Inteligencia Artificial y Aprendizaje Automático

Actividad Semanas 3 y 4:
Modelado, Balanceo e Importancia de Factores

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

Nombre: _____ Matrícula: _____

Esta Tarea se deberá resolver de manera individual y es parte de lo que estarás estudiando en las semanas 3 y 4 del curso. Deberás generar un archivo de Jupyter-Notebook con los análisis y comentarios que se te piden en los ejercicios.

La rotación de personal es uno de los problemas que afecta actualmente a muchas empresas y organizaciones, grandes o pequeñas y de cualquier tipo de negocio. En esta actividad usaremos una base de datos generada por IBM para estudiar cómo enfrentar dicho problema. Deberás descargar el archivo de la siguiente liga de Kaggle, la cual consta de 1470 registros y 35 columnas:

https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

PARTE 1: Análisis descriptivo y preprocesamiento de los datos:

1) Incluye una breve introducción sobre lo que se entiende por el problema de rotación de personal en las organizaciones (*employee attrition problem*).

2) Carga la base de datos y realiza los análisis necesarios para responder a las siguientes preguntas:
    a) ¿Cuántas y cuáles de las variables son numéricas?
        i. ¿Cuántas y cuáles de estas variables numéricas son de valor entero?
        ii. ¿Cuántas y cuáles de estas variables numéricas son de valor real (flotante)?
        iii. De existir el caso, ¿cuántas y cuáles de estas variables numéricas se pueden anular del problema? Justifica tu respuesta.
    b) ¿Cuántas y cuáles de las variables son categóricas?
        i. ¿Cuántas y cuáles de estas variables son binarias?
        ii. ¿Cuántas y cuáles de estas variables son nominales? Indica el total de niveles que tiene cada una de estas variables.
        iii. ¿Cuántas y cuáles de estas variables son ordinales? Indica el total de niveles que tiene cada una de estas variables.
        iv. De existir el caso, ¿cuántas y cuáles de estas variables categóricas se pueden anular del problema? Justifica tu respuesta.
    c) En esta base de datos y con base a la información dada, ¿existe alguna o algunas variables cuya clasificación en algún tipo de dato dependa del analista? ¿Cuáles y por qué?

d) De existir, elimina del problema todas las variables que consideraste que no están aportando información alguna.

3) Realiza una partición de los datos en Entrenamiento, Validación y Prueba, del 70%, 15% y 15%, respectivamente. Llama a dichos conjuntos Xtrain, Xval, Xtest, ytrain, yval, ytest, para los datos de entrada y de salida, respectivamente. Asegúrate que dicha partición conserve la estratificación de las clases de la variable "Attrition".
    a) Despliega la dimensión obtenida de los tres conjuntos: Entrenamiento, Validación y Prueba.

4) Usando solamente el conjunto de Entrenamiento, obtener los histogramas de las variables numéricas.
    a) Con base a estos gráficos ¿qué tipo de transformaciones sugieres llevar a cabo en dichas variables?
        i. Aplica las transformaciones que hayas determinado realizar, evitando el filtrado de información (*data-leakage*). A estas nuevas variables transformadas llamarlas XtrainT, XvalT y XtestT.

5) Aplica la transformación LabelEncoder() de sklearn a todas las variables binarias, evitando el filtrado de información.
    a) En particular, obtener la distribución de las clases de la variable de salida "Attrition". Con base a dicha distribución, ¿podemos considerar que tenemos un problema de datos no balanceados?

6) Realiza una inspección de las variables ordinales y determina qué transformaciones aplicarles, en caso de aplicar alguna. Justifica la decisión que tomes.

7) Obtener la matriz de correlación de los factores obtenidos hasta el momento. Debes incluir la variable "Attrition".
    a) Indica las correlaciones positivas "fuertes" entre pares de factores que encuentres dentro de la matriz.
    b) Indica las correlaciones negativas "fuertes" entre pares de factores que encuentres dentro de la matriz.

8) Aplica la transformación get_dummies() de Pandas a las variables nominales, evitando el filtrado de información y usando el argumento "drop_first" para generar "k-1" variables "dummies", de las "k" que cada variable.

