EVALUACIÓN 2 - PARTE 2 - MODELOS DE CLASIFICACIÓN
INTELIGENCIA ARTIFICIAL

Docente Jazna Meza Hidalgo
Octubre 2024

CONTEXTO DEL NEGOCIO

La evaluación busca predecir la "attrition" (deserción o salida de empleados) en una empresa. La retención de empleados es un aspecto crucial para la sostenibilidad y eficiencia organizacional, ya que la pérdida de talento puede generar altos costos de reemplazo, disminución de la moral del equipo y pérdida de conocimiento especializado. Por lo tanto, contar con un modelo que permita identificar a los empleados con mayor probabilidad de abandonar la empresa ayudaría a tomar medidas preventivas, como ofrecer mejores incentivos, programas de desarrollo o mejoras en las condiciones laborales.

El conjunto de datos proporciona información sobre:

1. Employee ID: representa el identificador del empleado
2. Age: edad del empleado
3. Gender: género del empleado
4. Years at Company: cantidad de años en la empresa
5. Job Role: rol o cargo específico del empleado dentro de la organización
6. Monthly Income: ingreso mensual
7. Work-Life Balance: percepción del equilibrio entre vida personal y trabajo
8. Job Satisfaction: nivel de satisfacción del empleado con su trabajo
9. Performance Rating: evaluación del desempeño del empleado
10. Number of Promotions: cantidad de promociones que el empleado ha recibido.
11. Overtime: indica si el empleado trabaja horas extra
12. Distance from Home: distancia en kilómetros entre el hogar del empleado y la empresa.
13. Education Level: nivel educativo del empleado
14. Marital Status: estado civil del empleado
15. Number of Dependents: número de dependientes que tiene el empleado.
16. Job Level: nivel jerárquico del trabajo del empleado
17. Company Size: tamaño de la compañía
18. Company Tenure: años que el empleado ha estado en la organización, incluyendo tiempo total de servicio.

19. Remote Work: indica si el empleado trabaja de forma remota
20. Leadership Opportunities: indica si tiene oportunidades de liderazgo
21. Innovation Opportunities: indica si tiene oportunidades de innovación
22. Company Reputation: reputación de la compañía desde la perspectiva del empleado
23. Employee Recognition: indica si el empleado recibe reconocimiento en su trabajo
24. Attrition: indica si el empleado ha dejado la empresa

REQUERIMIENTOS DE LA EVALUACIÓN

1. Construcción de Modelos de línea base
   (a) Construir una línea base que contenga 4 modelos para predecir si un empleado dejará la empresa o no; los 4 modelos se deben obtener considerando los algoritmos revisados en clases (Logistic Regression, Random Forest, DecisionTreeClassifier, NaiveBayes) . Los modelos de esta línea base deben tener TODOS un accuracy superior a 0.7 y un ROC_AUC superior a 0.7.
   (b) Calcular todas las métricas para cada modelo de la línea base mejorada: *accuracy, precision, recall, f1-score, matriz de confusión, roc_auc*
   (c) **Nota.** El conjunto de datos debe dividirse en conjuntos de entrenamiento y prueba para evaluar el rendimiento del modelo. En caso de que esto se omita entonces se considera que el modelo ha sido entrenado de forma incorrecta.
2. Construcción de Modelos de línea base mejorada
   (a) Considerando la línea base anterior, es decir, manteniendo las mismas configuraciones de los modelos (hiper parámetros), aplicar alguna técnica de las revisadas en clases que le permita mejorar el rendimiento considerando el ROC_AUC que se requiere que sea superior a 0.78 para cada modelo.
   (b) Calcular todas las métricas para cada modelo de la línea base mejorada: *accuracy, precision, recall, f1-score, matriz de confusión, roc_auc*
3. Evaluación de Modelos
4. Seleccionar y justificar la elección de una métrica para seleccionar el mejor modelo
   (a) Evaluar el rendimiento del mejor modelo de la línea base mejorada utilizando métricas:
      i. accuracy,
      ii. precision,
      iii. recall,
      iv. f1-score,
      v. roc_auc
      vi. matriz de confusión
   (b) Los resultados anteriores se deben incluir en una tabla que debe tener el aspecto:

Table 1: Tabla Resultados

Modelo	**accuracy train**	**accuracy test**	**recall**	**precision**	**f1-score**	**roc-auc**
Modelo	9,999	9,999	9,99	99,99%	99,99%	99,99%

(c) Interpretar cada una de las métricas de la tabla anterior en el contexto del negocio.

(d) Interpretar la matriz de confusión del modelo seleccionado.

(e) Obtener una conclusión acerca de existencia/inexistencia de overfitting del modelo seleccionado.

