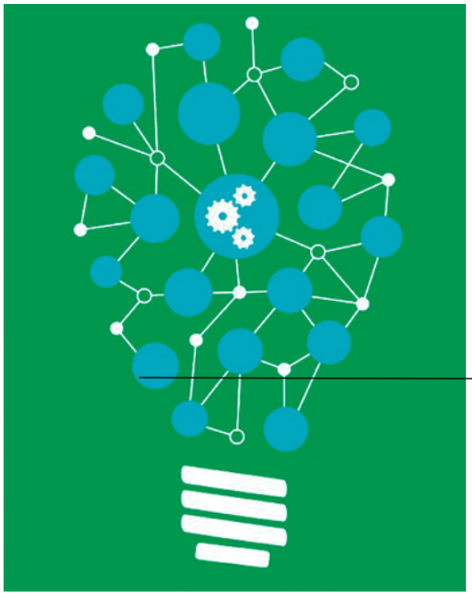
OBS Business School | THREEPOINTS

De:
Planeta Formación y Universidades



Modelación de Potenciales Usuarios que Consumen Energía de manera Fraudulenta

Modelos de Machine Learning: optimización y aplicaciones

Master en Inteligencia Artificial

EQUIPO 7
Wilber Jiménez
Ricard Nogues
Cristian Blanco
Oswaldo Quiñonez
Sergio Ayala
Jhoan Flores

PROFESOR
Alfonso Ibáñez Martín

octubre, 2020

Barcelona

Tabla de Contenidos

1 INTRODUCCIÓN AL PROBLEMA

En el mundo, el sector eléctrico pierde miles de millones de euros al año por culpa de los fraudes. Según Emergin Markets Smart Grid, las pérdidas se estiman para el año 2015 en alrededor de 90 mil millones de dólares anuales. Estos actos ilícitos generan mayores tarifas y peor servicio para los clientes actuales y ponen en riesgo la sostenibilidad de la infraestructura eléctrica. Estos robos pueden resultar en alteraciones de la corriente eléctrica, sobrecargas de los sistemas eléctricos, pérdidas de ingresos para las compañías de energía y peligros a la seguridad pública (incendios o cortocircuitos)

Las empresas de energía gastan recursos tratando de detectar a los infractores, sin embargo, es cada vez más difícil ya que por un lado los mecanismos de detección siguen procesos manuales complejos y muchas veces el infractor cuenta con la complicidad de los funcionarios de las empresas distribuidoras de energía.

Los robos de energía se pueden materializar esencialmente bajo dos metodologías: i) conexión irregular directa a la red del distribuidor; o ii) alteraciones de los equipos de medida y control (contadores y demás dispositivos).

El objetivo de este caso es detectar aquellos clientes potenciales de consumos fraudulentos de energía de cara a facilitar las visitas comerciales a dichos lugares, resolviendo la problemática del negocio a través de la Inteligencia artificial y más específicamente utilizando técnicas de aprendizaje automático.

Lo que se busca en este caso es:

- Desarrollar un mecanismo de predicción capaz de monitorear el fraude energético de los usuarios.
- Presentar a la compañía de una manera gráfica los resultados que el modelo genere, producto de los consumos mensuales de los usuarios.
- Proporcionar un porcentaje de los posibles fraudes que se presentan, con el grado de precisión que el modelo entregue.
- Ejecución mensual y predicción para el mes siguiente basados en los consumos reales.
- Programar visitas semanales a los infractores y lanzar campañas de prevención al fraude.

2 DETALLE DEL ENFOQUE ANALÍTICO PROPUESTO

En nuestro estudio se propone el enfoque de Clasificación binaria (Fraude/ No fraude) con lo que nos centraremos en el problema expuesto desde un punto de vista de aprendizaje SUPERVISADO y algoritmos de CLASIFICACIÓN.

Para ello, nos valdremos de toda clase y tipología de datos, desde datos históricos de consumo de energía de la base de datos de clientes e históricos de clientes que cometieron fraude en el pasado, pasando por tipología de contadores (contador-medidor de energía), región donde se ubica el contador, categoría de clientes (si son particulares, empresas, etc..), lecturas de consumo

totales, lecturas de consumo mensuales, gracias a los datos de facturación, entre otros muchos datos.

