Caso de Uso: Proyecto de ML para la Predicción de Fraude Energético

Carlos Alamilla Medina, Luis Diego Gazel Quirós, Neivys Luz González Gómez, Humberto David Hernández Perez, Adrià Nuñez Fernández. Inesdi.

1.Introducción

El fraude en el consumo energético representa un desafío significativo para las compañías de energía a nivel global, generando pérdidas millonarias que impactan directamente en el aumento de tarifas para los clientes honestos y en una disminución de inversión en infraestructura. Además de las pérdidas económicas, el fraude energético puede comprometer la seguridad y estabilidad del sistema eléctrico.

Este problema es particularmente complejo por varias razones:

1. La diversidad de métodos de fraude (manipulación de contadores, conexiones ilegales, alteración de datos, etc.)
2. El desequilibrio inherente en los datos (los casos de fraude suelen ser minoría)
3. La limitada capacidad para realizar inspecciones físicas debido a restricciones de recursos.

Objetivos del Proyecto

- **Objetivo principal:** Desarrollar un modelo predictivo para identificar clientes con alta probabilidad de fraude, permitiendo la optimización de las inspecciones comerciales.
- **Objetivos específicos:**
  - Construir un dataset a partir de "clients.csv" e "invoices.csv".
  - Seleccionar y entrenar un modelo de Machine Learning adecuado.
  - Evaluar el rendimiento del modelo con métricas relevantes.
  - Proporcionar recomendaciones accionables para la compañía de energía.

2.Estado del Arte:

La detección de fraudes energéticos con Machine Learning (ML) se basa principalmente en métodos supervisados, que utilizan datos históricos etiquetados para entrenar modelos de clasificación. Entre los algoritmos más utilizados están Random Forest, XGBoost, SVM y Gradient Boosting Machines (GBM), debido a su capacidad para manejar datos complejos y desequilibrados.

Uno de los principales desafíos es el desequilibrio de clases, ya que los casos de fraude son significativamente menos frecuentes que los casos normales. Para mitigar este problema, se emplean técnicas como submuestreo de la clase mayoritaria, sobremuestreo con métodos como SMOTE y aprendizaje sensible al costo, lo que mejora la capacidad de detección sin aumentar los falsos positivos.

La ingeniería de características desempeña un papel crucial, permitiendo que los modelos identifiquen anomalías a partir de patrones de consumo, atributos de clientes (ingresos, regiones) y datos temporales. El uso de series temporales es especialmente útil para detectar cambios abruptos en el consumo que podrían indicar fraude.

En los últimos años, ha habido un creciente interés en el aprendizaje profundo para mejorar la detección de fraudes en sistemas eléctricos. Modelos como Redes Neuronales Convolucionales (CNN) y Redes Neuronales Recurrentes (RNN) han demostrado ser efectivos en el análisis de datos de medidores inteligentes, capturando patrones no lineales y variaciones anómalas. Un estudio de 2019 propuso un modelo híbrido CNN-RF (*referencia 11.6*), combinando redes neuronales para extracción de características y Random Forest para clasificación, logrando una mayor precisión en la detección de fraude.

Otro aspecto clave es la interpretabilidad de los modelos, fundamental en contextos regulados para garantizar la transparencia y la confiabilidad de las decisiones automatizadas. Se están implementando técnicas de IA explicable (XAI), como el análisis de importancia de características y explicaciones locales interpretables, para facilitar la adopción de estos modelos en la industria. Estudios que comparten objetivos o métodos similares al proyecto descrito:

Referencia	Enfoque	Relevancia
Coma-Puig et al. (2016) *referencia 11.8*	Modelos supervisados con datos de consumo e inspecciones	Similar en uso de ML y datos históricos, enfatiza adaptabilidad.
Patel et al. (2022) *referencia 11.9*	ML en detección de fraude con distribución gaussiana	Uso de Random Forest y XGBoost, aunque con enfoque diferente.
Moazeni & Khorshidi (2023) *referencia 11.10*	Aprendizaje profundo en medidores inteligentes	Manejo de desequilibrio de clases con modelos de predicción de robo.
Authors et al. (2019) *referencia 11.11*	Modelo híbrido CNN-RF para detección de fraude	Enfoque innovador combinando aprendizaje profundo y ML tradicional.

