Ficha Técnica: Proyecto 5 Profundización "Machine learning"

Título del Proyecto
Proyecto 5 Profundización

Objetivo
Desarrollar un modelo de machine learning supervisado para anticipar y gestionar la rotación del personal, fortaleciendo así la capacidad de nuestra empresa para retener y desarrollar el talento clave.

Equipo

Herramientas y Tecnologías
Listado de herramientas y tecnologías utilizadas en el proyecto:

1) Google Colab (Python).
2) Workspace Google (Presentaciones, Chat GPT y Documentos).
3) Videos Looms.
4) Git Hub.

Procesamiento y análisis

1) **Proceso de Análisis de Datos:**

1.1 Procesar y preparar base de datos

a) Conectar/importar datos a otras herramientas.
b) Identificar y manejar valores nulos.
c) Identificar y manejar valores duplicados.
d) Identificar y manejar datos fuera del alcance del análisis.
e) Identificar y manejar datos discrepantes en variables categóricas.
f) Identificar y manejar datos discrepantes en variables numéricas.
g) Comprobar y cambiar tipo de dato.
h) Dividir base en train y test.
i) Crear nuevas variables.

1.2 Análisis exploratorio

a) Agrupar datos según variables categóricas.
b) Visualizar las variables categóricas.
c) Aplicar medidas de tendencia central.
d) Visualizar distribución.
e) Aplicar medidas de dispersión.

1.3 Aplicar técnica de análisis

a) Machine learning (supervisado).

Resultados y Conclusiones

Durante el desarrollo del proyecto se utilizó el proceso de desarrollo del modelo de machine learning, desde la recopilación y preparación de los datos hasta la implementación y evaluación del modelo.

El análisis inicial de base de datos fue Dividir la base en train y test, luego se entrenaron 3 modelos de regresión (Logistica, XG-Bosst Y RandomForest), en dos oportunidades, para comparar la exactitud de cada algoritmo en la base de test. Cuyos resultados fueron los siguientes:

1. En el primera iteración: XGBoost y Random Forest tienen una precisión muy alta en este conjunto de datos, lo que sugiere que pueden estar sobreajustados a los datos de entrenamiento. Regresión Logística, siendo un modelo más simple, ofrece una precisión decente y es menos propenso a sobreajustar. Sin embargo, no captura tanta complejidad en los datos como XGBoost o Random Forest. Como la precisión fue más alta que en el conjunto de prueba, es necesario hacer una validación cruzada.
2. Segunda iteración : XGBoost es el mejor modelo hasta ahora (88.32%), y pequeños ajustes en los hiperparámetros podrían mejorar aún más su rendimiento. Logistic Regression es un modelo más sencillo, pero su desempeño es consistente y confiable. Random Forest es competitivo, pero no supera a XGBoost en este caso, aunque tiene un buen equilibrio y generaliza bien.

Seguir ajustando hiperparámetros en XGBoost y Random Forest para intentar mejorar la generalización. Considerar reducir ligeramente la complejidad del modelo para asegurarte de que no esté memorizando partes del conjunto de entrenamiento.

Limitaciones/Próximos Pasos

Sin observaciones.

Enlaces de interés

Proyecto 5 Profundización: Machine learning.ipynb

Proyecto 5 Profundización: Machine learning

Presentación Video

GitHub

Hitos pendientes para Proyecto 5 ML