9) Usa la instrucción XtrainT.head().T, para desplegar los primeros registros de tus datos de entrenamiento con todas las transformaciones realizadas hasta ahora.
NOTA: Hasta aquí, los nombres de tus variables deben seguir siendo los mismos: XtrainT, XvalT, XtestT, ytrainT, yvalT, ytestT.
    a) Despliega las dimensiones de los conjuntos XtrainT, XvalT y XtestT.

PARTE 2: Análisis exploratorio entre factores

10) Se pueden analizar una buena cantidad de factores para buscar relaciones entre los factores y la variable de salida "Attrition", la decisión de dejar un puesto. A a manera de ejemplos, veamos solo algunas de dichas relaciones gráficamente y que pueden aportar información al problema que enfrentamos.
    a) Las variables "Age" y "Attrition".
    b) Las variables "Department" y "Attrition".
    c) Las variables "Gender" y "Attrition".
    d) Incluye alguna otra relación o relaciones que consideres relevantes.
    e) Con base a los gráficos obtenidos incluye tus conclusiones al respecto.

NOTA: Puedes seleccionar el tipo de gráfico que consideres más adecuado. En particular te puedes apoyar en la librearía de seaborn. A manera de ejemplo puedes ver: https://seaborn.pydata.org/generated/seaborn.countplot.html

PARTE 3: Modelado

11) Utiliza los conjuntos de entrenamiento y validación para generar el mejor modelo no sobreentrenado de regresión logística usando la función LogisticRegression() de sklearn.
    a) Despliega los valores de la exactitud (accuracy) de los conjuntos de Entrenamiento y Validación.
    b) Utiliza los datos de validación para desplegar la matriz de confusión y el reporte dado por la función classification_report() de sklearn.
    c) Con base a estos resultados, ¿podemos decir que el modelo está subentrenado (underfitting)? ¿o sobreentrenado (overfitting)? Justifica tu respuesta.
    d) ¿Consideras que tenemos un problema desbalanceado? Justifica tu respuesta.
    e) Interpreta el valor numérico de la "precision" de la clase positiva.
    f) Interpreta el valor numérico del "recall" de la clase positiva.
    g) Con base al contexto de este problema, de rotación de personal, ¿cuál de las métricas, "precision" o "recall" consideras que es más importante disminuir su valor. Es decir, si no se puede disminuir el valor de ambos al mismo tiempo y debieras sacrificar uno de ellos, ¿cuál sería el que buscarías que fuera más cercano a cero? Justifica tu respuesta con base al contexto del problema.

12) Utiliza los conjuntos de entrenamiento y validación para generar el mejor modelo no sobreentrenado de los vecinos más cercanos kNN, usando la función KNeighborsClassifier() de sklearn.
    a) Despliega los valores de la exactitud (accuracy) de los conjuntos de Entrenamiento y Validación.
    b) Utiliza los datos de validación para desplegar la matriz de confusión y el reporte dado por la función classification_report() de sklearn.
    c) Con base a estos resultados, ¿podemos decir que el modelo está subentrenado (underfitting)? ¿o sobreentrenado (overfitting)? Justifica tu respuesta.
    d) Interpreta el valor numérico de la "precision" de la clase positiva.
    e) Interpreta el valor numérico del "recall" de la clase positiva.

f) Compara los resultados con los del modelo de Regresión Logística y escribe tus conclusiones.

PARTE 4: Balanceo de Clases

13) Utiliza el argumento "class_weight" de la función LogisticRegression() de sklearn y los valores de los hiperparámetros que consideres más adecuados para obtener un modelo no sobreentrenado.
    a) Despliega los valores de la exactitud (accuracy) de los conjuntos de Entrenamiento y Validación.
    b) Utiliza los datos de validación para desplegar la matriz de confusión y el reporte dado por la función classification_report() de sklearn.
    c) Compara los resultados con los modelos anteriores y escribe tus conclusiones.