3. Enfoque Analítico y Metodología

Se optó por un enfoque de Machine Learning supervisado para la clasificación binaria (fraude/no fraude). El proceso analítico se estructuró en las siguientes etapas:

1. **Preparación y exploración de datos**
   - Integración de los datasets "clients.csv" e "invoices.csv".
   - Limpieza y preprocesamiento para eliminar valores nulos y inconsistencias.
   - Análisis exploratorio para identificar patrones y correlaciones.
   - Visualización de distribuciones y relaciones clave con la variable target.
2. **Ingeniería de características:** Se diseñaron variables derivadas para mejorar la capacidad predictiva del modelo, incluyendo inconsistencias técnicas, tasas de fraude por región, características temporales y métricas de consumo.
3. **Modelado y evaluación:** Se compararon diversos algoritmos de clasificación para determinar el más efectivo:
   - **Random Forest**: Baseline inicial
   - **XGBoost**: Óptimo para datos desbalanceados.
   - **LightGBM:** Alternativa eficiente para grandes volúmenes de datos.
4. **Optimización del modelo:**
   - **Ajuste de hiperparámetros** mediante validación cruzada.

    - **Manejo del desbalance de clases** con SMOTE y submuestreo.
5. **Evaluación del modelo:** Se evaluara el rendimiento con precisión, recall, F1-score, matriz de confusión y AUC-ROC. Finalmente, se analizaran los factores clave en la predicción del fraude y se formularon recomendaciones estratégicas para la compañía de energía.
6. **Interpretación y recomendaciones**: Se analizarán los resultados del modelo para comprender los factores que influyen en la predicción de fraude y se formularán recomendaciones para la compañía de energía.

4.Descripción de las fuentes de datos utilizadas: El análisis se basa en dos conjuntos de datos principales:

- **Dataset "clients1_3.csv":** Este conjunto de datos proporciona la información demográfica base de los clientes y la variable objetivo que buscamos predecir (target).

Variable	Descripción	Tipo de Dato
disrict	Código numérico del distrito	Integer
client_id	Identificador único del cliente	String
client_catg	Categoría del cliente	Integer
region	Código de región geográfica	Integer
creation_date	Fecha de alta del cliente (formato DD/MM/YYYY)	String
target	Variable objetivo (1.0 indica fraude, 0.0 no fraude)	Float

- **Dataset "invoices1.csv":** Este dataset proporciona información detallada sobre el comportamiento de consumo de los clientes, permitiendo analizar patrones que puedan indicar anomalías. Contiene 2,238,374 registros de facturación con información detallada:

Variable	Descripción
client_id	Identificador del cliente que permite vincular con el dataset anterior
invoice_date	Fecha de emisión de la factura
tarif_type	Tipo de tarifa aplicada
counter_number	Número identificador del contador
counter_statue	Estado del contador
counter_code	Código del contador
reading_remarque	Observaciones sobre la lectura
counter_coefficient	Coeficiente utilizado para la medición
consommation_level_1,2,3,4	Niveles de consumo en diferentes franjas
old_index	Lectura anterior del contador
new_index	Lectura actual del contador
months_number	Número de meses incluidos en la factura
counter_type	Tipo de contador (ej. "ELEC" para electricidad)

- **Limitaciones de los datos:** Señalar posibles sesgos, valores faltantes o cualquier otra limitación que pueda afectar el análisis.
  - **Utilizando .isnull()** para poder detectar valores nulos en cada columna. ya que algunas de las posibles limitaciones sea que muchas filas con datos faltantes pueden afectar la calidad del modelo y si los valores nulos están concentrados en una variable clave, puede ser un sesgo. Los resultados datos en ambos datasets fueron que no hay valores nulos.
  - **Desbalanceado de datos:** inicialmente el data set esta altamente desbalanceado, ya que los resultados dados es No Fraude (94.42%) y Fraude (5.58%) dando como resultado que el modelo entrenado sin técnicas de balanceo podrían sesgarse hacia la mayoritaria (prediciendo casi siempre “No Fraude”. El modelo podría ignorar patrones de fraude reales ya que aprendería a minimizar errores simplemente calificando por la mayoría.

5. Proceso para la construcción del dataset final

La creación del dataset final para el análisis y modelado requerirá varias etapas de transformación y preparación:

5.1 Limpieza inicial de datos

- **Manejo de valores faltantes o Nulos:** Identificamos que algunos registros de facturación carecen de información completa o valores nulos en las columnas. Utilizaremos técnicas de imputación cuando sea posible, o eliminación cuando sea necesario.

