Bancolombia

Prueba de conocimiento analítico

DiCAGI 2022

Gracias por su interés en participar en esta convocatoria.

Se buscan personas comprometidas, analíticas, con buena actitud de servicio y que se destaquen por realizar un trabajo oportuno y de calidad... además, personas con unos sólidos conocimientos en técnicas de modelación, estadística, matemáticas, minería de datos y bases de datos y como si fuera poco, capaces de definir adecuadamente un problema, generar las hipótesis, asumir supuestos, obtener conclusiones poderosas y comunicar los resultados de los modelos de una forma práctica y amigable para el usuario final, de tal forma que se asegure que los resultados de los modelos se utilicen en la estrategia del banco.

El propósito de la prueba es medir sus capacidades de análisis y desarrollo de modelos analíticos. La idea es que no le dedique más de 15 horas en total, incluyendo el tiempo para documentar lo que hizo.

Es posible usar cualquier herramienta que quiera (Python, R, SAS Guide/Miner, SPSS, etc.), y cualquier recurso del internet, pero no se permite consultar directamente con otras personas por ningún medio.

Puede realizar los supuestos que considere necesarios. No es necesario utilizar todos los conjuntos de datos o todas las variables. Esto depende de la forma en que usted aborde el problema. No hay una solución única. Inclusive puede darse el caso en que no se tenga un modelo viable.

Introducción

Las organizaciones estan constantemente expuestas a las pérdidas económicas por fraudes, ya sean transacciones o solicitudes fraudulentas, suplantación de identidad, entre otras. Esto conlleva a mejorar continuamente los procesos de gestión y prevención de dicho fraude. Para esto es necesario contar con estrategias efectivas que permitan identificar el fraude de manera oportuna y certera.

Este es un problema que conlleva a que las entidades tengan perdidas económicas, deterioro de imagen y desconfianza de los clientes. Con el boom de los datos y sacando el mayor provecho de estos mediante analitica, las entidades se han dado cuenta que es posible atacar estas actividades fraudulentas mediante modelos analíticos que permiten identificar actividades fraudulentas con gran efectividad y poco riesgo, trayendo grandes beneficios.

Contexto y descripción del problema

El **contexto** es el siguiente, en un proceso donde se atienden requerimientos particularmente devolución de cierta cantidad de dinero transada a través del canal D y por alguna falla del canal no se logra completar la transacción, este requerimiento es solicitado por el "reclamante" el cual estaba realizando alguna transacción de envió de dinero hacia el

"destinatario" por el canal D y pide que le reembolsen el dinero a la cuenta del "beneficiario", en este escenario es importante notar que cada uno de estos roles pueden ser la misma persona, personas diferentes, dos de ellos pueden ser los mismos, etc.

El **problema** es que, al se un proceso que devuelve dinero es propenso a que se presente fraude, pues las personas pueden simular, fingir que hubo una falla cuando nunca existió y solicitar la devolución de dinero.

A continuación, se describe información que puede ser relevante para el entendimiento del problema:

- Estos roles ocupados por personas pueden tener asociados productos (cuentas) y desde un punto de vista transaccional pueden ser "origen" o "destino" de transacciones, lo cual es equivalente a entrada y salida de dinero.
- El comportamiento transaccional de una persona se puede dar a través de diferentes canales A, B, C, D, E, F, entre otros.
- En el histórico de requerimientos las personas pueden estar varias veces, es decir, estas personas pueden estar implicadas en diferentes requerimientos fungiendo diferentes roles a lo largo de la historia.
- Se tienen logs del canal, mediante el cual se pueden identificar diferentes mensajes de error como transacción cancelada, host no responde, entre otros.

Para la **solución**, se necesita saber si es posible construir un modelo de ML que permita identificar cuando un requerimiento es un posible fraude y cuando no.

Definición de la población objetivo y variable respuesta

Como se mencionó anteriormente, lo que se busca es saber si un requerimiento es fraudulento o no. En el conjunto de entrenamiento ***entrenamiento_fraude.xlsx*** existe una columna ***fraude*** la cual es una variable dicotómica donde 1 es fraude y 0 es no fraude. El resto de las variables serian explicativas de la población. En el mismo archivo se encuentra las correspondiente metadata.

