Natalia Angarita
Johnny Castañeda
Carolina Osorio
Marcell Piraquive

Universidad de los Andes

Modelo de riesgo para clientes hurtadores de energía

1. Resumen.

Una de las problemáticas actuales más comunes en la distribución de energía eléctrica (sobre todo para países en desarrollo), es la vulneración del sistema mediante el cual las empresas prestadoras de servicio suministran energía a los usuarios, que incluye cualquier tipo de manipulación a las herramientas usadas por las empresas para alterar la tarifa cobrada a los usuarios. Lo anterior, produce un servicio riesgoso y por supuesto pérdidas para las compañías en el mercado, por lo que cada vez es de mayor interés para las entidades prestadoras la implementación de estrategias que ayuden a minimizar e identificar oportunamente este tipo de conductas fraudulentas. Mediante una revisión exhaustiva de la literatura relacionada con el robo de energía en hogares, seguida de un análisis detallado de los datos de consumo eléctrico, se propondrá una metodología que combina el aprendizaje supervisado y no supervisado, utilizando algoritmos de clustering para agrupar a los usuarios según comportamientos similares de consumo y técnicas de aprendizaje supervisado para predecir la probabilidad de fraude y determinar variables relevantes para la problemática. La implementación de esta metodología tiene como objetivo detectar y prevenir el fraude de energía en el sector eléctrico, mejorando los indicadores de interés y optimizando la prestación de servicios a los clientes.

Para empezar, realizamos la exploración de datos con el fin de entender las variables que hacen parte de la data y su relevancia, teniendo en cuenta la lógica del negocio. Después, llevamos a cabo un análisis de la calidad de la información mediante la preparación de los datos, identificando valores perdidos.

A continuación, trabajamos en el preprocesamiento de los datos y considerando ese primer análisis, decidimos eliminar algunas columnas que no tenían data relevante para el modelo como el número de observación y la unidad de consumo. Además, realizamos conteo de valores nulos y calculamos el porcentaje con respecto al número total de observaciones para después eliminar las columnas con más del 40% de información faltante pues no aportan en la predicción porque difícilmente podremos hacer un trabajo previo de imputación o eliminación. Las variables ENERGIA_A_INCREMENTAR y ENERGIA_A_RECUPERAR para el conjunto de datos abordado, fueron excluidas debido a que tienen una dependencia directa con la variable resultado.

En esta primera aproximación del proyecto, también realizamos un estudio de las estadísticas descriptivas y algunas gráficas descritas más adelante.

2. Introducción.

Las compañías de servicios públicos en todo el mundo enfrentan de manera constante el acoso de fraudes relacionados con la pérdida no técnica (NTL). En la red eléctrica convencional, el robo de electricidad es la forma principal de fraude de NTL. El robo de electricidad se refiere al uso de energía eléctrica, con o sin un contrato con un proveedor, utilizando un sistema de medición con desvíos totales o parciales, también se considera robo la manipulación de este sistema para alterar sus mediciones. Los métodos comunes de robo varían desde comprometer la seguridad física de los medidores hasta conectar cargas directamente a las líneas de distribución de electricidad.

Históricamente, el robo de energía ha sido característico de los países en desarrollo, donde puede alcanzar gran parte de sus regiones. En los países desarrollados, las pérdidas no técnicas en el sector eléctrico son mínimas o inexistentes, ya que la mayoría de la población puede pagar tarifas que reflejan los costos de suministro. Sin embargo, en los países en desarrollo, muchos servicios públicos eléctricos siguen experimentando grandes pérdidas (Antmann, 2009). Como resultado, las empresas afectadas por este problema están utilizando cada vez más técnicas de gestión del riesgo que les permitan identificar patrones que afecten sus resultados.

Las empresas de distribución de energía tienen acceso a una gran cantidad de información de sus usuarios, que puede utilizarse para abordar problemas internos y externos. Este proyecto tiene como objetivo aprovechar la información técnica y comercial de los usuarios de una empresa de distribución de energía para intentar predecir mediante algoritmos de aprendizaje supervisado y no supervisado, si es probable que estos clientes cometan fraudes de energía o no.

En el mercado de energía eléctrica, el riesgo es un indicador que permite predecir la probabilidad de robo en un grupo de instalaciones. El desafío para las empresas es desarrollar modelos que puedan identificar patrones en los datos históricos y perfilar a posibles infractores.

