Análisis de datos para generar esfuerzos coordinados ante el robo de identidad en el sector financiero de México

Adolfo Bravo Hernández [1], Fernando Vázquez Torres[2] y Eric Manuel Rosales Peña Alfaro[3]

[1,2,3] Instituto Politécnico Nacional, Unidad Profesional Interdisciplinaria de Ingeniería y Ciencias Sociales y Administrativas, Sección de Estudios de posgrado, Av. Té 950, Granjas México, 08400 Ciudad de México, CDMX

ofloda@gmail.com, fvazquezt@gmail.com y ericmrp@hotmail.com

Resumen. En este trabajo se busca analizar los datos proporcionados por el gobierno mexicano, acerca de fraudes relacionados con el robo de identidad, y encontrar relaciones que identifiquen posibles medios de combate a este delito. Se utiliza una metodología CRISP-DM por ser muy difundida y porque genera un análisis estructurado de los datos, con el apoyo de la herramienta RapidMiner que es líder en "Plataformas de Ciencia de Datos", examinando el reporte R27 proporcionado la CNBV. No se cuentan con resultados experimentales, sin embargo, se obtiene resultados que apoyan la búsqueda de estrategias para combatir el robo de identidad.

Palabras clave: Robo de Identidad, Reporte R27 CNBV, CRISP-DM.

Abstract. This paper pretends to analyze the data provided by the Mexican government, about frauds related to identity theft, and find relationships that identify possible ways of combating this crime. A CRISP-DM methodology is used because it is usually used and because it generates a structured analysis of the data, with the support of the RapidMiner tool that is a leader in "Data Science Platforms", examining the R27 report provided by the CNBV. There are no experimental results; however, results are obtained that support the search for strategies to combat identity theft.

Keywords: Identity theft, report R27, CRISP-DM.

1 Introducción

“En México, el delito de robo de identidad va en aumento día con día, según datos del Banco de México, nuestro país ocupa el octavo lugar a nivel mundial en este delito; en un 67% de los casos, el robo de identidad se da por la pérdida de documentos, 63% por el robo de carteras y portafolios, y 53% por información tomada directamente de una tarjeta bancaria. Comúnmente, el delito de robo de identidad se usa de manera ilegal para abrir cuentas de crédito, contratar líneas telefónicas, seguros de vida, realizar compras e incluso, en algunos casos, para el cobro de seguros de salud, vida y pensiones” [1].

Uno de los indicadores del robo de identidad es el reporte 27 generado por la Comisión Nacional Bancaria y de Valores. “El reporte Reclamaciones recaba información referente a las reclamaciones de operaciones monetarias realizadas por los clientes, agrupadas por productos y canales transaccionales de las Instituciones” [2].

El presente proyecto realizará la extracción de los datos de las reclamaciones usando este medio, y se obtendrá conocimiento bajo la luz de la metodología CRISP-DM, con lo que se puede fortalecer el combate a estos delitos.

2 Estado del arte

El gobierno mexicano difunde información a través de varios medios, como lo es la página de la CONDUSEF, donde señala acciones para: asegurar documentos y fuentes de datos que expongan información que identifique a la persona, protección de datos al no difundir información confidencial, o hacerlo a medios confiables. Además de que invita a la denuncia en caso de ser víctima de un delito por mal uso de la identidad de la persona.

Se cuenta con el Instituto Nacional de Transparencia, Acceso a la Información y Protección de Datos Personales (INAI), que es el organismo que debe garantizar el correcto uso de los datos de las personas, así como la protección de estos. "El INAI impulsa la Transparencia en tres vertientes:

Transparencia Reactiva: Se refiere a los procedimientos de acceso a la información y a los recursos de revisión que propician la entrega de la información solicitada.

Transparencia Activa: Es la publicación de información por parte de los sujetos obligados de acuerdo con lo establecido en el artículo 70 de la Ley General de Transparencia y Acceso a la Información Pública (Ley General).

Transparencia Proactiva: Es el conjunto de actividades que promueven la identificación, generación, publicación y difusión de información adicional o complementaria a la establecida con carácter obligatorio por la Ley General" [3].

Adicionalmente se está trabajando para fortalecer la seguridad, con medios biométricos para la identificación o verificación de la identidad de las personas. "La CNBV da un plazo del 30 de agosto de 2018 al 31 de marzo de 2020, para que los bancos conformen sus propias bases de datos biométricos, que estima ofrecerían ahorros a los bancos por 235 millones 200 mil pesos" [4].

