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 riesgo 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 propone una metodología que aborda el aprendizaje no supervisado, utilizando algoritmos de clustering para agrupar a los usuarios según comportamientos similares de consumo y asignar una probabilidad de fraude. La implementación de esta metodología tiene como objetivo detectar y recuperar económicamente parte de la energía perdida a través del fraude, mejorando los indicadores de interés y optimizando la prestación de servicios a los clientes.

Los datos corresponden a la información técnica y comercial de una empresa prestadora del servicio de energía eléctrica de Centroamérica, más específicamente a los atributos que configuran el insumo principal para el modelamiento del comportamiento de sus usuarios. Estos atributos brindan información relacionada con cada unidad de consumo, tal como la clase, el tipo, diferentes cálculos realizados diariamente sobre su consumo histórico, etc. 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 los valores pertinentes para desarrollar el modelo.

El modelo actual usado por la empresa presenta una baja correlación entre las inspecciones y los fraudes corroborados, es poco flexible ya que es una jerarquización de determinadas características de la unidad de consumo que históricamente se han utilizado sin tener un sustento estadístico, lo que termina castigado a usuarios aversos al riesgo por vecinos amantes al riesgo. Obtuvimos un modelo basado en clustering mediante PCA y utilizando K-Means, tenemos 18 clústeres que obedecen al doble de las clases o tipos de servicio que integran el set de datos logrando una aceptable distribución de las unidades de consumo y unas probabilidades de riesgo más altas que las obtenidas con el modelo anterior, en donde en la mayoría de los casos, casi el 100%, asignaban riesgos bajos y muy bajos. Ahora podemos presentar todas las asignaciones de riesgo, a saber: Muy bajo [0, 0.1), Bajo [0.1, 0.2), Medio [0.2, 0.3), Alto [0.3, 0.4) y Muy Alto [0.5, 1].

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. A nivel nacional, la Asociación Colombiana de Distribuidores de Energía Eléctrica (ASOCODIS) estima que el promedio ponderado por este hecho es de 15,5% dentro de las empresas miembros de la asociación, la zona del país en la que más incidencia tiene el hurto es la Costa Caribe. Esto, coincide con las cifras de hurto que tiene la empresa Air-e, que reemplazó la operación de Electricaribe en Atlántico, La Guajira y Magdalena. El índice que tiene la empresa es de 32% (Morales Soler, El Heraldo, 2022).

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 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.

De hecho, hay amplia literatura con modelos y metodologías para la detección de pérdidas no técnicas que explican diferentes técnicas de aprendizaje supervisado y no supervisado como el proyecto de grado de la Universidad Tecnológica de Pereira, que finalmente asigna etiquetas mediante el a algoritmo de Adaboost y Bagging (Trejos, 2014). Por otra parte, también está el documento entregado por Giraldo que realiza clustering para segmentar a los clientes en tres grupos para realizar etapas de entrenamiento y clasificación final de usuarios (Giraldo, 2018).

Para el caso de estudio, el riesgo es calculado utilizando operaciones pasadas y 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 está centrado en la creación de un modelo de aprendizaje 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 uso fraudulento del servicio. El modelo está basado en técnicas de aprendizaje no supervisado y busca una mejor caracterización de los clientes objetivo en el marco de la lógica de negocio de la compañía cliente.

El modelo resultante cuenta con 18 clústeres que obedece a la lógica del negocio, puesto que hay dentro de los datos 9 tipos diferentes de servicio, estadísticamente hay una menor cantidad, pero castigando mucho la asignación del riesgo. La lógica del negocio que enmarca al modelo desarrollado obedece a hacer una buena caracterización y agrupación de instalaciones, incluyendo las que presenten comportamientos atípicos, y basándose en la historia almacenada para poder asignar una probabilidad de riesgo. Cuando la cantidad de clústeres disminuye, sacrifica la agrupación de instalaciones con consumos o atributos no convencionales que hacen parte de la empresa y que pueden configurar un grupo de interés para el caso.

En los resultados del modelo, tenemos un clúster con una buena cantidad de unidades de consumo y una probabilidad de riesgo superior al 50%, riesgo muy alto, es decir, que históricamente en este grupo de instalaciones que presentan ciertas características similares, de cada dos inspecciones realizadas una de ellas ha sido corroborada como fraude. Esto quiere decir que si vamos a asignar riesgo a una instalación que resulta pertenecer a ese clúster, cuenta con una alta probabilidad de ser un cliente fraudulento, que son la clase objetivo del modelo de riesgo.

