UNIVERSIDAD TORCUATO DI TELLA

Introducción a Aplicaciones de Ciencia de Datos en Finanzas
Maestría en Finanzas 2025

Prof. Pablo Roccatagliata
Universidad Torcuato di Tella
E-mail: proccatagliata@gmail.com

La evaluación del curso será mediante el siguiente trabajo domiciliario[1]. Pueden armar grupos de hasta 3 personas. La fecha de entrega será a elección grupal hasta el 17 de septiembre de 2025 a las 23.59hs. El viernes 15/08 a las 19.15pm tendremos horas de consultas sobre las consignas propuestas.

Motivación

Es importante que las compañías puedan detectar correctamente transacciones fraudulentas, en particular aquellas de alto valor. ¿Qué tipos de errores pueden cometerse al clasificar automáticamente a las transacciones según su riesgo estimado de ser fraudulentas? ¿Cómo podríamos calibrar estos costos en base a las características de cada cliente?

Contenido

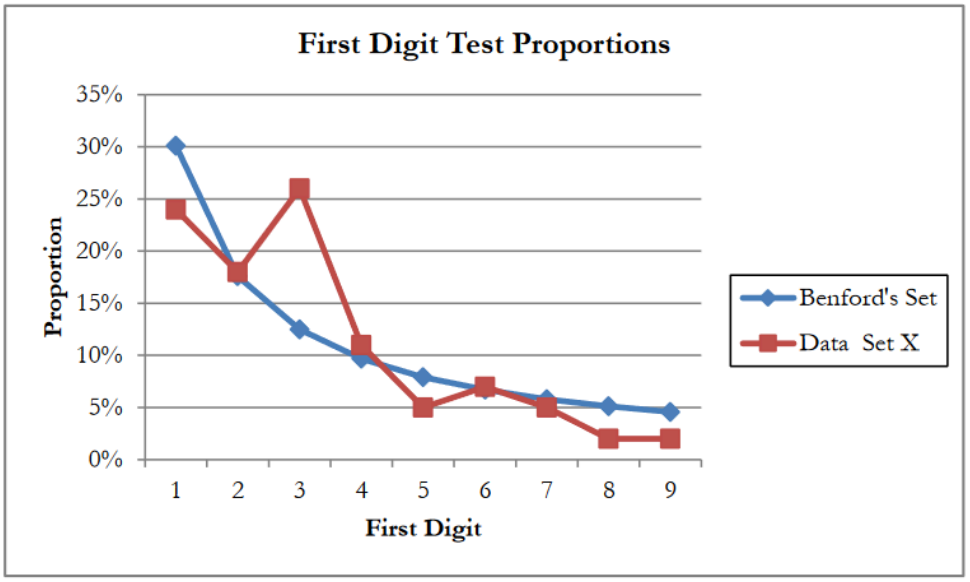
El dataset [2]contiene información de aproximadamente 1.5 millones de registros de transacciones de tarjeta de crédito divididos en Train.cvs y Test.csv. Recuerde usar el widget CSV File Import para importar los datos. La variable de target (isFraud) indica los casos etiquetados como fraude. El target tiene más de 99% de negativos por lo que se encuentra fuertemente desbalanceada. Además, el dataset contiene predictores cuantitativos y cualitativos. Los atributos que se encuentran en este conjunto de datos son el monto de la transacción, información de contexto, id del cliente, etc.

Una primera aproximación al estudio del fraude puede realizarse mediante los tests identificados por Mark Nigrini en base a la conocida ley de Benford (first digit test, second digit test, first two digits test, first three digits test, last two digits test). Una guía posible es el siguiente caso Using Benford´s Law to detect fraud de ACFE(Association of Certified Fraud Examiners)

---

[1] ***Es importante que todos los participantes de un equipo trabajen activamente en la solución del caso. Desde la dirección del MFIN nos solicitan que en caso de evaluación grupal exista una instancia de evaluación individual como una defensa por parte de algún integrante del grupo si existieran dudas al respecto.***

[2] Pueden descargarlo de Kaggle (requiere cuenta gratuita) o de este link.



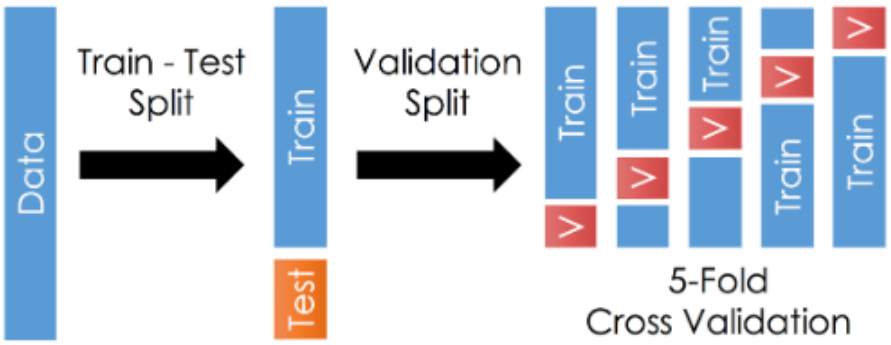
Fuente: Using Benford´s Law to detect fraud

Por otro lado, los costos asimétricos de cada error en la predicción (dejar pasar un fraude vs cancelar una transacción no fraudulenta) los pueden calibrar en base a las consecuencias de cada clasificación.