3 Metodología

CRISP–DM (Cross Industry Standard Process for Data Mining), que se muestra en la figura 1, "fue creada por el grupo de empresas SPSS, NCR y Daimer Chrysler en el año 2000, es actualmente la guía de referencia más utilizada en el desarrollo de proyectos de Data Mining. Estructura el proceso en seis fases: Comprensión del negocio, Comprensión de los datos, Preparación de los datos, Modelado, Evaluación e Implantación" [5], la Fig. 1 muestra esta estructura. En 2007, fue considerada por la encuesta KDnuggets, como la más utilizada en la industria. Se puede consultar a detalle en la página de Internet http://crisp-dm.eu/home/crisp-dm-methodology/

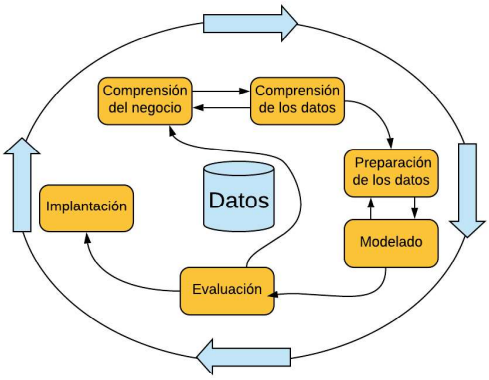

Fig. 1. Metodología CRISP-DM [6].

La metodología cuenta con tareas en cada una de sus fases. A continuación se describe de manera muy sucinta cada una de dichas fases, aplicada al caso de estudio.

3.1 Comprensión del negocio

Como parte de la esta fase se deben plantear los objetivos a perseguir, los cuales son:

- Revisar las tendencias de los últimos años de las reclamaciones con base en los datos recabados del reporte R27.
- Verificar el planteamiento de las tendencias hasta 2017 del robo de identidad. Con un fundamento cuantitativo e independiente al reporte de Condusef.

En los últimos meses se ha invertido muchísimo esfuerzo en tratar de contener y reducir los fraudes. Los bancos están en proceso de implementar mayores medidas de seguridad con el uso de biométricos. Con lo que se obtienen mayores controles, pero aún la cantidad de suplantaciones de identidad es muy severa.

3.2Comprensión de los datos

Al revisar la R27 y extraer datos se obtuvieron los siguientes puntos:

Hay cuatro atributos esenciales en la categorización de los casos:

Producto. Nos dice si se trata de una tarjeta de crédito, débito u algún instrumento al que se asoció el problema.

Canal: Nos dice si se realizó la transacción por un cajero automático, o por internet u algún otro medio.

Motivo. Clasifica si se trata de un cobro no reconocido, o si es un depósito no acreditado, con lo que podemos identificar posibles fraudes.

Hay muchos datos en blanco, esto debido a que hay tres tipos de monto principales: Procedente, improcedente y pendiente. Cuando se reporta con alguna cantidad estos montos, aparece un número, de lo contrario aparecen en blanco, la opción adecuada será cambiar estos por ceros, por ser la cantidad adecuada. Seguramente se reportan en blanco por eficiencia en el transporte de datos.

Después de 2010, las reclamaciones las reclamaciones de un banco desaparecen, esto debido a que se trata de IXE, el cual se fusionó con Banorte en ese año.

Hay una serie de reclamaciones con folio -1, las cuales no aportan información para nuestro caso pues no está asociada a un canal por el que se haya hecho el posible fraude o a un producto.

3.3 Preparación de los datos

Se revisaron los datos después de limpiarlos, se harán transformaciones para ser más adecuados al análisis.

- El período se separa en dos campos, pues al estar compuesto por año y mes, es más útil tenerlo separado.
- Se cambian las claves de banco por su nombre pues tienen mayor significado al revisarlo un especialista en el dominio.

Los datos inicialmente se construyeron con un modelo de estrella, todos estos se integrarán en una sola tabla por practicidad para su análisis.

3.4 Modelado

A partir de los datos se construyó una base de datos.

Fig. 2.