3. Materiales y Métodos.

Después de realizar una exploración de los datos, decidimos trabajar con una imagen de la información alojada en MySQL tomada el 25 de septiembre de 2023 (actualizada para la entrega final del proyecto). Los datos históricos que se encuentran disponibles son calculados diariamente y los valores de los atributos no son necesariamente iguales a los obtenidos en el momento en que se identificó y corroboró el estado de la instalación, ya fuese fraude o normal.

En el análisis de la calidad de la información realizado mediante la preparación de los datos, definimos las fases de preprocesamiento de la información, en primera instancia, hallamos atributos que contaban con una gran cantidad de valores nulos, determinamos un umbral del 40% para conservar la mayor información disponible, es decir, eliminamos todos los atributos con más del 40% de nulos en sus valores del dataframe. Segundo, realizamos una clasificación de los atributos haciendo una distinción entre atributos numéricos y atributos categóricos, determinamos como procedimiento de imputación reemplazar los valores nulos de los atributos categóricos con la moda y para los datos numéricos la media. Tercero, identificamos las columnas cuya data no era distinta entre observaciones y eliminamos dichas variables ya que no son dicientes para el modelo. Por último, realizamos un análisis detallado para eliminar las variables con información redundante bajo el marco del negocio. La tabla en la Ilustración 1 brinda información de los datos y además, el porcentaje de fraude de cada clase después del preprocesamiento de los datos.

CLASE DE SERVICIO	DESCRIPCIÓN	CANTIDAD UCS			
		NORMAL	FRAUDE	TOTAL	% FRAUDE
CS	Consumo Social	333	70	403	17,4%
MC	Medidor Colectivo	35	3	38	7,9%
TG	Comercial	11786	2775	14561	19,1%
TIN	Industrial	530	106	636	16,7%
TMT	Media Tensión	29	1	30	3,3%
TMTB	Media Tensión B	10		10	0,0%
TPROM	Promocional	11	1	12	8,3%
TR	Residencial	48973	16077	65050	24,7%
TRH	Residencial Horaria	390	159	549	29,0%
TOTAL		**62097**	**19192**	**81289**	**23,6%**

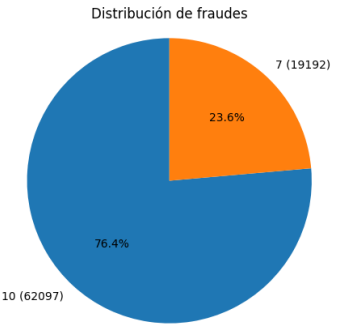

Ilustración 1. Clases de Servicio y Variable Objetivo

Las clases con mayor número de instalaciones son la Residencial (TR), Comercial (TG) e Industrial (TIN).

Posterior al procesamiento de datos obtuvimos 329 variables, por esta razón, técnicas de reducción de dimensionalidad serian útiles en estos casos pues capturan la varianza en menos dimensiones y, por ende, optimizan el modelo. En este caso, decidimos utilizar Análisis por Componentes Principales (PCA). Para esto, desarrollamos los siguientes pasos:

a. Estandarización de variables. Esto es necesario porque las variables pueden tener escalas diferentes y por lo tanto no tendrían varianzas comparables. Sin estandarización, las variables con mayor magnitud podrían dominar los componentes principales, lo cual llevaría a una interpretación errónea. Al tener los datos estandarizados con media cero y varianza 1, nos aseguramos de capturar la mayor información posible al momento de realizar PCA en las diferentes direcciones.

b. Cálculo de Eigenvalores. Los eigenvalores representan la cantidad de varianza que cada componente principal captura de los datos. Los componentes principales con eigenvalores más altos capturan más varianza.
c. Cálculo de Eigenvectores. Los eigenvectores representan la dirección de cada componente principal en el espacio original de las variables. Los coeficientes de los eigenvectores indican cómo cada variable original contribuye a cada componente principal.
d. Selección de componentes principales. Para determinar cuántos componentes principales considerar, utilizamos el criterio del "codo" (proporción de varianza explicada), donde seleccionamos los componentes que capturan una cantidad significativa de varianza antes de que los eigenvalores comiencen a disminuir rápidamente.