Generaremos un score a los clientes sobre la posibilidad de que hayan defraudado a la compañía de energía eléctrica en el pasado y dichos patrones nos servirán para predecir si los clientes actuales pueden estar cometiendo fraude o no. Dicho score será el que usaremos como 'target' para nuestro modelo predictivo. Aquellos clientes con mayor score serán los que consideraremos candidatos a confirmar que realmente estén realizando fraude eléctrico.

Para discriminar que regiones y/o distritos son más propensos a realizar fraude, podremos comparar la energía suministrada por las distintas redes de distribución y subestaciones eléctricas versus la energía facturada en dichas redes y subestaciones, este resultado debería darnos el porcentaje de "energía fugada", por supuesto algunas veces se deberá a averías, y pérdidas justificables, pero en otros casos nos darán la zona exacta (distrito y región) donde se estuviera cometiendo el fraude, y poder así enviar técnicos de la compañía eléctrica a realizar revisiones periódicas de las instalaciones propias de la compañía así como en casos demostrados realizar revisiones de contadores y aparatos de medición en las viviendas o locales y verificar 'in situ' el trucaje de éstos.

En ocasiones puede que se detecte un aumento inusual en el consumo de energía de los clientes, este indicador puede decirnos que ese cliente puede ser víctima de que algún vecino se haya 'enganchado' ilegalmente a su instalación lo que nos indicaría que habría fraude, aunque este no se habría cometido por el cliente al que se detecte el aumento en el consumo, sino de algún vecino cercano que deriva la energía para su uso particular. Este fraude podría tener como origen, la ocupación de viviendas abandonadas de forma "ilegal", con el consiguiente "enganche" a los contadores de los vecinos de estas viviendas abandonadas, esto podría justificar la aparición de Outliers en el data set ya que dichas viviendas no tendrían contador propio por lo que su seguimiento sería un poco más complejo.

La tipología de los contadores también puede darnos información valiosa en referencia al fraude, ya que los antiguos contadores de consumo, requerían la lectura real por técnicos de la compañía; la realización de lecturas mensuales por técnicos de la compañía no se podían realizar de forma regular cada mes, sino que las lecturas reales servían de muestra para realizar estimaciones en el consumo en los meses en que el empleado de la compañía no realizaba las lectura 'in situ' frente al contador, esto podía generar que las anomalías en el consumo de energía no se detectaran pasados unos meses haciendo también que los datos de la compañía eléctrica basados en el consumo real de los usuarios de la red no fuera real generando datos inconsistentes en las bases de datos. Estos mismos contadores eran fácilmente manipulables con lo que se podían trucar 'los limitadores' de potencia obteniendo más potencia de la que realmente hubiera contratado el cliente y por el que pagaría unos costes fijos en base a la potencia contratada. Con los nuevos medidores digitales inteligentes de consumo se obtienen lecturas de consumo en tiempo real, con lo que se abandonan las practicas de facturación sobre estimaciones, y dificulta mucho más el trucaje de los aparatos ya que cualquier cambio en su configuración se detectaría instantáneamente por la compañía eléctrica.

La detección de anomalías en el consumo eléctrico será una de las bases para determinar el posible fraude cometido, aunque como se ha explicado anteriormente, esto no nos dará la prueba final de que un cliente esté cometiendo el fraude supuesto, por lo que hay que tener en cuenta los falsos positivos y cualquier predicción obtenida por el modelo se debería acompañar de procedimientos y medidas que corroborarán dicha predicción, como visitas presenciales

hasta el contador de energía, posibles fugas de energía debidas a causas externas al cliente como mantenimiento deficiente de las instalaciones de la compañía, averías, entre otras, a su vez esto podría ayudar al plan general de mantenimiento de la compañía, evitando o anticipándose a posibles deterioros de las instalaciones y realizar un mantenimiento proactivo con la ayuda de conexiones IoT de sensores que remitan información del estado de la red en tiempo real.

En nuestro modelo queremos potenciar la métrica Precision ya que realmente queremos identificar aquellos clientes que realizan fraude, ya que las medidas que deberán acompañar a las predicciones implicarán visitas presenciales a fin de descartar falsos positivos, aunque las métricas finales se determinarán una vez analizado el dataset y el resultado obtenido.