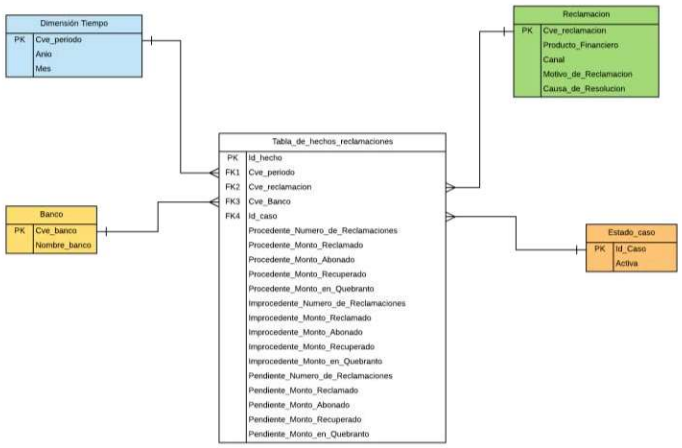

Fig. 2. Modelo de base de datos en Estrella.

Utilizando RapidMiner se creó el modelo para el análisis de los datos, como se muestra a continuación, en la Fig. 3. Se utilizó esta herramienta por ser gráfica y práctica, y fue nombrado líder en el "Magic Quadrant for Data Science and Machine Learning Platforms 2019" por sexto año consecutivo [7]:

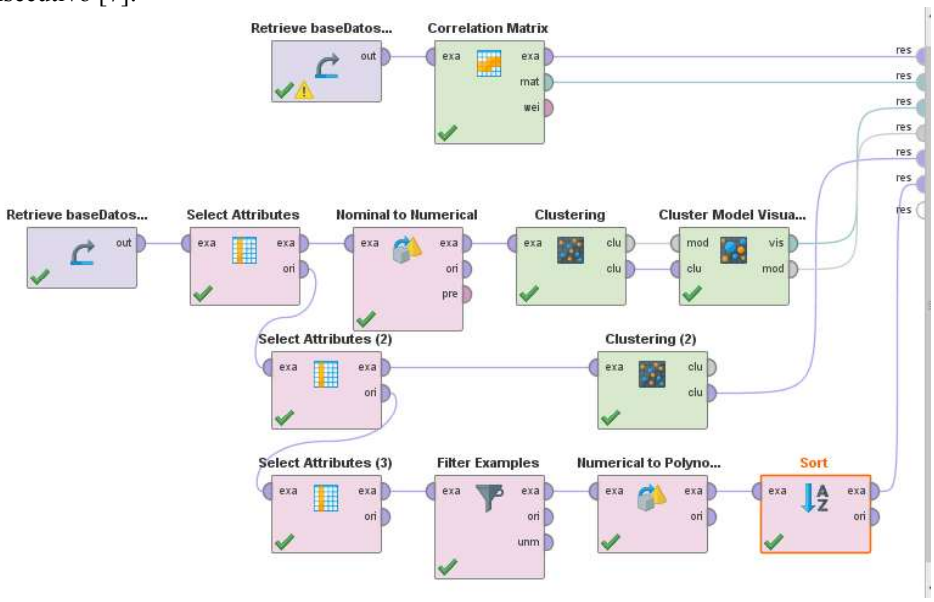

Fig. 3. Modelo en RapidMiner.

3.5 Evaluación

Los productos con promedios más alto en los montos reclamados son las tarjetas de crédito y de débito, que son por los que se generan transacciones de mayor valor monetario, lo que se muestra en la Fig. 4.

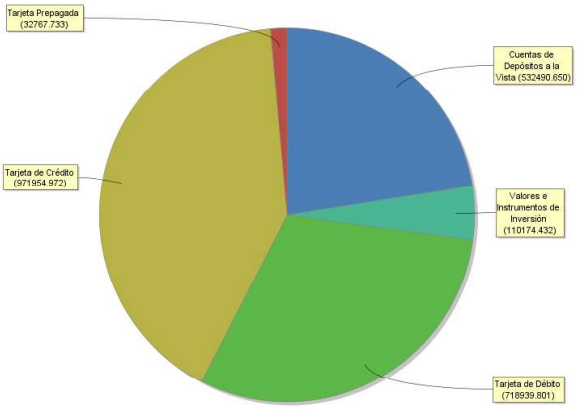

Fig. 4. Los productos con promedios más alto.

En las gráficas de caja de Producto en la dimensión año, observamos que el caso de "Valores e instrumentos de inversión", tiene una mayor concentración en años pasados, mientras que las "Cuentas de depósito a la vista" y "Tarjetas de crédito" tienen un auge en los años recientes. Mostrado en la Fig. 5.