La Ilustración 2 nos ayuda a determinar el número de componentes principales a considerar. La línea horizontal roja representa el criterio común de seleccionar componentes con un eigenvalor mayor que 1 (Martínez, C., 2018). Basándonos en este criterio, seleccionamos los primeros 25 componentes principales que tienen eigenvalores mayores que 1, asimismo, sabemos que explican un 82.5% de la varianza en los datos teniendo en cuenta el criterio de Kaiser.

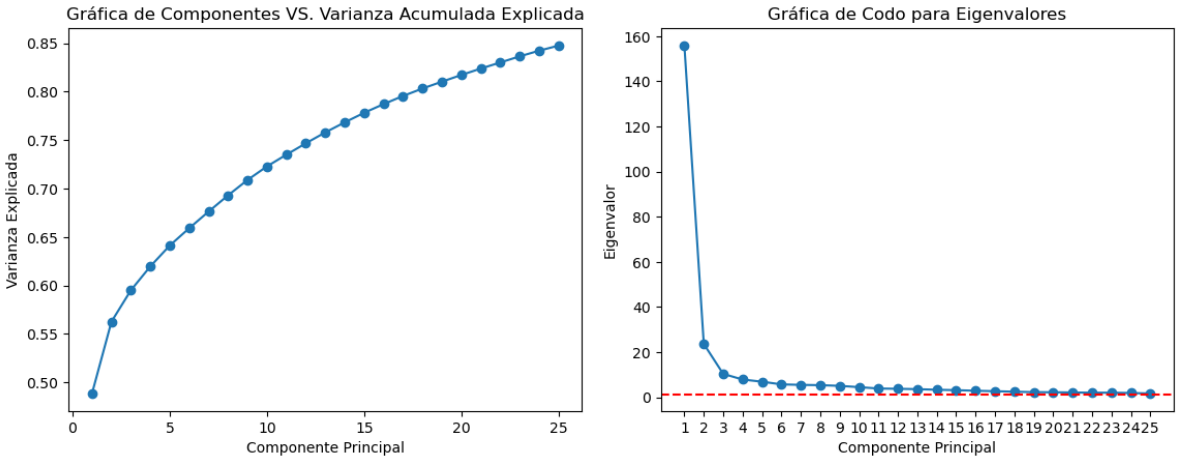

Ilustración 2. Componentes vs Varianza Acumulada Explicada y Eigenvalores

El objetivo del proyecto es desarrollar una herramienta que permita a entes interesados la fácil identificación de fraudes para así administrar mejor sus recursos de control y mejorar la calidad en el servicio. Por lo que decidimos implementar un algoritmo de aprendizaje no supervisado que busque clasificar mediante clustering las observaciones según sus similitudes e intentar determinar el porcentaje de fraudes por clúster. Evaluamos distintos algoritmos, dentro de los cuales optamos por utilizar K-Means, debido a su simplicidad e interpretabilidad, ya que al basar su funcionamiento en centroides permite una fácil comprensión de las características que definen cada grupo.

El algoritmo K-Means agrupa las observaciones en un número predefinido de 18 clusters buscando que la suma de las varianzas internas de los clusters, sea lo menor posible porque de hecho esperamos que las observaciones sean similares y así, permitan segmentar a los diferentes usuarios. *n_clusters: 18 clusters que se van a generar.*

4. Resultados y Discusión.

Dado el contexto del negocio, es de interés para el proveedor de servicios la categorización de 18 clústeres, debido a las 9 clases de servicio expuestas en la tabla superior y el consumo alto o bajo, ya que consideran que este es un factor determinante para cometer fraude. Aunque este valor puede variar en la práctica, los resultados obtenidos son de utilidad para varias entidades interesadas, teniendo en cuenta que según ASOCODIS, el promedio ponderado por este hecho es de 15,5% (por lo que decidimos emplear 20% como umbral de referencia).

