CASO DE ÉXITO

LA CIENCIA DE LOS DATOS

New Era Technology Le Puede Brindar Datos En Los Que Podrá Confiar Plenamente.

EL RETO

Una empresa líder en seguros necesitaba ayuda para identificar los posibles siniestros de auto-móviles que podían considerarse fraudulentos y que se encontraban sepultados bajo bases de datos con volúmenes de información del orden de cientos de Gigabytes, lo que operacionalmente se convertía en una tarea titánica e imposible de revisar manualmente, ya que se requeriría de tener múltiples regimientos de agentes para realizar esta labor.

PREOCUPACIONES FUNDAMENTALES

- **Información heterogénea**. Además de los Gigabytes de volumen, la información de pólizas se encontraba dispersa a través de bases de datos como Fox Pro, MySQL, Oracle en distintas versiones, así como también en archivos de Excel.

- **Información física.** Parte de los expedientes para cada siniestro consistía en un conjunto de fotografías, las cuales no todas estaban digitalizadas.

- **Información Incompleta**. La calidad de la información para ejecutar análisis implicaría un proceso de depuración y limpieza preliminar, que permitiera completar los huecos de información.

OBSTÁCULOS NOTABLES

- Era necesario tomar todas las fuentes de información, filtrar, clasificar y asignarles una probabilidad de manera tal que se obtuviera como resultado el top 15 de siniestros, que es el volumen que podía ser procesado manualmente por los agentes de antifraude.

- Manipular una base de datos de cientos de Gigabytes no es factible con ordenadores portátiles, así como no es deseable que la solución dependa de los recursos de unos pocos colaboradores.

- Las pólizas (contratos) se caracterizan por series de números y letras que cambian con los endosos (cambios en el contrato), o al renovarse, independientemente de que se trate de la misma persona. Por lo que no era posible analizar la frecuencia de siniestros asociados a un mismo individuo, si solo se contaba con dicha informa-

ción, por esta razón era necesario trabajar con otros datos que complementaran los modelos de análisis creados.

- No existía una base de datos histórica de grupos delictivos.

Technology Used

Power BI · axure Data Lake · SPSS · python R Language · SQL, SSIS

LO QUE LOGRAMOS

- Identificamos los mejores predictores que debían ser utilizados para la obtención del resultado esperado.
- Integramos de manera eficiente las bases de datos con:
  - Las variables necesarias para alimentar a los modelos matemáticos utilizados.
  - La ingeniería de algunos atributos.
- Creamos una base del histórico de fraudes usando la teoría de grafos.
- Entrenamos una red neuronal que clasificaba las imágenes de los diferentes siniestros e incorporaba esta información de manera oportuna.

SOLUCIÓN

Ciencia de Datos: Para identificar los grupos que comenten estos fraudes, a través de formalismo matemático de la teoría de grafos, se construyó inicialmente una "matriz de incidencia" basada en los números de siniestros.

Posteriormente, a partir de esta matriz de incidencia hay un algoritmo para generar la "matriz de adyacencia" y con esta es posible extraer lo que se denomina como cliques, que son grupos en el que todos se encuentran relacionados con todos los integrantes del mismo grupo. Debido a que existen cliques de diferentes órdenes y características, mediante la utilización de NetworkX de Python se logró realizar la identificación de los posibles individuos o grupos relacionados a un posible fraude.

Arquitectura de Datos: Se diseño un pipeline transportable en Java para automatizar todos los pasos necesarios para entregar diariamente el top 15 de posibles fraudes más recientes. Esta arquitectura incluía la extracción de datos integrando Oracle y MySQL, el procesamiento de estos datos, y la generación de atributos mediante Pandas de Python.

Machine Learning: Mediante Keras en Python, se logró construir y entrenar una red neuronal convolucional para identificar las características similares que pueden presentar los fraudes, e incorporar estos al conjunto de imágenes de los expedientes. De igual manera, se entrenó el modelo de Extreme Gradient Boosting (XGBoost de Python) para asociar diferentes probabilidades a los diferentes siniestros y una vez identificados estos, el proceso entregaba los 15 posibles candidatos de fraude para ser analizados por los agentes anti fraude.

Ingeniería de Datos: Se realizó una ingeniería de atributos, principalmente usando las fechas y las diferentes regiones, así como información de los asegurados, para facilitar el entrenamiento de los modelos.

Conclusión

A partir del análisis exhaustivo de los cambios e inconsistencias en los patrones de comportamiento de los clientes, la Ciencia de Datos, la IA, el machine learning y agentes inteligentes, se cuenta con la capacidad de predecir las transacciones fraudulentas en tiempo real. Al mismo tiempo, se reduce la tasa de positivos falsos, lo que redunda en la satisfacción de los asegurados, la protección de los ingresos y la reducción de los costos.

Si su organización puede beneficiarse de estos talentos, estamos a sus órdenes. New Era Technology puede poner más de dos décadas de experiencia a su servicio. No es por presumir, pero podemos mejorar su negocio, es lo que mejor hacemos.