Para el caso de estudio, el riesgo se calcula utilizando operaciones pasadas y se asigna a las instalaciones actuales según su perfil. La empresa distribuidora de energía utiliza una serie de atributos para determinar las operaciones que cumplen con los requisitos para formar parte de la simulación del riesgo. Estos atributos se relacionan lógicamente entre sí y con ciertas constantes mediante reglas que se utilizan en los perfiles y proporcionan un historial de efectividad, que representa el porcentaje de fraudes e irregularidades encontrados en el volumen de inspecciones generadas o simuladas por la regla. Por ejemplo, si una regla tiene una efectividad del 32%, significa que de cada 100 inspecciones que cumplen con los criterios de la regla, 32 resultan en un fraude o irregularidad.

El modelo de riesgo actual de la empresa distribuidora de energía se basa en la efectividad de las reglas utilizadas en los perfiles. Sin embargo, este modelo presenta una baja correlación entre el riesgo asignado a la instalación y la efectividad en campo. Además, el modelo actual tiende a generar falsos positivos al castigar a clientes que no tienen predisposición al fraude debido a las características del modelo.

El proyecto se centra en la creación de un modelo de aprendizaje supervisado/no supervisado que permita identificar de manera eficaz y acertada un mal uso del servicio por parte de los usuarios mediante la simulación de datos históricos de la empresa distribuidora de energía, para así mejorar los indicadores de interés y prestar un mejor servicio. El objetivo es desarrollar un modelo de clustering que permita identificar grupos de usuarios con características similares para identificar así patrones que puedan indicar un mal uso del servicio. El modelo se basa en técnicas de aprendizaje automático y busca una mejor caracterización de los clientes objetivo en el marco de la lógica de negocio de la compañía cliente.

3. Revisión preliminar de antecedentes en la literatura.

En la red eléctrica tradicional, el robo de electricidad y los fraudes constituyen la forma principal de fraudes NTL, incluso superando las pérdidas técnicas que pueden presentarse y se estima en alrededor del 1% del consumo mundial de electricidad (Targosz, 2009). Por esta razón, las empresas prestadoras de servicio de energía han estado interesadas por mucho tiempo en un método efectivo para minimizar los fraudes relacionados con NTL. Las dos técnicas más usadas para este control son: Instalación de medidores electrónicos (medición inteligente) y aplicación de modelos de estimación.

De acuerdo al estudio realizado por Sridharan y Schulz, la medición inteligente es un método efectivo pero costoso, además es necesario trabajar en la infraestructura para permitir la instalación de medidores eléctricos. Por otra parte, la técnica de modelos de estimación, propuesta por Fourie también se considera efectiva, pero requiere mucho trabajo y al mismo tiempo una alta inversión.

Por eso, actualmente con el desarrollo de nuevas tecnologías y los avances en machine learning, las empresas buscan otras alternativas prácticas que permitan realizar predicciones con cierto grado de exactitud.

Por otra parte, en los estudios de Giraldo y Trejos se evidencia que las efectividades obtenidas en diferentes modelos que abordan el problema de perfilar a los clientes de las empresas prestadoras

del servicio de energía eléctrica como posibles hurtadores, son bajas a la luz de sus cifras, una efectividad superior al 30% para un modelo de riesgo en sector de utilities es buena, por lo general se manejan efectividades por debajo del 20%. Pese a esto en ambos trabajos se superan las efectividades de los modelos anteriores a los propuestos por ellos.

4. Descripción detallada de los datos.

Dentro de la data obtenida se cuenta con 80741 observaciones y 466 variables. Aunque se cuenta con un gran número de datos, se tiene el 27% de nulos a nivel total de la data. Por lo que se optó por mantener las variables que tuvieran menos del 40%. La data fue reducida a 340 variables y se trataron los nulos remanentes. A continuación, en la *Tabla 1* tenemos una descripción estadística de las variables más relevantes para intentar entender la problemática de una mejor manera:

	Promedio de facturación mensual	Revisiones ejecutadas último año	RESULTADO	Suma de los consumos leídos de los últimos 3 meses (kWh)
count	80741.000000	80741.000000	80741.000000	80741.000000
mean	549.539411	0.059226	9.289580	1201.070522
std	2734.419417	0.267345	1.275376	6267.888338
min	0.031250	0.000000	7.000000	0.000000
25%	167.250000	0.000000	10.000000	212.000000
50%	289.125000	0.000000	10.000000	519.000000
75%	536.625000	0.000000	10.000000	917.000000
max	288892.312500	6.000000	10.000000	617300.000000