3 DESCRIPCIÓN DE LAS FUENTES DE DATOS UTILIZADAS

Para este trabajo se hace referencia a una base de datos de la compañía The Tunisian Company of Electricity and Gas (Túnez), disponible a través de código abierto en la plataforma Kaggle, a la cual se puede acceder a través del link [4]. Los datos se bajaron de la red y que consta de la principal variable es el consumo de energía registrada por usuario mensualmente. Adicionalmente, se adjuntará a este informe un Notebook con el dataSet respectivo

Esta base de datos está compuesta por registros de consumo energético para 135 493 clientes, con su respectivas variables de zonas geográficas, tipos de tarifas, regiones, entre otras.

A su vez, cuenta con la información asociada a 201 893 contadores eléctricos con los cuales se lleva la contabilidad de los flujos energéticos.

Acerca de ellos, a continuación, una breve descripción:

- En el contador. Existe una diversidad de métodos para modificar cómo mide el consumo el contador. Los nuevos equipos de medida han aumentado significativamente la protección contra manipulaciones e intrusiones, y resulta mucho más probable la detección de las manipulaciones. Además, si están conectados a una Red eléctrica inteligente, el contador dará aviso a la compañía distribuidora de que se está llevando a cabo una manipulación.
- En el Interruptor de control de potencia. En la mayoría de los suministros con contadores de generaciones previas, el ICP está separado del contador. La manipulación del ICP busca consumir energía por encima de la potencia que se haya contratado. En los suministros con contadores digitales con telegestión, el ICP se haya integrado en el propio contador.
- En el suministro de otros usuarios. El defraudador puede hacer una conexión desde el suministro de un vecino al suyo. En dicho supuesto, toda la energía que llegue a la red del defraudador (calculable mediante la Ley de tensiones de Kirchoff) se medirá en el contador de la víctima de fraude, a quien le llegará el cargo correspondiente en su factura.

4 PROCESO PARA LA CONSTRUCCIÓN DEL DATASET FINAL

Con tal información base, se generan tratamientos de los datos para facilitar el cálculo de la probabilidad de que un usuario sea fraudulento. Para ello se generan, para cada usuario, las variables asociadas a sus consumos mínimos, máximos, medios y su desviación estándar, para detectar anomalías usuario por usuario. Adicionalmente se crean variables asociadas a las variaciones de consumos en el tiempo del comienzo y fin de un periodo de facturación para contrastarlo con los valores de los clientes. Finalmente se crea una variable asociada a los números de transacciones energéticas por fracciones de tiempos para revisar picos y caídas anómalas, a la vez que se analizan las correlaciones entre las principales variables y se descartan las de menor incidencia.

4.1 Data Acquisition

Es el proceso comenzó con la obtención, filtrado y limpieza los datos antes de que estos sean procesados por los algoritmos de Machine Learning (ML).

Se espera tener en esta etapa:

- Gran volumen de información: consumo de energía de los usuarios, sectores, tipo de usuario, sector, etc.
- Información del CRM de la compañía – Históricos de fraudes, reconexiones, sitios de difícil acceso, zonas de tolerancia, etc.
- Fuentes adicionales y externas a la compañía, sectores vulnerables, sectores industriales de empresas pequeñas y medianas.

4.2 Data Comprehension

La data comprehension es lo que se realizó luego de la preparación de la data completa, la data ha sido visualizada, en esta etapa se logró:

- Un análisis exploratorio inicial sobre los datos de ingesta.
- Univariante (solo una variante está involucrada): diagramas de caja, histogramas, diagramas de barras, dispersiones, datos estadísticos (promedio, media, máximos, mínimos), etc.
- Multivariante: gráficos de dispersión, matrices correlación, barras apiladas, dispersiones, importancia de las variables, etc.

4.3 Data Preparation

Se procesó antes de pasar los datos a los algoritmos de ML. Aquí se obtuvo:

- Generación de variables derivadas, asociadas al consumo de energía de los meses pasados.
- Consumo de los usuarios (familias/pequeñas, medianas y grandes empresas).
- Tratamiento de variables (filtros, transformaciones, valores outliers como casas o apartamentos desocupados).