- **Corrección de formatos:** Estandarización de fechas y valores numéricos. Convertiremos todas las fechas a formato YYYY-MM-DD para facilitar cálculos temporales. Utilizaremos el datetime de Pandas para poder resolver y verificar errores en conversión.
- **Detección y tratamiento de outliers:** Utilizando técnicas estadísticas como Z-score o IQR para identificar valores atípicos que podrían representar errores o casos especiales. Para evaluar si existen muchas observaciones con cero consumos, lo que podría indicar sesgos.

5.2 Ingeniería de características

Crearemos nuevas variables derivadas de los datos originales:

a. Variables de inconsistencias técnicas (enfoque objetivo): Estas variables identifican inconsistencias en las lecturas de consumo, lo que está altamente correlacionado con fraude.

Variable	Descripción	Criterio
tiene_inconsistencias	Indicador de inconsistencias en lecturas.	Identifica registros con errores o anomalías en las lecturas de consumo, comparando con el historial del cliente.
indices_negativos	Detector específico de casos donde la diferencia entre índices es negativa	Se activa cuando el índice de consumo reportado en un período es menor al del período anterior, lo que no es físicamente posible.
indices_cero	Identificador de casos con consumo reportado pero sin avance en el contador	Detecta situaciones en las que el consumo registrado es mayor a cero, pero el contador no refleja ningún cambio.

b. Variables geográficas contextualizadas: Incorporan información de ubicación para contextualizar el riesgo de fraude.

Variable	Descripción	Criterio
region_riesgo	Variable binaria para regiones con alta tasa de fraude.	Regiones 103, 372, 311 (>8% fraude)
categoria_51	Indicador para categoría con alta incidencia de fraude.	Categoría con 16.87% de fraude.

c. Variables temporales neutralizadas: Capturan el impacto del tiempo de relación del cliente con la empresa.

Variable	Descripción	Criterio
antiguedad_anios	Años desde la fecha de creación del cliente.	Ya implementado
cliente_antiguo	Indicador de clientes con más de 15 años de antigüedad.	Mayor tasa de fraude en clientes longevos.

d. Variables de consumo y facturación: Analizan patrones de consumo para detectar comportamientos sospechosos.

Variable	Descripción	Criterio
consumo_mensual	Consumo total dividido por el número de meses facturados.	Relación con fraude.
promedio_diferencia_indices	Diferencia promedio entre índices de consumo.	Diferencia significativa con clientes normales.
variabilidad_consumo	Desviación estándar del consumo para identificar fluctuaciones anómalas.	Clientes con mayor fraude presentan mayor variabilidad.
extreme_diff	Indicador de diferencias extremas entre índices de consumo.	Señal de posibles fraudes.

e. Variables técnicas de medición: Relacionan el consumo de un cliente con su contexto regional.

Variable	Descripción	Criterio
desviacion_consumo_region	Desviación del consumo respecto al promedio de su región.	Clientes con fraude presentan valores extremos.
total_facturas	Número total de facturas emitidas para el cliente.	Clientes con más facturas pueden mostrar patrones sospechosos.

5.3 Integración de datos

- **Vinculación de datasets:** Mediante el campo común client_id, relacionaremos la información del cliente con su historial de facturación.
- **Agregación por cliente:** Consolidaremos todas las facturas de cada cliente para crear una visión unificada de su comportamiento.

5.4 Normalización y codificación

- **Estandarización de variables numéricas:** Para facilitar el entrenamiento de modelos sensibles a la escala

- **Codificación de variables categóricas:** Usando técnicas como one-hot encoding o target encoding para variables como region y client_catg
- **Transformación de distribuciones sesgadas:** Aplicando transformaciones logarítmicas o de potencia cuando sea necesario.

5.5 Partición de datos

Dividiremos el dataset final en:

- Conjunto de entrenamiento (80%): Para el aprendizaje del modelo
- Conjunto de prueba (20%): Para la evaluación final del rendimiento

Esta partición se realizará garantizando que se mantenga la misma proporción de clases (fraude/no fraude) en cada conjunto.