Tabla 1. Estadísticas descriptivas

- El promedio de facturación mensual oscila entre $0.03 y $288,892 unidades, con una media de $549.53 unidades, esto nos da un indicio de que hay muchas observaciones con promedios de facturación bajos. El 50% de la data tiene un promedio de facturación menor a $289.12 unidades.
- El 75% de las revisiones ejecutadas al último año tienen el valor de cero, lo que da un indicio de que no se está haciendo una buena gestión de control y revisión periódica.
- El 75% de los datos tienen un consumo acumulativo de los últimos 3 meses de 917 kWh o menos y presentan un máximo valor de 617,300 kWh. Se aprecia un rango muy amplio en estos consumos.

Nuestra variable objetivo es la presencia de fraude (catalogada como “7”) y no fraude (catalogada como “10”) las cuales tienen la siguiente distribución en los datos representada en la *Ilustración 1*:

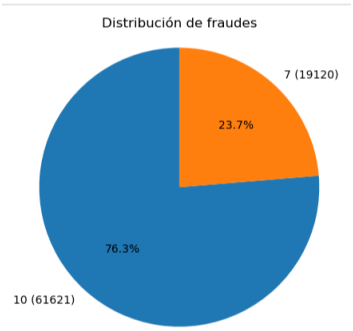
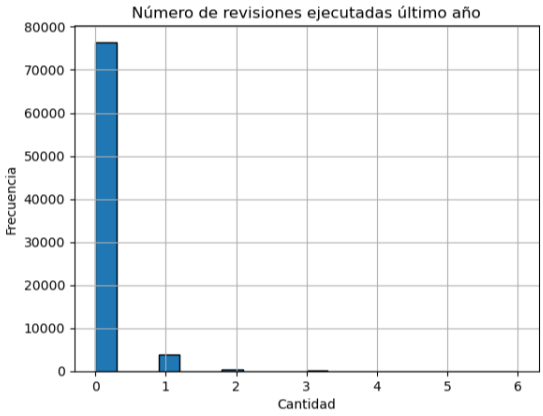


Ilustración 1. Variable objetivo

Podemos observar que 19,120 casos de fraude fueron detectados, lo que es de alto impacto para las métricas de la empresa y se convierte en un problema para evaluar desde la analítica. Dentro de las variables de interés, se encuentra el número de revisiones ejecutadas en los últimos 12 meses (N294) cuya distribución de frecuencia está en la *Ilustración 2*:



Revisiones	Cantidad
0	76465
1	3864
2	344
3	50
4	12
5	4
6	2

Ilustración 2. Distribución de frecuencias revisiones

La gráfica anterior muestra que las revisiones realizadas por la empresa son limitadas ya que estas consumen recursos económicos y personal, de las observaciones únicamente el 5.29% ha sido revisada por lo menos una vez. A continuación, observamos la suma de los consumos en kWh de los últimos 3 meses para las observaciones con valores menores a 100.000.

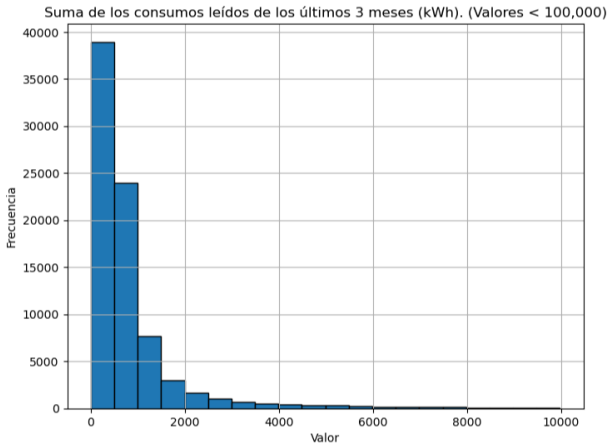


Ilustración 3. Consumos últimos 3 meses

Asimismo, en la *Ilustración 3* es posible visualizar que la distribución de la suma de los consumos en los últimos 3 meses. Se puede evidenciar la complejidad de la problemática a abordar ya que un bajo consumo no siempre implica un fraude y un alto consumo no lo descarta.

