Inteligencia Artificial y Aprendizaje Automático

Actividad Semanas 5 y 6: Riesgo Crediticio

Maestría en Inteligencia Artificial Aplicada — Tecnológico de Monterrey
Prof. Luis Eduardo Falcón Morales

Nombre: _____ Matrícula: _____

Esta Tarea deberás resolverla de manera individual.

Esta actividad se complementa con el archivo "**MNA_IAyAA_semana_5_y_6_Actividad.ipynb**" que se encuentra en Canvas y donde deberás ir respondiendo los ejercicios. Este documento PDF simplemente complementa la información de los ejercicios. Al final deberás entregar la liga de tu GitHub donde se encuentra tu archivo de JupyterNotebook con las respuestas.

El asignar un crédito que conlleva un riesgo para el prestamista en caso de que el deudor no pague al final la cantidad asignada, o bien, al equivocarnos en negarle el préstamo a alguien que sí era confiable. Durante décadas se ha tratado de resolver dicho problema desde muchas áreas del conocimiento y en particular las técnicas de Aprendizaje Automático (Machine Learning) han brindado y siguen proporcionando nuevas formas de enfrentar este problema.

Existen pocas bases de datos abiertas bien documentadas sobre este problema. Una de ellas son los datos de la página de la UCI llamada "South_German_Credit" y sobre la cual se ha hecho mucha investigación en torno a la asignación de créditos. En esta actividad trabajarás con los datos del archivo "**SouthGermanCredit.asc**", el cual se encuentra dentro del archivo **south+german+credit.zip** que puedes descargar de la liga : https://archive.ics.uci.edu/dataset/522/south+german+credit

En ese mismo archivo zip se encuentra el archivo **codetable.txt**, donde puedes encontrar información detallada sobre el significado y tipo de cada variable, además de la que se encuentra en dicha página.

En la página de Kaggle puedes encontrar información adicional de estos datos:

https://www.kaggle.com/competitions/south-german-credit-prediction/overview

Pero sobre todo puedes apoyarte en el estudio publicado en el siguiente artículo de la IEEE:

https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9239944

Estos datos son una actualización de unos previos que se estuvieron usando durante décadas para investigación, pero como estaban en idioma alemán, no se habían percatado de varios errores que se habían generado al codificar las variables.

Parte I: Partición, análisis y pre-procesamiento de los datos.

1. Descarga los datos, los cuales nos llevan a un arreglo de 1000 registros y 21 variables. Cambia los títulos de las columnas al nombre en inglés (originalmente están en alemán). La información la puedes encontrar en cualquiera de las ligas dadas arriba.

2. Contrario a lo que sucede en analítica de datos, la clase de los buenos clientes están etiquetados con el valor de 1 y los malos clientes con el valor de 0. Como este no es el proceder dentro del área de ciencia de datos, aplica alguna transformación para invertir dichos valores, de manera que en lo sucesivo la clase negativa de los buenos clientes estén etiquetados con el valor de 0 y los malos clientes o clase positiva, con el valor de 1.
3. Realiza una partición de los datos en los conjuntos de entrenamiento, validación y prueba, del 70%, 15% y 15%, respectivamente.
4. Describe el significado de las 21 variables de acuerdo con la información que se encuentra en las ligas previas. Además, indica el tipo de variable que se tiene en cada caso, numérica o categórica. En particular, para las variables categóricas indica el número de niveles que tiene cada una.
5. Utilizando el conjunto de entrenamiento solamente, realiza un análisis descriptivo sobre el conjunto de datos e indica el tipo de transformaciones que aplicarás a las columnas. NOTA: Utilizaremos la clase Pipeline de Sklearn para aplicar dichas transformaciones en un siguiente ejercicio, por lo que aquí solamente tienes que mencionar las transformaciones que has decidido aplicar a cada columna.

Parte II: Modelos de aprendizaje automático con los conjuntos originales de la partición realizada.

6. Utiliza las clases Pipeline() y ColumnTransformer() de Sklearn para definir y conjuntar todas las transformaciones que hayas decidido aplicar a cada variable en el ejercicio anterior.
7. Como vamos a utilizar validación cruzada, concatena los conjuntos de entrenamiento y validación en un nuevo conjunto llamado trainval, que tendrá el mismo número de columnas, pero con el total de renglones la suma de ambos.
8. En este ejercicio deberás encontrar los hiperparámetros de cada modelo que consideres más adecuados para empezar a buscar los mejores modelos. Recuerda que debes buscar que no estén sobreentrenados o subentrenados los modelos.
9. De acuerdo a la información de la página de los datos antes de ser actualizados, se tiene una matriz de costo que pondera diferente los Falsos Positivos y los Falsos Negativos. Ver: https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data De acuerdo a esta información como una matriz de costo, contesta las siguientes preguntas:
    a. ¿Qué error se considera más costoso por parte del banco?
    b. ¿Cuál o cuáles serían entonces las métricas a considerar como más importantes?
    c. Investiga qué otras métricas se pudieran considerar, de la gran familia de métricas que existen.

       NOTA: Puedes consultar la siguiente página de Sklearn:

       https://scikit-learn.org/stable/modules/model_evaluation.html

10. Obtener un diagrama de caja y bigotes (boxplot) múltiple de todos los modelos, utilizando los resultados obtenidos con la métrica que consideraste más importante en el ejercicio anterior. Es

decir, en un mismo gráfico deben estar los siete diagramas de caja. Incluye tus conclusiones al respecto, en particular indica cuáles consideras son los tres mejores modelos obtenidos.

Parte III: Modelos con técnicas para clases no balanceadas

11. Selecciona una técnica de sobremuestreo, submuestreo o sobremuestreo+submuestreo para clases no balanceadas que consideres adecuada, en combinación con los tres mejores modelos de la Parte II, para entrenar y desplegar todas las métricas que se desplegaron en la Parte II. Puedes incluir aquellas métricas que también consideraste eran importantes en el ejercicio 9 anterior.

    Puedes consultar la página de Sklearn para la selección de alguno de los modelos (https://imbalanced-learn.org/stable/references/over_sampling.html) e igualmente te puedes apoyar en los resultados presentados en la investigación del artículo de la IEEE que se incluye al inicio de este documento para que busques aquellos que encontraron como mejores modelos de submuestreo y sobremuestreo.

Parte IV: Mejor modelo

12. Selecciona y justifica cuál consideras es el mejor modelo que has obtenido hasta ahora.
13. Con dicho mejor modelo y utilizando la técnica de validación cruzada, busca los mejores hiperparámetros de dicho modelo y despliega todas las métricas que se han estado desplegando. Verifica que tu modelo no esté sobreentrenado e indica cuáles son los mejores valores obtenidos de los hiperparámetros.
14. Con el mejor modelo y los mejores hiperparámetros encontrados en el ejercicio anterior, utiliza ahora (por primera vez en la actividad) el conjunto de prueba (test set) para:
    a. Obtener los valores de todas las métricas que se han estado desplegando.
    b. Obtener la matriz de confusión.
    c. Realiza un análisis de importancia de variables (feature importance) de este mejor modelo con el conjunto de prueba (test) e incluye tus conclusiones al respecto.
15. Escribe tus conclusiones finales de la actividad. En particular puedes comparar tus resultados con los que se muestran en el artículo de la IEEE.