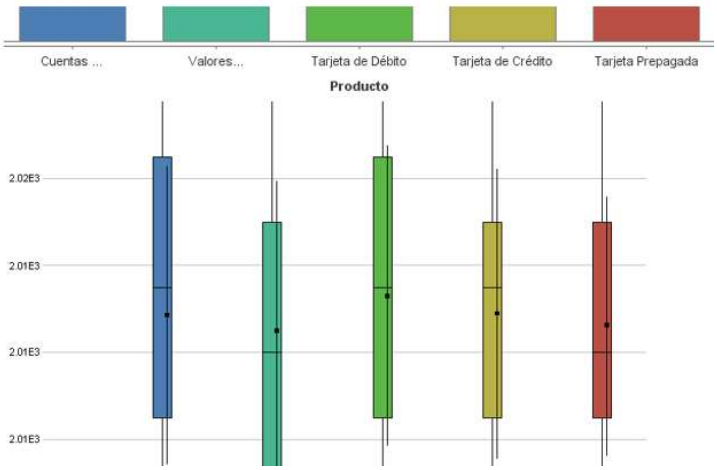

Fig. 5. Gráficas de caja de producto.

Al ejecutar algoritmos de agrupamiento de k-means sobre productos y canales, se obtienen conjuntos que podrían ser protegidos en conjunto, pues se relacionan estrechamente en su comportamiento. Por ejemplo los fraudes por pagos por celular y la banca por teléfono, probablemente tengan una forma de operar semejante por parte de los defraudadores. Cuando las iniciativas de seguridad sólo se enfocan a mejorar la tecnología del canal, pierden de vista trabajar en conjunto sobre los procesos utilizados por estos grupos. Se muestran dichos resultados en la Fig. 6.

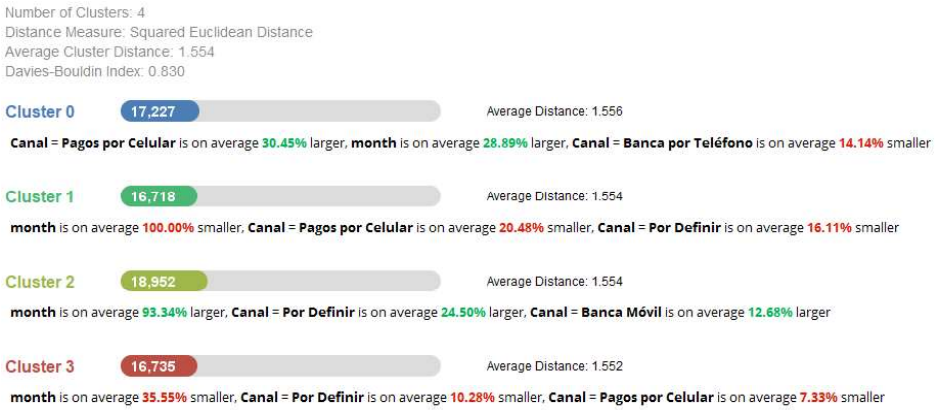

Fig. 6. Agrupación por k-means.

Los casos de robo de identidad tuvieron una baja en 2017, probablemente los procesos nuevos como biométricos tuvieron un efecto inhibidor sobre estos los fraudes. Lo cual se representa en la Fig. 7.

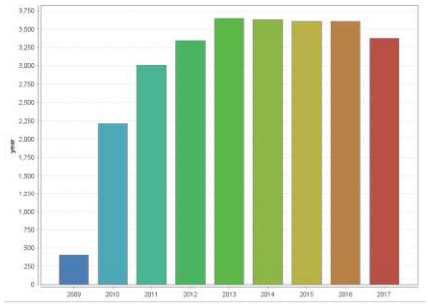

Fig. 7. Conteo de reportes del R27 por año.

3.6 Implantación

Para representar el proceso de manera muy sucinta, se creó el diagrama que se muestra en la Fig. 8.

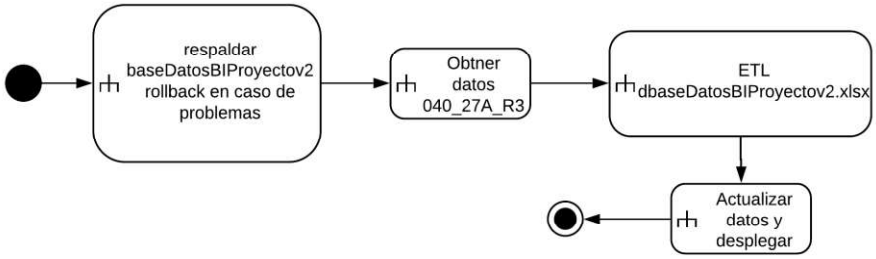

Fig. 8. Diagrama de actividades de la implantación.