Poniendo en contexto lo anterior, se muestran los resultados obtenidos por clustering mediante PCA utilizando K-Means y DBSCAN:

	conteo_0	conteo_1	porcentaje_0	porcentaje_1
Cluster 0	8449	2965	74.02313	25.97687
Cluster 1	32257	8901	78.37358	21.62642
Cluster 2	20	1	95.23810	4.76190
Cluster 3	2	0	100.00000	0.00000
Cluster 4	942	95	90.83896	9.16104
Cluster 5	1131	550	67.28138	32.71862
Cluster 6	3	0	100.00000	0.00000
Cluster 7	7644	1706	81.75401	18.24599
Cluster 8	74	6	92.50000	7.50000
Cluster 9	88	12	88.00000	12.00000
Cluster 10	6	0	100.00000	0.00000
Cluster 11	239	15	94.09449	5.90551
Cluster 12	3527	1223	74.25263	25.74737
Cluster 13	932	217	81.11401	18.88599
Cluster 14	2169	2362	47.87023	52.12977
Cluster 15	3644	823	81.57600	18.42400
Cluster 16	13	0	100.00000	0.00000
Cluster 17	957	316	75.17675	24.82325

RESULTADO CLUSTER	Conteo_0	Conteo_1	Porcentaje_0	Porcentaje_1
-1	519.00000	54.00000	90.57592	9.42408
0	61461.00000	19117.00000	76.27516	23.72484
1	10.00000	4.00000	71.42857	28.57143
2	1.00000	1.00000	50.00000	50.00000
3	3.00000	2.00000	60.00000	40.00000
4	4.00000	1.00000	80.00000	20.00000
5	4.00000	3.00000	57.14286	42.85714
6	5.00000	0.00000	100.00000	0.00000
7	3.00000	4.00000	42.85714	57.14286
8	7.00000	0.00000	100.00000	0.00000
9	11.00000	0.00000	100.00000	0.00000
10	13.00000	0.00000	100.00000	0.00000
11	6.00000	0.00000	100.00000	0.00000
12	11.00000	1.00000	91.66667	8.33333
13	5.00000	1.00000	83.33333	16.66667
14	5.00000	0.00000	100.00000	0.00000
15	5.00000	1.00000	83.33333	16.66667
16	5.00000	0.00000	100.00000	0.00000
17	4.00000	1.00000	80.00000	20.00000
18	5.00000	1.00000	83.33333	16.66667
19	5.00000	1.00000	83.33333	16.66667
20	5.00000	0.00000	100.00000	0.00000

Tabla 1. Resultados K-Means y DBSCAN

Podemos observar que, aunque DBSCAN presenta métricas altas de Porcentaje_1 (Fraude), está agrupando la gran mayoría de observaciones en un solo clúster, lo cual no es información relevante. Por otro lado, la mejor métrica de K-Means logra un 52.1% de detección de fraudes en el clúster 14 evaluando 2362 observaciones.

Una de las limitaciones de este resultado, es que por la naturaleza del algoritmo pueden presentarse outliers de instalaciones que sean clasificadas en un clúster no apropiado, sin embargo, esta problemática podría ser abordada complementando el estudio con la implementación de un modelo supervisado (XGBoost por ejemplo) para lograr así una doble validación de fraude y tomar mejores decisiones.

5. Conclusión.

Abordamos una problemática social vigente (sobre todo en países en desarrollo) como lo es el fraude en instalaciones eléctricas por parte de los consumidores de distintos tipos, procesamos una data con características de las instalaciones y una etiqueta que identificaba la presencia o no de fraude. El algoritmo implementado es clustering por K-Means con reducción de dimensionalidad por PCA.

Los resultados por DBSCAN no fueron seleccionados para el proyecto porque agrupaban la mayoría de instalaciones en un solo cluster y esto no es deseable. Por lo tanto, consideramos K-Means.

Teniendo en cuenta que no hay una solución definitiva para la problemática abordada, podemos decir que los resultados obtenidos por clustering con K-Means son de utilidad para un prestador de servicios ya que permiten focalizar puntos donde puede presentarse fraude de una manera más optima a la convencional. Según los resultados, recomendamos aplicar políticas de control más fuertes en las instalaciones pertenecientes al clúster 14 porque 1 de cada 2 serán encontradas como fraude. También, es recomendable monitorear las instalaciones pertenecientes a los clústeres 0, 5 y 12. Estos clústeres logran agrupar el 36% del fraude total con una probabilidad superior al 25% (considerando que no esta incluido el clúster 1, pues contiene la mayoría de los datos y de incluirse lograría agrupar el 83% del fraude total con una probabilidad superior al 21.6%).

6. Bibliografía.

La República. Daniela Morales Soler (2023). Radiografía de la pérdida de energía por robo. URL: https://www.larepublica.co/empresas/air-e-y-afinia-son-las-empresas-con-el-mayor-indice-de-perdida-de-energia-por-hurto-3297960

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