- Dataset de usuarios: Definición de las variables predictoras.
- Selección variables: mRMR – mínima redundancia y máxima relevancia.
- Dataset de usuarios – Reducción de variables predictoras (las más importantes).

5 JUSTIFICACIÓN DEL ALGORITMO SELECCIONADO

Con el problema planteado, buscamos predecir la probabilidad que un cliente cometa o no alguna actividad fraudulenta. Esto nos lleva ante un problema de clasificación binaria (Fraude o no Fraude); por lo que algoritmos de clasificación serían los ideales.

Entre estos podemos mencionar, por ejemplo, a la regresión logística, los árboles de clasificación, los SVM, los Random Forest, las redes neuronales, los modelos de gradient boosting, etc. El cuadro adjunto muestra un resumen de estos algoritmos y algunas de sus características:

Algoritmo	Tipo	Algunas características:
1.- De Facil Interpretabilidad ("White Box" Models)		
Regresión Logistica	S	El aporte marginal de cada varible sobre la probabilidad estimada es completamente concido (parámetros), capaz de reconocer algunas relaciones no lineales, los datos requieren una limpieza previa (missings, outliers, etc.).
Árbol de Clasificación	S	Interpretación fácil y directa, lidia muy bien con problemas en los datos (missings y outliers) por lo que no requiere tratamiento previo, propensos a overfitting (se debe prever a traves de la poda)
2.- De difícil interpretabilidad ("Black Box" Models)		
Máquina de Soporte Vectorial - SVM	S	Reconoce muy bien patrones no lineales (Kernel), no son tan propensos al overfitting, requiere un soporte computacional intenso.
Random Forest	S	Reconoce muy bien patrones no lineales, es paralelizable.
Red Neuronal	S	Resuelven problemas complejos, muy bueno reconociento comportamientos no lineales (a través de la capas ocultas), computacionalmente muy costoso.

Figura 1: Algoritmos de clasificación. Fuente: elaboración propia

Los dos primeros algoritmos ofrecen una interpretabilidad bastante alta que permite conocer el impacto (o peso) de cada una de nuestras variables y como estas afectan a tener una menor o mayor probabilidad de fraude. Pero esa ganancia en interpretabilidad hace que sean menos eficientes al momento de explicar relaciones no lineales en los datos. Mientras que los algoritmos del grupo 2, lidian bastante bien con las no linealidades, pero ofrecen poca o nula interpretabilidad.

Para el problema que estamos abordando asumiremos el no interés en la interpretabilidad (básicamente por dos razones: el fenómeno a estudiar tiende a ser un low-rate model, lo que significa que no se tienen muchas observaciones de fraude positivo; y por la existencia de no linealidades que son más persistentes en el mundo real), por lo que modelos del segundo grupo

son los principales candidatos.

Si bien hemos decidido optar por un modelo ML poco o nada interpretable, se evalúan de igual forma modelos del primer grupo. Así, se presentarán varias opciones de modelos que competirán entre sí y se precisarán sus ventajas y desventajas al momento de definir el modelo final. Su comparativo global se hará a nivel de su métrica de evaluación (definido en el apartado siguiente al igual que las métricas obtenidas por cada modelo).

Una vez seleccionado el tipo de algoritmo que propone dar solución a nuestro problema, el siguiente paso es su ejecución. Para ello podemos hacer uso de librerías que faciliten el proceso de modelamiento como PyCaret. Es más, esta opción nos permite manejar más algoritmos y compararlos entre sí; abriendo el abanico de posibles mejores soluciones a nuestro problema.

6 JUSTIFICACIÓN DE LA MÉTRICA DE EVALUACIÓN UTILIZADA

Para medir la efectividad de detectar el fraude estimada por nuestro modelo, podemos utilizar varias métricas. Para nuestro caso utilizaremos la matriz de confusión para evaluar el resultado del modelo y las métricas que de este devengan (véase figura 2 con gráfico adjunto):