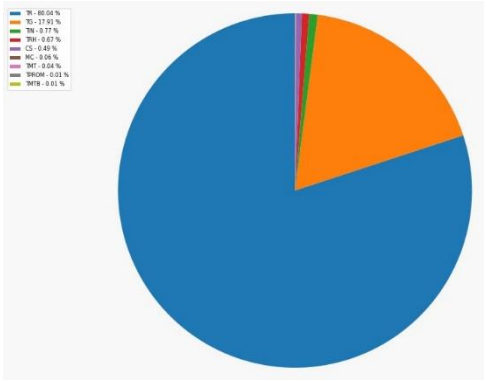


Ilustración 4. Distribución de los usuarios por clase de servicio

Por otra parte, en la *Ilustración 4* es evidente que la mayoría de las instalaciones pertenecen a dos clases, TR y TG, que corresponden a los sectores Residencial y Comercial, sumando entre ellas aproximadamente el 98% de las unidades de consumo. Es una variable que se debe tener en cuanta

debido a que los hábitos de consumo de la clase residencial distan mucho de los comerciales y se reflejaran en los valores de los atributos.

Fuente de datos: Este conjunto de datos pertenece a una empresa prestadora de servicios públicos, específicamente de energía eléctrica, de Centroamérica. Corresponden a simulación histórica realizada el día 3 de septiembre de 2023.

Los datos históricos alojados en la base de datos corresponden a un valor calculado diariamente, no es una imagen fiel del valor de los atributos en el momento en el cual se tomó la decisión de direccionar una visita en campo a una instalación en particular. Esto significa que el valor encontrado en la vista histórica no es necesariamente el mismo al que hubo cuando se realizó el direccionamiento (revisión de la instalación), lo cual introduce variaciones que pueden sesgar el modelo.

5. Propuesta metodológica.

Una de las grandes problemáticas de la data es la similitud entre observaciones y presencia de outliers, por lo que se usará un algoritmo robusto frente a estas dos características. Se elige DBSCAN por su capacidad de identificar outliers junto con un método de reducción de dimensionalidad para identificar variables relevantes en la problemática. También, se intentará involucrar un algoritmo de aprendizaje supervisado para identificar la probabilidad de ocurrencia de fraude en la operación de la empresa ya que los recursos son limitados y no tienen capacidad de realizar revisiones con una frecuencia alta.

6. Bibliografía.

Antmann, P. (2009). Reducing Technical and Non-Technical Losses in the Power Sector. URL: https://openknowledge.worldbank.org/handle/10986/20786.

Targosz , R. (13 de julio de 2009). Electricity theft - a complex problem. URL: http://www.leonardo-energy.org: http://www.leonardo-energy.org/resources/460/electricity-theft-a-complex-problem-581307167ced1

K. Sridharan and N. N. Schulz, “Outage management through amr systems using an intelligent datafilter, ”Power Delivery, IEEE Transactions on, vol. 16, pp. 669–675, 2001.

E. Gontijo, A. Delaiba, E. Mazina, J. E. Cabral, J. O. P. Pinto et al. ,“Fraud identification in electricity company customers using decision tree,” in Systems, Manand Cybernetics, 2004 IEEE International Conference on, 2004.

https://repositorio.utp.edu.co/server/api/core/bitstreams/77fbeabd-e0dc-4fd4-80dd-21b01162ee17/content

https://biblus.us.es/bibing/proyectos/abreproy/70923/fichero/TFM_MSEE_Modelo+basado+en+mineria+de+datos+para..._+Franna+Quezada+Mateo.pdf

Trejos, R (mayo de 2014). “Metodología para la detección de perdidas no técnicas en sistemas de distribución utilizando métodos de minería de datos” URL: https://repositorio.utp.edu.co/server/api/core/bitstreams/c93646cb-19e3-44a6-82bc-3d3167e32504/content

Giraldo, A (enero de 2018). “Desarrollo y aplicación de la metodología bagging y ADABOOST para la detección de pérdidas no técnicas en el sistema de distribución de la empresa de energía de Pereira S.A. ESP” URL: https://repositorio.utp.edu.co/server/api/core/bitstreams/77fbeabd-e0dc-4fd4-80dd-21b01162ee17/content