Instrucciones importantes

A continuación, se dan ciertas instrucciones importantes para la correcta ejecución de la prueba:

- ***entrenamiento_fraude.xlsx:*** Este conjunto de datos tiene características de los requerimientos y si este fue fraudulento o no.
- ***testeo_fraude.xlsx:*** Contiene exactamente las mismas columnas del archivo ***entrenamiento_fraude.xlsx***, exceptuando la columna ***fraude***.
- ***base_evaluada.xlsx***: Este archivo contienen los mismos radicados del archivo ***testeo_fraude.xlsx***. Usted debe poner aquí únicamente el radicado, la probabilidad de fraude y si es fraude o no, valores predichos por el modelo para cada uno de los casos de prueba. No cambie el nombre de este archivo ni los nombres de las columnas.

Entregables

A continuación, se describen los entregables:

- Se debe entregar un archivo ***base_evaluada.xlsx*** con la columna de radicado y ***fraude_pred***. La columna ***fraude_pred*** debe contener 1 si es clasificado como fraude y 0 en caso contrario. No aceptaremos valores nulos, NaNs, N/A, N/D, vacíos, o mensajes de texto como, por ejemplo: "datos incompletos". Por favor haga todo lo posible por conservar el formato del archivo (csv separado por comas, no otro carácter; el orden de las columnas; la línea de encabezado etc. El orden de las filas no es crítico).
- También nos debe entregar la implementación de su modelo (archivos de código con comentarios en caso de usar un lenguaje de programación convencional o el archivo de proyecto que incluya documentación, en caso de usar SAS Miner, Azure ML Studio, u otra herramienta parecida).
- Un archivo de texto (en .txt , .doc, .html, .rmd, .md, notebooks, .pdf) que contenga una descripción del proceso que siguió para generar el modelo (incluyendo exploración, transformaciones de variables, selección de variables, etc.) y luego generar sus estimaciones de fraude. Ahora, es posible que llegue a la conclusión de que no es posible desarrollar un buen modelo predictivo a partir de la información proporcionada o dada la calidad de esta. Si este es el caso, queremos evaluar el mejor modelo que pueda producir y también que nos dé una sustentación de esa conclusión.
- De manera opcional nos podría hacer saber qué otros datos o atributos adicionaría idealmente al conjunto de datos, para un modelo analítico más efectivo. Aquí, tenga en cuenta la factibilidad y el costo de obtener esos datos. Con esta información y sus propios análisis que utilidad extraería. Especifique para quien y de qué modo y el porqué es útil.
- Con el uso anterior en mente, diseñe un sistema de manera teórica que bosqueje una solución que permita hacer disponible los resultados de su modelo analítico y que sean fácilmente consumibles por el proceso, ya sean servicios externos, páginas web, servicio móvil, etc. No se tiene que desarrollar. Desarrollar una aplicación o sistema de información no da ningún punto extra y no será tenida en cuenta para la calificación total por lo que recomendamos no desarrollarla sino únicamente elaborar el bosquejo y describir la estrategia.

Evaluación

Para la evaluación se tendrán varios criterios, mencionados a continuación:

- Documentación del código y código limpio.
- Tratamiento de los datos, limpieza e imputación de la información, balanceo en caso de ser necesario.
- Descriptivo de los datos, con insights hacia la solución del problema.
- Gráficos descriptivos del problema y dicientes.
- Metodologías para transformación y selección de variables.

- Metodologías de selección de modelos de Machine Learning (Evaluación train-test, cross validation, tuneo de hiperparametros, overfitting, entre otros).
- Como es un problema de clasificación se evaluarán: método para selección de threshold, interpretación de los errores tipo I y II (¿Cuál nos interesa mas en el contexto del problema?) e interpretación de la curva ROC, AUC y matriz de confusión.
- La métrica para evaluar el modelo será el Accuracy entre la categoría de fraude real de cada requerimiento, que sólo nosotros conocemos, y el valor predicho por su modelo y consignado en el archivo ***base_evaluada.xlsx*** por medio de la columna ***fraude_pred.***