TRAXIÓN

Detección de Conductores con Alta Probabilidad de Retirarse

Defensa de prueba técnica

Autora: M. en C. Gabriela Durán Meza

TRAXIÓN

Índice

TRAXIÓN

Objetivos y Requerimientos del Proyecto

Objetivo general: **Identificar conductores en riesgo de abandono mediante un modelo predictivo basado en datos sintéticos.**

- Objetivos:
  - Emular datos reales de quejas de conductores
  - Realizar análisis exploratorio y modelado predictivo
  - Documentar el proceso completo en notebooks ejecutables
- Requerimientos:
  - Generar base de datos sintética (.csv, mínimo 3000 registros)
  - Desarrollar notebooks para: generación de datos, EDA y modelado

TRAXIÓN

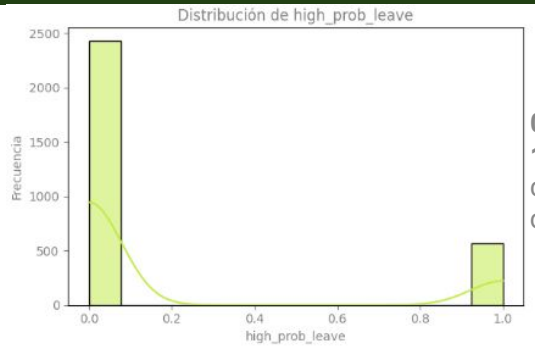
Generación de Datos Sintéticos

- Diseñando variables que generan **patrones realistas del sector**, lo que refuerza la credibilidad y utilidad del dataset sintético para simular un entorno laboral verosímil.
- Estableciendo la correlación lógica entre variables (por ejemplo, la relación entre edad y años de experiencia).
- Generando de textos mediante **GenAI**, implementando **langchain** y el modelo de **GPT-4o** para obtener la columna 'message'.
- Implementando **Feature Engineering** para obtener nuevas variables (e.g. derivando "years_experience" en función de la edad).

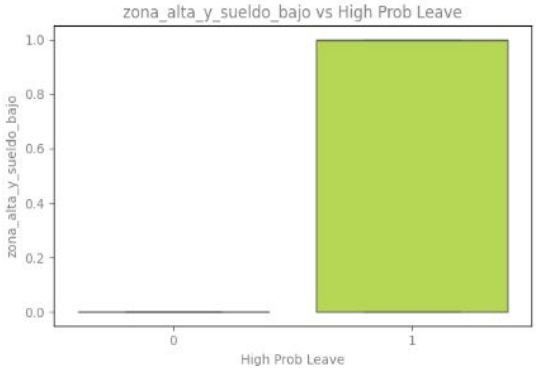
Pandas&Numpy							GenAI	Feature engineering	
driver_id	tag	age	salary	education	risk_zone	high_prob_leave	message	salario_por_km	zona_alta_y_sueldo_bajo
3571958	operaciones	48	7390	primaria	baja	0	Hola, me gustaría que revisaran mi salario, ya que con 10 años de experiencia y el costo de vida actual, 7390 pesos al mes no es suficiente.	449.326288	0
3880410	recursos humanos	30	5444	secundaria	alta	0	Con un salario de 5444 pesos y viajes de casi 3 horas para solo 6.6 km, es frustrante que no se valore más nuestro tiempo y esfuerzo.	826.633723	0

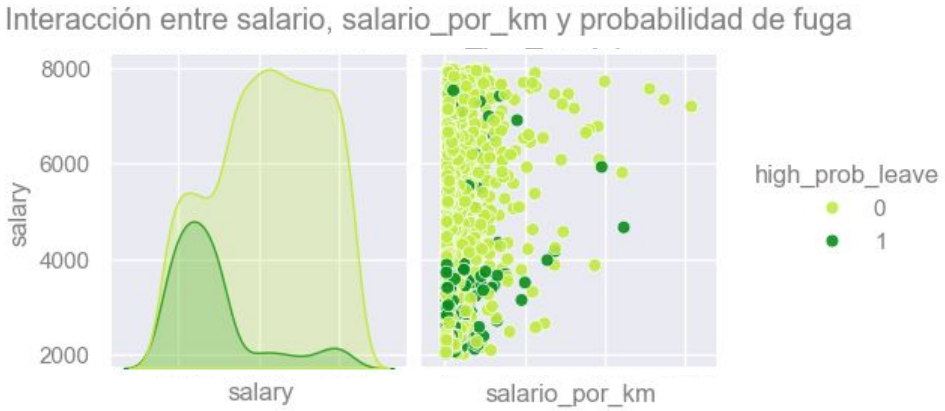
TRAXIÓN

Análisis Exploratorio de Datos (EDA)



0 ~83%
1 ~17%
clases desbalanceadas



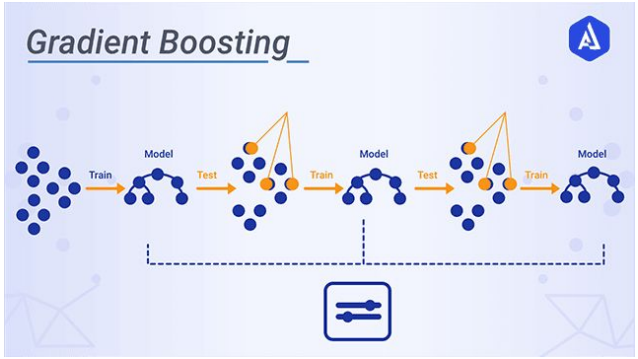


Principales Factores que Aumentan el Riesgo de Renuncia

- **Salario Insuficiente**
  - Conductores con menor salario presentan una mayor propensión a renunciar.
- **Zonas de Alto Riesgo**
  - Mayor probabilidad de churn en zonas clasificadas como “altas” por inseguridad o complejidad.
- **Salario vs Duración**
  - