TALLER 2

Objetivo

Evaluar las capacidades del estudiante para desarrollar modelos analíticos supervisados contemplando las etapas de la metodología ASUM-DM, como son el análisis y entendimiento de los datos, preparación de datos, creación de modelos, evaluación y análisis de resultados.

Planteamiento del problema

Una empresa del sector farmacéutico desea utilizar la información que posee de sus empleados para estimar el nivel de satisfacción laboral que puedan tener en algún momento y usarlo como insumo para detectar posibles fugas de personal. El CEO de la farmacéutica le ha pedido a su equipo de científicos de datos que construyan un modelo para tal fin y le proporcionen un conjunto de estrategias que permitan aumentar el nivel de satisfacción laboral de sus empleados y evitar futuras fugas de personal.

A continuación, se describe el diccionario de datos del dataset con el cual se debe conducir el análisis:

Field	Description
Age	Age of the employee
Attrition	employee attrition
BusinessTravel	how frequently an employee travels for business purpose
DailyRate	Daily wage of an employee
Department	Employee department
DistanceFromHome	Distance from home to office in KM's
Education	Qualification of employee (masked, higher is better)
EducationField	Stream of Education
EmployeeCount	EmployeeCount
EmployeeNumber	employee number
EnvironmentSatisfaction	Environment (higher is better)
Gender	Gender of employee
HourlyRate	employee hourly rate
JobInvolvement	Job involvement (higher is better)

JobLevel	level of Job (higher is more important)
JobRole	job role of an employee
JobSatisfaction	if employee is satisfied?
MaritalStatus	employee is married or not
MonthlyIncome	income of an employee
MonthlyRate	monthly rate of an employee
NumCompaniesWorked	number of companies worked for
Over18	age over 18
OverTime	employee works over time
PercentSalaryHike	salary hike
PerformanceRating	performance rate
RelationshipSatisfaction	Relationship satisfaction
StandardHours	per week standard work hours
StockOptionLevel	company stock option level
TotalWorkingYears	total working years
TrainingTimesLastYears	Training time
WorkLifeBalance	Work life balance
YearsAtCompany	total years at current company
YearsInCurrentRole	total years in current role
YearsSinceLastPromotion	years since last promotion
YearsWithCurrManager	Years worked under current manager

Actividades

A continuación, se describen más a fondo los hitos mínimos esperados por la empresa:

Limpieza y preparación de datos (15 pts)

Incluya como mínimo:

- Corrección de formatos.
- Búsqueda y corrección de valores atípicos, valores faltantes y duplicados. **No se deben eliminar registros.**

Análisis exploratorio de datos y selección de features (25 pts)

Debe consistir de lo siguiente:

- Análisis univariado de cada una de las columnas del dataset.
- Análisis bivariado de las relaciones más importantes.
- Utilice visualizaciones siempre que sea posible.

- Selección de las features más relevantes.

Responda las preguntas:

- ¿Qué variables impactan en mayor nivel la satisfacción de los empleados?
- ¿Qué variables parecen no ser relevantes para el análisis?

Construcción y selección del mejor modelo (30 pts)

Implemente 5 modelos de clasificación utilizando como variable objetivo "JobSatisfaction". Algunos algoritmos recomendados son: Regresión Logística, Árbol de Decisión, Random Forest. Para el proceso de experimentación debe ser evidente lo siguiente:

- Correcto manejo de los conjuntos de entrenamiento, validación y prueba.
- Implementación del proceso de entrenamiento y validación para la selección del mejor modelo. Se recomienda utilizar validación cruzada.
- Búsqueda de hiper-parámetros para encontrar el mejor modelo.
- Una vez seleccionado el mejor modelo, este deberá ser evaluado en el dataset de prueba para obtener las métricas de error finales.

Interpretación (10 pts)

- ¿Son Age y MonthlyIncome features significativas para la estimación del nivel de satisfacción del empleado?
- Elija las 3 features más importantes de su modelo e interprete cuales son las posibles reglas y/o efectos sobre la variable objetivo.

Evaluación (10 pts)

Con base en los resultados previos, concluya cuál es el modelo que debe ser usado para la estimación de la satisfacción del empleado. Exponga sus razones respondiendo las siguientes preguntas:

- ¿Qué métrica utilizó para escoger el mejor modelo? ¿Por qué?
- Adicional a las métricas de evaluación, ¿por qué lo considera un buen modelo?
- Defina al menos 3 estrategias concretas para presentarlas al CEO de la farmacéutica que permitan mejorar el nivel de satisfacción de los empleados y evitar la fuga de personal.

Estimación del nivel de satisfacción de nuevos empleados (10 pts)

En el dataset entregado existen 12 empleados para los cuales se desconoce su nivel de satisfacción (NewEmployee = 1). Estime el nivel de satisfacción de estos empleados, la calificación total del punto se tomará a partir de los éxitos en la estimación utilizando la siguiente tabla:

Respuestas correctas	pct del punto
[10-12]	100%
[7-9]	75%
[4-6]	50%
[1-3]	25%
[0]	0%

Bono (10 pts)

- Utilizar adecuadamente alguna técnica de reducción de dimensionalidad.
- Entrenar un cuarto modelo utilizando XGBoost. Debe incluirse dentro de la etapa de selección de modelo.

Mecanismo de entrega

- El taller debe ser desarrollado en grupos de 2 a 3 estudiantes.
- Debe ser entregado en los tiempos estipulados y solo a través de BloqueNeón. No se admiten entregas por otros medios como correo electrónico.
- El entregable debe consistir de un notebook subido a un repositorio público de GitHub, el cual debe incluir los outputs de la ejecución de cada bloque, pero también deberá poder ser ejecutado en su totalidad. En BloqueNeón se debe subir solo la URL del repositorio, no se admitirán commits posteriores a la fecha máxima de entrega.
- Dentro del notebook, haga uso de celdas de texto tipo markdown para exponer sus resultados y/o conclusiones de cada punto. También puede utilizar el archivo Readme del repositorio para concluir lo que considere necesario.

- Debe utilizar únicamente el dataset provisto en este taller.