente 14	24H	15	LSTM	(100, 'tanh', 25)	0,47	0,54	0,69	9,49	-0,6
Cliente 14	24H	30	LSTM	(100, 'tanh', 25)	0,56	0,61	0,75	10,39	-0,78
Cliente 14	48H	15	LSTM	(100, 'tanh', 25)	2,27	1,16	1,51	10,45	-0,6
Cliente 15	24H	15	LSTM	(100, 'tanh', 25)	2,68	1,37	1,64	9,84	-0,05
Cliente 15	24H	30	LSTM	(100, 'tanh', 25)	3,42	1,47	1,85	10,26	-0,29
Cliente 15	48H	30	LSTM	(100, 'tanh', 25)	15,08	2,96	3,88	12,02	-0,11
Cliente 16	24H	30	LSTM	(100, 'tanh', 25)	0,63	0,35	0,79	4,2	-0,02
Cliente 16	48H	15	LSTM	(100, 'tanh', 25)	1,23	0,57	1,11	2,74	-0,04
Cliente 16	48H	30	LSTM	(100, 'tanh', 25)	1,51	0,6	1,23	2,95	-0,03
Cliente 17	48H	15	LSTM	(100, 'tanh', 25)	1,44	0,74	1,2	3,18	0,05
Cliente 17	48H	30	LSTM	(100, 'tanh', 25)	1,78	0,73	1,33	3,28	-0,01
Cliente 17	12H	30	XGB	(300, 5, 0.1, 0.1)	0,91	0,5	0,95	4,25	1
Cliente 18	24H	15	LSTM	(100, 'tanh', 25)	1,28	0,7	1,13	6,14	-0,08
Cliente 18	48H	15	LSTM	(100, 'tanh', 25)	2,48	1,01	1,58	3,96	0,03
Cliente 18	48H	30	LSTM	(100, 'tanh', 25)	3,53	1,25	1,88	5,04	-0,13
Cliente 19	48H	30	LSTM	(100, 'tanh', 25)	1,39	0,3	1,18	1,82	-0,04
Cliente 19	8H	30	XGB	(100, 7, 0.1, 0)	0,67	0,23	0,82	2,19	0,99
Cliente 19	8H	30	XGB	(300, 10, 0.05, 0)	0,72	0,23	0,85	2,2	0,99
Cliente 2	12H	30	XGB	(100, 5, 0.1, 0.1)	1,22	0,87	1,11	2,91	1
Cliente 2	12H	30	XGB	(100, 5, 0.1, 0)	1,22	0,87	1,1	2,92	1
Cliente 2	12H	30	XGB	(100, 5, 0.1, 0.2)	1,23	0,87	1,11	2,93	1
Cliente 20	12H	30	XGB	(300, 7, 0.05, 0.1)	17,48	1,81	4,18	9,99	0,91
Cliente 20	12H	30	XGB	(500, 7, 0.05, 0.1)	17,48	1,81	4,18	9,99	0,91

Cliente	Frecuencia	Rezago	Modelo	Hiperparametros	MSE	MAE	RMSE	MAPE	R2
Cliente 20	12H	30	XGB	(300, 7, 0.05, 0)	17,42	1,83	4,17	10,09	0,91
Cliente 21	12H	15	XGB	(300, 5, 0.01, 0)	0,05	0,09	0,22	18,07	0,35
Cliente 21	12H	15	XGB	(100, 5, 0.05, 0)	0,05	0,09	0,23	18,18	0,32
Cliente 21	12H	30	XGB	(500, 5, 0.01, 0)	0,05	0,1	0,21	18,32	0,29
Cliente 22	12H	15	XGB	(500, 10, 0.01, 0.1)	0,11	0,2	0,33	11,62	0,38
Cliente 22	12H	15	XGB	(100, 7, 0.1, 0.1)	0,11	0,2	0,34	11,62	0,34
Cliente 22	12H	15	XGB	(300, 7, 0.1, 0.1)	0,11	0,2	0,34	11,62	0,34
Cliente 23	12H	30	XGB	(300, 3, 0.01, 0.8, 0.8, 0)	0,28	0,34	0,53	11,54	0,03
Cliente 23	12H	30	XGB	(300, 3, 0.01, 0.8, 0.8, 0.1)	0,28	0,34	0,53	11,53	0,04
Cliente 23	12H	30	XGB	(300, 5, 0.01, 0.8, 0.8, 0)	0,28	0,35	0,53	11,58	0,02
Cliente 24	12H	15	XGB	(100, 3, 0.05, 0)	0,07	0,14	0,27	13,73	0,52
Cliente 24	12H	15	XGB	(500, 3, 0.01, 0)	0,07	0,15	0,27	13,81	0,5
Cliente 24	12H	15	XGB	(100, 3, 0.1, 0.1)	0,07	0,14	0,26	13,84	0,55
Cliente 25	12H	30	XGB	(100, 3, 0.05, 0)	0,07	0,17	0,26	11,94	0,39
Cliente 25	12H	30	XGB	(500, 3, 0.01, 0)	0,07	0,17	0,26	11,95	0,39