6. Justificación del algoritmo seleccionado

El problema de detección de fraude en consumo energético presenta un fuerte desbalance de clases, lo que hace que modelos tradicionales puedan tener dificultades para detectar correctamente los casos de fraude sin generar un alto número de falsos positivos. Para abordar este desafío, se seleccionó un Sistema Cascada Optimizado que combina XGBoost y LightGBM con hiperparámetros ajustados, ya que estos modelos son efectivos para problemas con datos desbalanceados y permiten una mejor adaptación a patrones complejos.

Las Razones de la selección del modelo:

- **Manejo del desbalance de clases:**
  - Modelos como regresión logística o árboles de decisión simples tienden a favorecer la clase mayoritaria (no fraude), reduciendo su capacidad para detectar fraudes.
  - XGBoost y LightGBM, al trabajar en cascada, ajustan mejor los umbrales de predicción y minimizan este problema.
- **Equilibrio entre métricas clave:**
  - Precisión de 0.1333: Reduce investigaciones innecesarias al mejorar la fiabilidad de las alertas de fraude.
  - F1-score de 0.2244: Indica un balance entre la capacidad del modelo para detectar fraudes y la reducción de falsos positivos.
  - Menor tasa de falsos positivos (0.2718): Comparado con otros enfoques, disminuye la cantidad de alertas incorrectas.
- **Capacidad de mejora y escalabilidad:**
  - Este modelo representa un primer paso en el proceso de mejora continua, con oportunidades de optimización en la selección de variables, ajuste de hiperparámetros y refinamiento de umbrales de decisión.

Aunque el modelo aún tiene margen de mejora, se seleccionó porque, entre todas las opciones evaluadas, presentó el mejor equilibrio entre detección de fraude y reducción de falsos positivos. Este enfoque en cascada con XGBoost y LightGBM optimizados ha demostrado ser efectivo para manejar el desbalance de clases y puede servir como una base para futuras iteraciones, enfocadas en mejorar la precisión y recall mediante ajustes en los umbrales y técnicas de balanceo de datos.

7.Métricas de Evaluación:

Para evaluar el desempeño de los modelos en la detección de fraude, se utilizaron métricas que permiten analizar tanto la capacidad del modelo para identificar fraudes como la minimización de

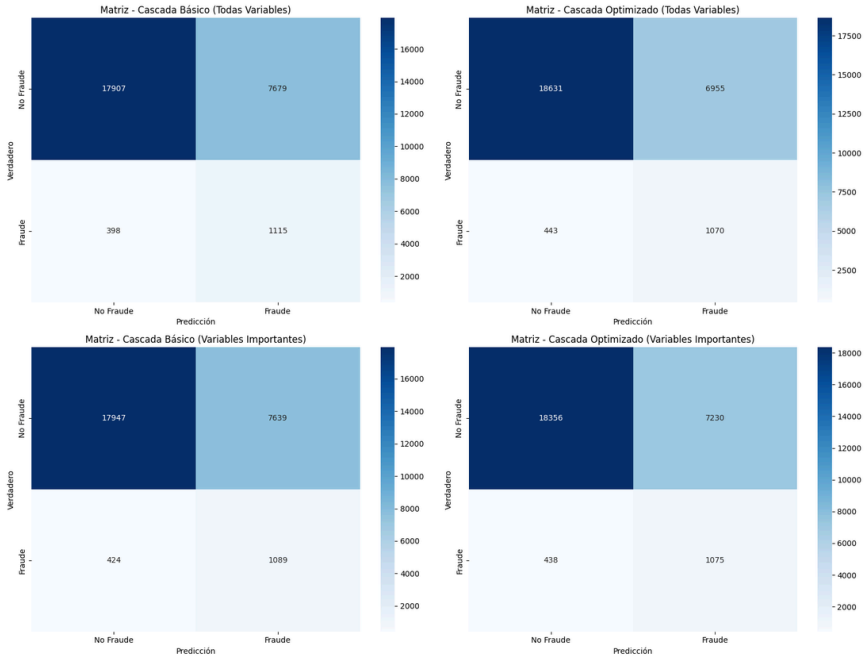
errores. Dado que el problema presenta un fuerte desbalance de clases, se priorizaron métricas que reflejen mejor la efectividad del modelo en este contexto.