(f) **Nota.** El modelo seleccionado debe cumplir, con al menos una métrica, con un rendimiento mínimo de 0.80.

5. Predicciones

   (a) Utilizar el mejor modelo seleccionado para realizar predicciones desde un archivo JSON que contenga valores para las variables independientes generando un archivo JSON que contiene el valor de las predicciones junto al valor de las variables independientes.

6. Respuesta a la pregunta al final del notebook

   (a) Responder con las justificaciones correspondientes a la pregunta al final del notebook base.

FORMATO DE LA ENTREGA

Debe entregar el notebook de base entregado completando cada una de las secciones contenidas en él. Todos los comentarios y las tablas de resultados que se requieren deben estar incluidas en el notebook.

Para el caso del archivo JSON que se pide con las entradas para realizar las predicciones, se debe incluir un enlace para cargar el archivo usando el comando *wget*.

PLAZOS DE ENTREGA

En la plataforma ADECCA, **LUNES 28 DE OCTUBRE** hasta las 18:00. Se aceptan entregas posteriores de acuerdo con los siguientes descuentos de la calificación final:

FECHA DE ENTREGA	**DESCUENTO A APLICAR**
28 octubre a las 18:01 horas hasta el 28 octubre a las 18:59 horas	Descuento = 2 puntos
28 octubre a las 19:00 horas hasta el 28 octubre a las 19:59 horas	Descuento = 3 puntos
Después de las 20:00 del 24 octubre	NO CONVIENE ENTREGAR

LISTA DE COTEJO

(a) Generales
   i. Respeta nombre del notebook entregado (620454-ModelosClasificación-Equipo-X.ipynb donde X corresponde a su número de equipo)
   ii. Indica información de los integrantes del equipo
   iii. Todas las interpretaciones y justificaciones aparecen como texto (evitando que aparezcan como comentarios del código)

(b) Modelos de línea base y cálculo de métricas considerando el rendimiento solicitado
   i. Construir y entrenar correctamente modelo usando LogisticRegression
   ii. Calcular todas las métricas del modelo.
   iii. Construir y entrenar correctamente modelo usando NaiveBayes
   iv. Calcular todas las métricas del modelo.
   v. Construir y entrenar correctamente modelo usando DecisionTreeClassifier.
   vi. Calcular todas las métricas del modelo.
   vii. Construir y entrenar correctamente modelo usando RandomForest.
   viii. Calcular todas las métricas del modelo.

(c) Modelos de línea base mejorada y cálculo de métricas considerando el rendimiento solicitado
   i. Construir y entrenar correctamente modelo usando LogisticRegression
   ii. Calcular todas las métricas del modelo.
   iii. Construir y entrenar correctamente modelo usando NaiveBayes
   iv. Calcular todas las métricas del modelo.
   v. Construir y entrenar correctamente modelo usando DecisionTreeClassifier.
   vi. Calcular todas las métricas del modelo.
   vii. Construir y entrenar correctamente modelo usando RandomForest.
   viii. Calcular todas las métricas del modelo.

(d) Mejor modelo usando una métrica seleccionada
   i. Justificar, correctamente, la métrica con la cual se selecciona el mejor modelo.
   ii. Seleccionar el mejor modelo desde la linea base mejorada considerando una de la métrica seleccionada.
   iii. Generar tabla de resultados del mejor modelo.
   iv. Interpretar, correctamente, en el negocio la métrica de accuracy en test.
   v. Interpretar, correctamente, en el negocio la métrica de f1-score.
   vi. Interpretar, correctamente, en el negocio la métrica de recall.
   vii. Interpretar, correctamente, en el negocio la métrica de precision.
   viii. Interpretar, correctamente, en el negocio la métrica de matriz de confusión.
   ix. Comentar, correctamente, acerca de la existencia/inexistencia de overfitting

(e) Predicciones
   i. Cargar, desde un archivo .json los valores de las variables independientes

ii. Realizar las predicciones de cada entrada contenida en el archivo cargado en el punto anterior.
iii. Generar un archivo .json que contenga las predicciones realizadas por el modelo

(f) Pregunta de cierre
i. Responde correctamente a la pregunta 1 planteada al final del notebook base
ii. Responde correctamente a la pregunta 2 planteada al final del notebook base

NOTA Cada ítem de la lista de cotejo se evalúa con presencia o ausencia.

CÁLCULO CALIFICACIÓN FINAL

La lista de cotejo define la calificación del proyecto. La calificación individual de cada estudiante va a depender de la respuesta a una pregunta asociada al desarrollo del proyecto:

1. Responde correctamente: calificación = calificación proyecto
2. Responde de forma parcial: calificación = 50% calificación proyecto
3. Responde de forma incorrecta: calificación = 40% calificación proyecto