Cliente 25	12H	30	XGB	(100, 3, 0.05, 0.1)	0,07	0,17	0,26	11,97	0,4
Cliente 26	12H	15	XGB	(100, 5, 0.1, 0.1)	1,11	0,55	1,05	4,26	0,97
Cliente 26	12H	15	XGB	(300, 5, 0.1, 0.1)	1,11	0,55	1,05	4,26	0,97
Cliente 26	12H	15	XGB	(500, 5, 0.1, 0.1)	1,11	0,55	1,05	4,26	0,97
Cliente 27	12H	15	XGB	(500, 5, 0.01, 0.2)	0,98	0,62	0,99	2,53	0,98
Cliente 27	12H	15	XGB	(100, 5, 0.05, 0.2)	1	0,62	1	2,53	0,97
Cliente 27	12H	15	XGB	(500, 5, 0.01, 0.1)	1	0,62	1	2,54	0,97
Cliente 28	12H	30	XGB	(100, 3, 0.05, 0.1)	0,63	0,6	0,79	3,45	0,98
Cliente 28	12H	30	XGB	(100, 3, 0.05, 0)	0,63	0,6	0,79	3,45	0,98
Cliente 28	12H	30	XGB	(500, 3, 0.01, 0.1)	0,63	0,6	0,79	3,45	0,98
Cliente 29	12H	30	XGB	(300, 3, 0.01, 0)	1,39	0,61	1,18	13,17	0,46
Cliente 29	12H	30	XGB	(300, 3, 0.01, 0.1)	1,39	0,61	1,18	13,18	0,46
Cliente 29	12H	30	XGB	(300, 3, 0.01, 0.2)	1,39	0,61	1,18	13,2	0,46
Cliente 3	12H	15	XGB	(500, 3, 0.01, 0)	2,84	1,19	1,68	4,56	0,99
Cliente 3	12H	15	XGB	(500, 3, 0.01, 0.1)	2,84	1,19	1,68	4,56	0,99
Cliente 3	12H	15	XGB	(500, 3, 0.01, 0.2)	2,84	1,19	1,68	4,56	0,99
Cliente 30	12H	15	XGB	(100, 5, 0.05, 0)	3,09	1,13	1,76	7,73	0,93
Cliente 30	12H	15	XGB	(500, 5, 0.01, 0.1)	3,09	1,13	1,76	7,73	0,93
Cliente 30	12H	15	XGB	(500, 5, 0.01, 0)	3,1	1,13	1,76	7,74	0,93
Cliente 4	24H	15	LSTM	(100, 'tanh', 25)	3,46	0,94	1,86	6,23	0,88
Cliente 4	12H	15	XGB	(500, 5, 0.05, 0.1)	0,87	0,64	0,93	6,34	1
Cliente 4	12H	15	XGB	(300, 5, 0.05, 0.1)	0,87	0,64	0,93	6,34	1
Cliente 5	24H	15	LSTM	(100, 'tanh', 25)	145,56	9,44	12,06	10,04	-0,04
Cliente 5	24H	30	LSTM	(100, 'tanh', 25)	160,87	10,11	12,68	10,75	-0,08
Cliente 5	12H	30	XGB	(300, 5, 0.01, 0.9, 0.9, 0)	118,06	7,97	10,87	13,76	0,78

Cliente	Frecuencia	Rezago	Modelo	Hiperparametros	MSE	MAE	RMSE	MAPE	R2
Cliente 6	24H	15	LSTM	(100, 'tanh', 25)	3,05	1,39	1,75	10,71	0,66
Cliente 6	24H	30	LSTM	(100, 'tanh', 25)	2,19	1,22	1,48	9,6	0,76
Cliente 6	48H	30	LSTM	(100, 'tanh', 25)	21,38	3,16	4,62	13,67	-0,04
Cliente 7	48H	30	LSTM	(100, 'tanh', 25)	1181,26	13,83	34,37	10,32	-0,05
Cliente 7	12H	30	XGB	(300, 6, 0.01, 0.8, 0.9, 0)	144,12	6,93	12,01	19,17	0,68
Cliente 7	12H	30	XGB	(300, 6, 0.01, 0.8, 0.9, 0.1)	143,96	6,93	12	19,17	0,68
Cliente 8	24H	15	LSTM	(100, 'tanh', 25)	11,34	2,42	3,37	9,5	-0,03
Cliente 8	48H	15	LSTM	(100, 'tanh', 25)	33,76	5,02	5,81	8,93	-0,47
Cliente 8	48H	30	LSTM	(100, 'tanh', 25)	33,92	4,51	5,82	8,38	-0,41
Cliente 9	12H	30	XGB	(100, 5, 0.1, 0)	0,59	0,3	0,77	6,75	0,99
Cliente 9	12H	30	XGB	(100, 7, 0.1, 0)	0,56	0,3	0,75	6,9	0,99
Cliente 9	12H	30	XGB	(100, 10, 0.1, 0)	0,62	0,32	0,79	6,94	0,99