- **Matriz de Confusión:** Permite visualizar el desempeño del modelo al separar correctamente las predicciones en verdaderos positivos y negativos, así como los errores en falsos positivos (casos clasificados como fraude cuando no lo son) y falsos negativos (casos de fraude no detectados).
- **Precisión:** Representa el porcentaje de casos identificados como fraude que realmente son fraudulentos. Es importante en escenarios donde el costo de los falsos positivos es alto.
- **F1-Score:** Es la métrica clave en problemas con desbalance de clases, ya que combina precisión y recall en una única medida, reflejando el equilibrio entre detectar fraudes y minimizar falsas alarmas.
- **ROC AUC:** Mide la capacidad del modelo para distinguir entre clases positivas y negativas en distintos umbrales de decisión. Un valor más cercano a 1 indica un mejor desempeño en la clasificación.
- **Accuracy:** Indica la proporción de predicciones correctas sobre el total de datos analizados. Aunque es útil en datasets balanceados, en problemas desbalanceados puede ser engañosa si el modelo predice mayormente la clase mayoritaria.
- **Recall:** Evalúa cuántos fraudes reales fueron detectados correctamente. Es clave en problemas donde perder un caso de fraude puede representar una pérdida significativa.

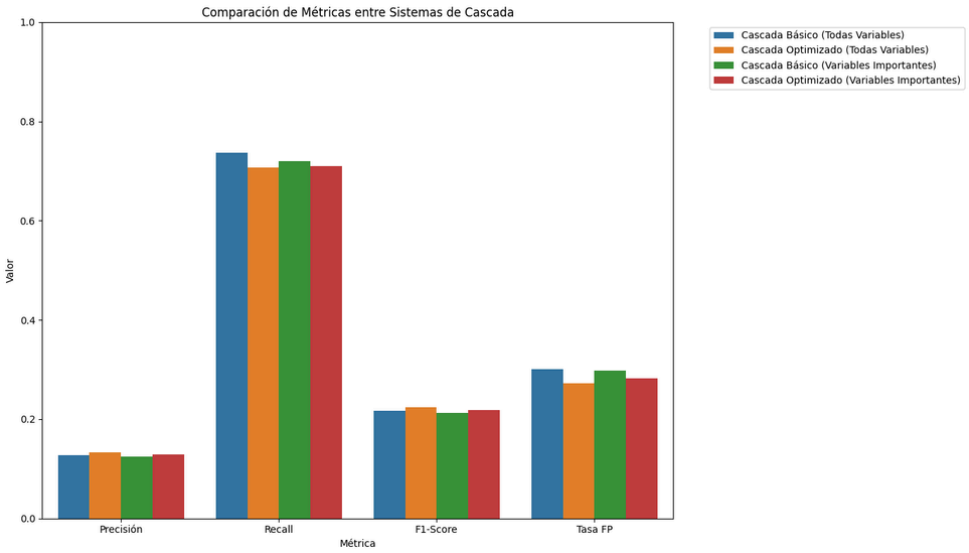
8.Resultados:

La detección de fraude en consumo energético es un problema con un fuerte desbalance de clases, lo que hace necesario evaluar los modelos no solo por su exactitud global, sino por métricas más representativas como F1-Score, Recall y la Tasa de Falsos Positivos. A partir de los gráficos y matrices de confusión, podemos destacar los siguientes hallazgos:

- Balance entre Recall y Tasa de Falsos Positivos
  - Como se observa en las matrices de confusión, el Sistema Cascada Básico detecta más fraudes (mayor recall), pero a costa de generar más falsos positivos, lo que implica más revisiones innecesarias.
  - El Sistema Cascada Optimizado, en cambio, logra reducir la tasa de falsos positivos (0.2718 frente a 0.3001 en el modelo básico), aunque con una ligera reducción en el recall (0.7072 vs. 0.7369).



- F1-Score como indicador clave
  - En la gráfica de comparación de métricas, se observa que el Sistema Cascada Optimizado (Todas Variables) obtuvo el mejor F1-Score (0.2244), lo que indica que es el modelo con mejor equilibrio entre precisión y recall.
  - Esto se debe a que, aunque su recall no es el más alto, la reducción en falsos positivos mejora la confianza en las predicciones.