14) Utiliza el método SMOTE de la librería "Imbalanced-learn" y los valores de los hiperparámetros que consideres más adecuados para obtener el mejor modelo posible.
    a) Despliega los valores de la exactitud (accuracy) de los conjuntos de Entrenamiento y Validación.
    b) Utiliza los datos de validación para desplegar la matriz de confusión y el reporte dado por la función classification_report() de sklearn.
    c) Compara los resultados con los modelos anteriores y escribe tus conclusiones.

15) Aplica alguno de los modelos combinados de sub y sobre entrenamiento y reporta los resultados del mejor modelo que hayas obtenido. Compáralo con los anteriores e incluye tus conclusiones.
    NOTA: https://imbalanced-learn.org/stable/references/combine.html

PARTE 5: La importancia de los factores

16) Con base al mejor modelo de regresión logística obtenido hasta ahora, utiliza la magnitud de los coeficientes como métrica para identificar aquellos factores que se consideran los más importantes al problema de rotación de personal.
    a) Generar un gráfico de barras de los coeficientes indicando el nombre de cada factor asociado a cada barra (bin).
    b) ¿Cuáles factores consideras que son los que influyen mayormente a que un empleado abandone su trabajo (attrition)?
    c) ¿Cuáles factores consideras que son los que influyen mayormente a que un empleado no abandone su trabajo (not attrition)?

17) Con base al mejor modelo que hayas obtenido hasta ahora ente regresión logística y el kNN, aplica la técnica de permutación de los factores con el método "permutation_importance()" de sklearn, y con la métrica "f1_weighted" del argumento "scoring" para identificar aquellos factores que se consideran los más importantes al problema de rotación de personal.

a) Generar un gráfico de barras de los coeficientes indicando el nombre de cada factor asociado a cada barra (bin).
b) ¿Cuáles factores consideras que son los que influyen mayormente a que un empleado abandone su trabajo (attrition)?
c) ¿Cuáles factores consideras que son los que influyen mayormente a que un empleado no abandone su trabajo (not attrition)?
d) Compara los resultados con el ejercicio anterior e incluye tus comentarios. En particular, comenta cuál método te da los mejores factores que tienen mayor impacto en el problema de rotación de personal.

18) Con base al mejor modelo que hayas obtenido hasta ahora ente regresión logística y el kNN y de los factores de mayor impacto que encontraste en el inciso anterior, utiliza la clase SelectFromModel de sklearn para reducir la cantidad de factores del problema y volver a entrenar el modelo con los datos de entrenamiento y validación, de manera que el desempeño con esta cantidad de datos reducida sea aproximadamente la obtenida previamente con todos los factores. Recuerda evitar el filtrado de información del conjunto de entrenamiento a los conjuntos de validación y de prueba.
    a) ¿A cuántos factores pudiste reducir el problema? Indica cuántos tenías y a cuántos se redujo, así como el porcentaje de reducción de factores.
    b) Despliega la matriz de confusión y el reporte dado por la función classification_report().

PARTE 6: Modelo final y conclusiones

19) Finalmente, con base a todos los resultados obtenidos hasta ahora, responde a los siguientes incisos para obtener el que consideres el mejor modelo para enfrentar el problema de rotación de personal a partir de los datos históricos iniciales.
    a) Forma un nuevo conjunto de Entrenamiento con los mejores conjuntos de entrenamiento y validación que hayas obtenido hasta ahora.
    b) Selecciona el mejor modelo de aprendizaje automático que hayas obtenido hasta ahora, entre regresión logística y kNN. Explica por qué lo consideras el mejor modelo.
    c) Entrena el modelo con el nuevo conjunto de entrenamiento aumentado y utiliza el conjunto de Prueba (Test) para obtener el desempeño final de tu mejor modelo. Para ello:
        i. Despliega la exactitud (accuracy) del conjunto de entrenamiento y del conjunto de Prueba para verificar que no esté sub o sobre entrenado.
        ii. Despliega la matriz de confusión y el reporte dado por la función classification_report() del conjunto de Prueba.
        iii. Incluye las conclusiones finales de la actividad. En particular interpreta y explica con base al problema de rotación de personal, los resultados obtenidos para las métricas “precision”, “recall” y “f1-score”.