Anexo 2 Estadísticas Energía Activa por Cliente

	Active_energy							
	count	mean	std	min	25%	50%	75%	max
Cliente								
Cliente 5	19681.0	4.998897	2.440656	0.004968	3.248524	4.779127	6.660707	14.622644
Cliente 18	19500.0	1.433431	1.759566	0.000000	0.339010	0.652623	1.025718	9.943944
Cliente 7	19681.0	2.426154	1.791167	0.000037	1.452873	1.791660	2.976583	8.574905
Cliente 3	19681.0	2.341707	1.458767	0.001658	1.629029	2.016040	2.385611	8.563269
Cliente 10	19681.0	3.545042	1.854900	0.011014	2.142763	2.564699	5.225753	8.343834
Cliente 2	19681.0	2.745060	1.368205	0.027246	2.105917	2.466698	2.788212	8.283679
Cliente 19	19500.0	1.129092	1.432507	0.127157	0.290365	0.550787	0.592899	7.534833
Cliente 8	19681.0	1.579517	1.229197	0.000250	0.941429	1.277382	1.733445	7.322165
Cliente 4	19681.0	1.270241	1.395533	0.000240	0.588655	0.875916	1.104245	7.204313
Cliente 1	19681.0	1.940788	1.449246	0.001715	1.197414	1.494509	1.892287	6.526612
Cliente 6	19681.0	1.407520	1.571658	0.000092	0.426879	0.738173	1.800236	6.365392
Cliente 16	19500.0	1.311767	1.664991	0.000052	0.301177	0.583925	0.721806	6.266824
Cliente 17	19500.0	1.279854	1.534047	0.000000	0.327545	0.618378	0.866101	5.827472
Cliente 20	19500.0	1.086441	1.413696	0.000000	0.231000	0.547359	0.581174	5.503240
Cliente 13	14000.0	0.967254	0.783729	0.001417	0.324044	0.997992	1.323173	5.157014
Cliente 11	14000.0	0.763829	0.791581	0.000620	0.292212	0.541020	0.885325	4.352005
Cliente 27	11415.0	1.870751	0.896604	0.000000	1.643705	1.841252	2.184151	4.328475
Cliente 28	11415.0	1.338501	0.698520	0.000000	1.125700	1.350335	1.563624	3.970402
Cliente 30	11415.0	1.054128	0.793689	0.000000	0.474998	0.854333	1.748458	3.591291
Cliente 12	11415.0	1.054128	0.793689	0.000000	0.474998	0.854333	1.748458	3.591291
Cliente 26	11415.0	1.024692	0.659504	0.000000	0.657796	0.802532	1.693354	3.237979
Cliente 15	14000.0	0.429494	0.375392	0.000106	0.031000	0.359226	0.693359	2.810176
Cliente 29	11415.0	0.296145	0.281160	0.000000	0.198037	0.287116	0.354299	2.731480
Cliente 9	19681.0	0.496010	0.581122	0.000014	0.158866	0.231000	0.491338	2.546763
Cliente 23	8925.0	0.206771	0.142244	0.000000	0.094087	0.209655	0.302123	0.814156
Cliente 14	14000.0	0.171379	0.150261	0.000011	0.031000	0.174064	0.231000	0.781830
Cliente 22	8925.0	0.119089	0.087325	0.000000	0.049028	0.116001	0.172845	0.542551
Cliente 25	8925.0	0.100950	0.075742	0.000000	0.039707	0.097080	0.146911	0.489633
Cliente 24	8925.0	0.065535	0.065195	0.000000	0.017709	0.055565	0.092455	0.445964
Cliente 21	8925.0	0.033676	0.044759	0.000000	0.007479	0.027524	0.039227	0.358380