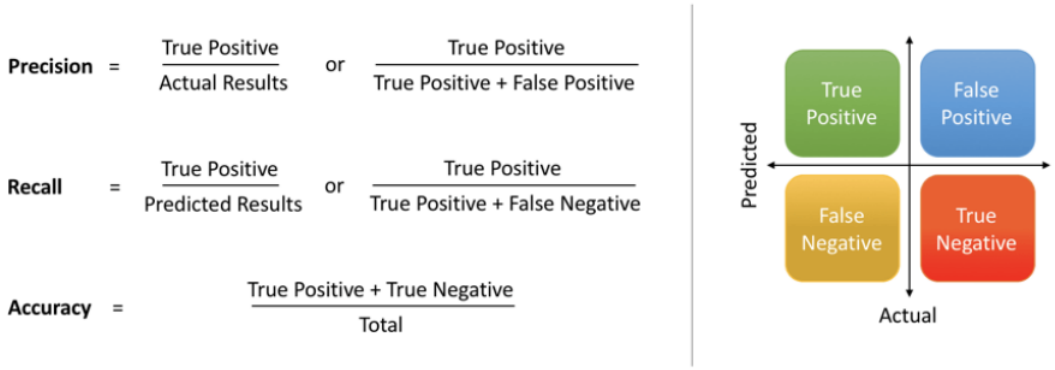


Figura 2: Gráfico accuracy-recall-precision. Fuente: [3]

Una de las principales métricas usadas para determinar qué tan bueno es un algoritmo comparándolo con el dato observado es el accuracy. Que está definido como el total de predicciones correctas (tanto para los casos positivos como negativos) sobre el total de predicciones.

Pero esta métrica tiene un problema para resolver nuestro caso en particular, principalmente por que los fraudes son eventos muy raros (se tiene una tasa baja del fenómeno) por lo que el accuracy nos podría dar una mala aproximación del desempeño del modelo. Por ejemplo, consideremos que nuestra tasa de fraude es de 2 %. Si creamos una simple regla y diríamos que ningún cliente comete fraude entonces nuestro accuracy sería de 98 %; es decir estamos acertando al 98 % de los datos de forma correcta; pero estaríamos dejando de lado a las observaciones que verdaderamente nos importan, que son los fraudes. Por esta razón el accuracy no es una métrica

óptima para nuestro problema.

Por otra parte, el recall y la precisión, serían métricas más apropiadas para nuestro problema. El recall nos permite cuantificar que tanto de los casos positivos (fraudes) verdaderamente son capturados por el modelo y la precisión, que proporción de los predichos como positivos lo son en verdad. Con estas métricas podemos cuantificar exactamente cuántos casos de los fraudes estamos atrapando con el modelo.

Adicional a estas métricas estamos incluyendo el AUC (Area under the Curve) usado ampliamente en modelos de clasificación binaria. El AUC proporciona una medición agregada del rendimiento en todos los umbrales de clasificación posibles. Una forma de interpretar el AUC es como la probabilidad de que el modelo clasifique un ejemplo positivo aleatorio más alto que un ejemplo negativo aleatorio. El AUC es conveniente por las dos razones siguientes:

- El AUC es invariable con respecto a la escala. Mide qué tan bien se clasifican las predicciones, en lugar de sus valores absolutos.
- El AUC es invariable con respecto al umbral de clasificación. Mide la calidad de las predicciones del modelo, sin tener en cuenta qué umbral de clasificación se elige.

Llevando estas ideas a la práctica, hemos realizado un ejercicio practico que resuelve el problema que nos estamos planteando. Se utilizó la librería PyCaret para su cálculo y los resultados son mostrados en la figura 3 adjunta (los detalles de las bases de datos, variables y el proceso pueden ser revisados en el código anexo a este documento):

```
compare_models()
```