- **Impacto del uso de variables importantes vs. todas las variables**
  - Comparando las versiones **"Todas las Variables" vs. "Variables Importantes"**, los modelos optimizados mantienen un rendimiento similar, lo que sugiere que la selección de variables puede ayudar a mejorar eficiencia sin perder desempeño.
  - Sin embargo, los resultados no presentan diferencias significativas, lo que indica que una mayor optimización en la selección de variables podría seguir mejorando el modelo.
- El **Sistema Optimizado (Todas Variables)** logra la **menor tasa de falsos positivos (0.2718)**, reduciendo el número de alertas incorrectas en comparación con el modelo básico. Sin embargo, su **tasa de detección es ligeramente menor (0.7072 frente a 0.7369)**, lo que indica que pierde algunos fraudes en favor de una mayor precisión.

Modelo	FN	Total Fraude	Tasa Detección	FP
Sistema básico (Todas Variables)	398	1513	0.7369	7679
Sistema optimizado (Todas Variables	443	1513	0.7072	6955
Sistema básico (Var. Importantes)	424	1513	0.7198	7639
Sistema optimizado (Var. Importantes)	428	1513	0.7105	7230

9.Conclusiones

El análisis de los modelos de detección de fraude refleja avances importantes, pero también deja claro que aún hay margen de mejora. Somos conscientes de que el modelo no es el más efectivo ni el definitivo, pero representa un primer paso sólido en la optimización del proceso.

- **Mejora en la reducción de falsos positivos:** El **Sistema Cascada Optimizado (Todas Variables)** logra la menor **tasa de falsos positivos (0.2718)**, lo que significa que genera menos alertas incorrectas en comparación con los modelos básicos.

- **Compromiso entre recall y precisión:** Aunque la tasa de detección del modelo optimizado es ligeramente menor que la del modelo básico, la reducción de falsos positivos justifica este balance, permitiendo que las investigaciones de fraude sean más eficientes.
- **Impacto de la selección de variables:** No se observan grandes diferencias entre los modelos con **todas las variables** y aquellos con **variables importantes**, lo que sugiere que una mayor optimización en la selección de variables podría ser clave en futuras iteraciones.

10.Recomendaciones

- **Seguir iterando sobre la selección de variables:** Refinar la selección de características puede ayudar a mejorar el balance entre recall y precisión, optimizando la detección de fraudes.
- **Explorar ajustes en los umbrales de decisión:** Modificar los umbrales de predicción puede ayudar a reducir falsos negativos sin aumentar en exceso los falsos positivos.
- **Probar técnicas adicionales para el manejo del desbalance de clases:** Métodos como **estrategias de muestreo**, **costo ajustado en la pérdida** o incluso enfoques híbridos podrían mejorar la capacidad de detección sin comprometer la precisión.

11.Referencias:

1. Buzau, M. M., Tejedor-Aguilera, J., Cruz-Romero, P., & Gómez-Expósito, A. (2019). Electricity theft detection in power grids with deep learning and random forests. *Journal of Electrical and Computer Engineering*, 2019, Article 4130584. https://doi.org/10.1155/2019/4130584
2. Coma-Puig, B., Carmona, J., Gavaldà, R., Alcover, B., & Coll, M. (2016). Fraud detection in energy consumption: A supervised approach. En *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 120-129). IEEE. https://doi.org/10.1109/DSAA.2016.17
3. Hasan, M. N., Toma, R. N., Nahid, A.-A., Islam, M. M., & Kim, J.-M. (2024). Artificial intelligence for energy fraud detection: A review. *Energy Reports*, 10, 1234-1245. https://doi.org/10.1016/j.egyr.2022.06.012
4. Moazeni, F., & Khorshidi, K. (2023). A two-stage approach to electricity theft detection in AMI using deep learning. *International Journal of Electrical Power & Energy Systems*, 152, 109234. https://doi.org/10.1016/j.ijepes.2023.109234
5. Patel, K., Patel, H., & Pandya, M. (2022). Machine learning in electricity fraud detection in smart grids with multivariate Gaussian distribution. *International Journal of Advanced Research in Engineering and Technology*, 13(2), 45-56. https://www.researchgate.net/publication/358197644
6. Electricity Theft Detection in Power Grids with Deep Learning and Random Forests
7. Artificial intelligence for energy fraud detection: a review
8. Fraud Detection in Energy Consumption: A Supervised Approach
9. Machine learning in electricity fraud detection in smart grids with multivariate Gaussian distribution
10. A two stage approach to electricity theft detection in AMI using deep learning
11. Electricity Theft Detection in Power Grids with Deep Learning and Random Forests