4 Resultados experimentales

El trabajo de investigación desarrollado tiene resultados que dan recomendaciones teóricas, para una futura implementación que apoye la búsqueda de estrategias para combatir el robo de identidad.

5 Conclusiones y futuras líneas de investigación

Uno de los indicadores del robo de identidad es el reporte 27 generado por la CNBV. Otros medios de información son difundidos por el gobierno mexicano, principalmente a través de CONDUSEF.

El INAI cuenta con sus procesos de Transparencia Reactiva, Transparencia Activa y Transparencia Proactiva, para garantizar el buen uso de los datos de las personas.

Se está trabajando para fortalecer la seguridad. Los casos de robo de identidad tuvieron una baja en 2017, probablemente los nuevos procesos como lo son los biométricos tuvieron un efecto inhibidor sobre estos los fraudes. La CNBV da un plazo del 30 de agosto de 2018 al 31 de marzo de 2020, para que los bancos conformen sus propias bases de datos biométricos. Lo cual podría inhibir aún más los fraudes.

Los productos con promedios más alto en los montos reclamados son las tarjetas de crédito y de débito, "valores e instrumentos de inversión", tiene una mayor concentración en años pasados, mientras que "cuentas de depósito a la vista" y "Tarjetas de crédito" tienen un auge en los años recientes.

Al ejecutar algoritmos de agrupamiento de k-means sobre productos y canales, se obtienen conjuntos que podrían ser protegidos en conjunto, pues se relacionan estrechamente en su comportamiento. No se debe perder de vista las relaciones que existen entre las debilidades y fortalezas de grupos de canales y productos, para generar esfuerzos coordinados para el combate del robo de identidad.

Trabajo futuro: con mayor información aplicar algoritmos más detallados para encontrar relaciones precisas y con esto enfocar esfuerzos en el combate del fraude.

Agradecimientos. La presente investigación es resultado del proyecto de investigación del Instituto Politécnico Nacional que lleva por nombre: "Las TIC´s para el diseño de un Laboratorio de Inteligencia Empresarial que ayude al análisis de información de los procesos de las PyMes para mejorar la toma de decisiones", por el apoyo a esta investigación, con clave del proyecto: 20196346.

Bibliografía

[1] Amigón, E. (2015). Robo de identidad, un delito en aumento. (Condusef) Recuperado el 30 de 05 de 2018, de http://www.condusef.gob.mx/Revista/PDF-s/2015/186/robo.pdf

[2] Mexicano, G. (30 de 05 de 2018). Instructivo del reporte serie r27 reclamaciones - gob.mx. Obtenido de Tu gobierno en un sólo punto: https://www.google.com.mx/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=0ahUKEwjN0b7V5K7bAhVDXqwKHaaSBksQFggoMAA&url=https%3A%2F%2Fwww.gob.mx%2Fcms%2Fuploads%2Fattachment%2Ffile%2F121850%2FBM_BD_R27_A-2701_Reclamaciones_Periodicidad_201601-a

[3] Instituto Nacional de Transparencia, A. a. (01 de 05 de 2019). Transparencia Proactiva. Obtenido de http://inicio.ifai.org.mx/SitePages/Transparencia-Proactiva-acciones.aspx

[4] México, A. d. (16 de 07 de 2018). Síntesis informativa. Obtenido de https://www.abm.org.mx/sala-de-prensa/sintesis/historial/sintesis_2018_07_16.pdf

[5] Europe, S. V. (2015). CRISP-DM Methodology. Obtenido de http://crisp-dm.eu/home/crisp-dm-methodology/

[6] Moine, I., Haedo, D., & Gordillo, D. (s.f.). Estudio comparativo de metodologías para minería de datos. Obtenido de Servicio de Difusión de la Creación Intelectual es el Repositorio Institucional de la

Universidad Nacional de La Plata: http://sedici.unlp.edu.ar/bitstream/handle/10915/20034/Documento_completo.pdf?sequence=1&isAllowed=y

[7] RapidMiner, Inc. (2019). Gartner Magic Quadrant for Data Science Platforms. Obtenido de https://rapidminer.com/resource/gartner-magic-quadrant-data-science-platforms/