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
0	Extreme Gradient Boosting	0.9442	0.6687	1.0000	0.9442	0.9713	0.0005	0.0048	20.7270
1	Light Gradient Boosting Machine	0.9442	0.6630	1.0000	0.9442	0.9713	0.0015	0.0151	1.3775
2	Naive Bayes	0.9441	0.5715	0.9998	0.9443	0.9712	0.0036	0.0225	0.1677
3	Ridge Classifier	0.9441	0.0000	0.9999	0.9442	0.9712	0.0018	0.0096	2.6344
4	Ada Boost Classifier	0.9441	0.6637	0.9999	0.9442	0.9713	0.0018	0.0131	7.0484
5	Gradient Boosting Classifier	0.9440	0.6679	0.9998	0.9442	0.9712	0.0002	0.0012	24.6799
6	CatBoost Classifier	0.9440	0.6697	0.9997	0.9442	0.9712	0.0020	0.0115	41.1231
7	Linear Discriminant Analysis	0.9431	0.6437	0.9987	0.9443	0.9707	0.0041	0.0138	3.8442
8	K Neighbors Classifier	0.9413	0.5508	0.9962	0.9447	0.9697	0.0166	0.0329	6.1708
9	Random Forest Classifier	0.9375	0.5832	0.9916	0.9449	0.9677	0.0238	0.0348	1.2178
10	Extra Trees Classifier	0.9349	0.5824	0.9883	0.9452	0.9663	0.0307	0.0403	11.8492
11	Decision Tree Classifier	0.9333	0.5753	0.9864	0.9453	0.9654	0.0331	0.0412	1.6756
12	SVM - Linear Kernel	0.7644	0.0000	0.7974	0.9535	0.7782	0.0046	0.0126	1.0424
13	Quadratic Discriminant Analysis	0.0683	0.5301	0.0140	0.9516	0.0275	0.0002	0.0036	2.9328
14	Logistic Regression	0.0558	0.5000	0.0000	0.0000	0.0000	0.0000	0.0000	0.5259

```
XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1,
              colsample_bynode=1, colsample_bytree=1, gamma=0,
              learning_rate=0.1, max_delta_step=0, max_depth=3,
              min_child_weight=1, missing=None, n_estimators=100, n_jobs=-1,
              nthread=None, objective='binary:logistic', random_state=5991,
              reg_alpha=0, reg_lambda=1, scale_pos_weight=1, seed=None,
              silent=None, subsample=1, verbosity=0)
```

Figura 3: Compare models - PyCaret. Fuente: elaboración propia

Como vemos, el resultado nos ordena los modelos por mejor AUC, Recall y Accuracy. Donde vemos que tenemos como modelo ganador al XGBoost (Extreme Gradient Boosting); ratificando lo mencionado en el apartado anterior al elegir como modelo optimo un modelo ML de caja negra.

```
extr_grad_bost=create_model('xgboost')
```

	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.9441	0.7032	1.0000	0.9441	0.9713	0.0000	0.0000
1	0.9444	0.6613	1.0000	0.9444	0.9714	0.0051	0.0505
2	0.9443	0.6588	1.0000	0.9443	0.9713	0.0000	0.0000
3	0.9443	0.6681	1.0000	0.9443	0.9713	0.0000	0.0000
4	0.9441	0.6476	1.0000	0.9441	0.9713	0.0000	0.0000
5	0.9440	0.6754	0.9998	0.9441	0.9712	-0.0003	-0.0030
6	0.9441	0.6839	1.0000	0.9441	0.9713	0.0000	0.0000
7	0.9441	0.6570	1.0000	0.9441	0.9713	0.0000	0.0000
8	0.9441	0.6840	1.0000	0.9441	0.9713	0.0000	0.0000
9	0.9441	0.6477	1.0000	0.9441	0.9713	0.0000	0.0000
Mean	0.9442	0.6687	1.0000	0.9442	0.9713	0.0005	0.0048
SD	0.0001	0.0170	0.0000	0.0001	0.0001	0.0015	0.0153

Figura 4: Resultados - PyCaret. Fuente: elaboración propia

Una vez seleccionado el modelo, que tiene un AUC de 66.9 % valores altos de recall y accuracy; verificamos su estabilidad aplicándolo sobre la base de test (para nuestro ejercicio hemos particionado en un 70 % – 30 % los datos).

El resultado nos muestra efectivamente que en la base de test se mantienen lo obtenido en la base de train; garantizando la extrapolación del modelo a poblaciones distintas.