Objetivos

- Realizar un análisis exploratorio del dataset.
    - Algunos gráficos pueden ayudarlo a entender mejor el problema.
    - ¿Es necesario realizar una reducción de dimensionalidad para alguna de estas visualizaciones?
    - ¿Existen outliers?
    - ¿Podemos encontrar clusters (sin usar el target) en los datos? Este punto es relevante porque podría ocurrir que la variable de target haya identificado como fraude solamente en algunos casos. Esto podría ocurrir porque confirmar el fraude requiere un proceso costoso que probablemente sólo se realiza sobre algunas observaciones identificadas previamente como sospechosas.
    - ¿Hay correlaciones entre los features? Recuerde el VIF.
    - ¿Qué observaciones identificaría como sospechosas de fraude según los tests de dígitos en base a la ley de Benford?
- Ingeniería de features.
    - ¿Necesita corregir valores faltantes?
    - ¿Encuentra valores de los features que parecen errores en los datos?
    - ¿Qué preprocesamiento necesitan las variables para los algoritmos de ensamble (bagging y random forest) que va a utilizar?

    - ¿Podríamos probar diferentes estrategias para recodificar las variables categóricas como one hot encoding y target encoding?
- La variable de target (fraude/no fraude) está relativamente desbalanceada. ¿Qué cuidados debe tener al hacer el Split en train y test?
  ¿Qué estrategias podría usar para rebalancear el dataset?



- Entrenamos ahora un random forest.
    - ¿Cuáles son los hiper parámetros en random forest?
    - ¿Cuáles de esos hiper parámetros pueden generar sobreajuste?
    - Recuerde optimizar los hiper parámetros mediante validación cruzada sobre el train set.
- Introducir una estimación de los costos de cada tipo de error en la clasificación y plantee la matriz de costos de clasificación.
    - Puede ser útil revisar el artículo de Correa para este punto, pero tenga en mente la simplificación que planteamos en clase.
    - Dadas las probabilidades pronosticadas, de fraude o no fraude, que salen del random forest para cada observación, use Solver para optimizar el umbral.
    - Este es el umbral tal que si la probabilidad pronosticada de fraude es menor que ese nivel se considera como no fraude mientras que si la probabilidad pronosticada es mayor se predice fraude para esa observación.
    - La idea es encontrar cuál debería ser ese umbral para minimizar la suma de costos de los errores. Dado un umbral y unas probabilidades se arma un vector de predicciones, cada predicción puede ser correcta o no, en caso de ser incorrecta se suma al costo económico según el tipo de error del que se trate.
    - En base a esta regla de detección de fraude (probabilidades del random forest+umbral) …. ¿Cuál es el accuracy obtenido? ¿Por qué cree que pasa esto?
        - ¿Es este umbral otro hiper parámetro?
        - Conceptualmente… ¿Cómo podría incorporar su optimización en el esquema de validación cruzada utilizado?

- Podemos utilizar el widget de feaure importance, recuerde instalar el add-on “Explain”, para calcular la importancia de los predictores vía permutation importance.
- Hagamos un análisis gráfico mediante los lift and gain charts en base a las predicciones obtenidas, sobre test, para el modelo y umbral seleccionado. Diseñe 1 slide para comunicar los resultados del modelo en cuanto a los accionables respecto a la estrategia de retención.
- Compare sus resultados en cuanto a las observaciones identificadas como sospechosas de fraude en base a tests de dígitos versus machine learning cost sensitive.

Herramientas y materiales

Se recomienda que utilicen la herramienta de programación visual Orange, pero pueden utilizar el lenguaje de programación o herramienta gráfica que deseen. Sin embargo, les recomiendo usar Solver en Excel para la optimización del umbral de crédito.

Algunos links útiles

Using Benford´s Law to detect fraud
Mark Nigrini sobre ley de Benford
A novel cost-sensitive framework for customer churn predictive modeling
Alejandro Correa Bahnsen * , Djamila Aouada and Björn Ottersten
Validación de modelos en Orange.
Gain and lift charts
Interpretable Machine Learning

Financial Shenanigans (para leer más adelante quizá)

Entregables

Un breve informe analizando el problema que dé respuesta a los puntos mencionados en la sección de objetivos.