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	Extreme Gradient Boosting	0.9441	0.6719	1.0	0.9442	0.9713	-0.0001	-0.0014

Figura 5: Resultado Extreme Gradient Boosting. Fuente: elaboración propia

7 ANÁLISIS DE REFERENCIAS CON PLANTEAMIENTO SIMILARES

7.1 Caso de aplicación de árboles de decisión

En [2] se presenta un modelo para identificar a posibles clientes fraudulentos que serán inspeccionados dentro de un grupo de usuarios sospechosos con el objetivo de minimizar los costos de inspección física. Se usó una base de datos que contiene el consumo mensual durante un año de 600 000 clientes y 52 atributos que los describen. En la etapa de filtrado de los datos sólo se seleccionaron los 5 atributos que representaban más correlación con la salida (3 variables discretas y 2 variables continuas) y los clientes con inspección de los últimos 12 meses. Este subconjunto se usó para entrenar el algoritmo. Los resultados obtenidos del modelo propuesto no fueron muy satisfactorios, debido a que sólo el 40 % de los clientes preseleccionados para ser inspeccionados

en campo resultaron con fraude, sin embargo, era más eficiente que el que manejaba la compañía. La tasa de éxito se incrementó de un 5 % a 40 %.

7.2 Caso de aplicación de redes neuronales artificiales

En [5] se aplica ELM (Extreme learning machine). El objetivo es extraer patrones de consumo que presentan una alta correlación con las actividades de fraudes a partir de los perfiles de carga de los clientes y otros datos generales. La base de datos usada contiene los perfiles de carga de 1 500 clientes comerciales obtenidos de los contadores inteligentes cada 30 minutos sobre un período de un año de una compañía eléctrica de Malasia, los cuales se agrupan de acuerdo a la similitud en el comportamiento de consumo y el día tipo de la semana con el fin de seleccionar los perfiles más representativos como medio de referencia para la clasificación, y además mejorar la precisión del algoritmo. La técnica de agrupamiento empleada fue C-means. Para categorizar el comportamiento de consumo en anormal o normal se usaron la media y la desviación como punto de referencia, además se implementó un módulo de clasificación para el análisis más profundo. El modelo empleado se evaluó con dos funciones de activación diferentes, la sigmoide y la RBF (radial basis function). La función que logró un mejor desempeño en el algoritmo fue la sigmoide, una precisión de 54.37 %.

7.3 Caso de aplicación de k vecinos más cercanos (k-nearest neighbors)

Esta técnica tiene una amplia gama de aplicaciones, la misma se puede usar para la clasificación de objetos, para el agrupamiento de datos con comportamientos similares y para la detección de atípicos. En [1] señalan esta técnica dentro de los cuatro modelos propuestos para evaluar el desempeño de la clasificación con respecto a la porción de las clases y las características o atributos seleccionados para el entrenamiento. La base de datos proporcionada para el experimento contiene los registros históricos de 700 000 clientes durante un período de enero 2011 a enero de 2015 y algunos atributos como ubicación, tipo de cliente, fecha de lectura, entre otras. Además, se cuenta con 400 000 datos de inspecciones. Para el entrenamiento se dividió el conjunto de instancias en 10 subconjuntos de tamaño similares con el fin de evitar el sobreajuste de los datos. La métrica de evaluación para verificar la robustez de los clasificadores ante estas condiciones fue el área bajo la curva (AUC). En este enfoque se observó que todos los modelos mejoran significativamente su desempeño cuando el conjunto de datos está equilibrado e incluye características tanto de las series de tiempo como datos claves del cliente.

Bibliografía

[1] Glauner, P., Meira, J. A., Dolberg, L., State, R., Bettinger, F., & Rangoni, Y. (2016). Neighborhood features help detecting non-technical losses in big data sets. *Proceedings of the 3rd IEEE/ACM International Conference on Big Data Computing, Applications and Technologies*, pages 253–261.

[2] Gontijo, E. M., Delaiba, A. C., Mazina, E., Cabral, J. E., & Pinto, J. O. P. (2004). Gontijo, E. M., Delaiba, A. C., Mazina, E., Cabral, J. E., & Pinto, J. O. P. (2004, October). Fraud identification in electricity company customers using decision tree. *IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No. 04CH37583)*, Vol. 4.

[3] Interviewbubble (2020). Gráfico accuracy-recall-precision. Interviewbubble. https://interviewbubble.com/accuracy-recall-precision-f1-score/.

[4] Kaggle (2020). Fraud Detection in Electricity and Gas Consumption. Kaggle. https://www.kaggle.com/mrmorj/fraud-detection-in-electricity-and-gas-consumption/notebooks.

[5] Nizar, A. H., Dong, Z. Y., & Wang, Y. (2008). Power utility nontechnical loss analysis with extreme learning machine method. *IEEE Transactions on Power Systems*, 23:946–955.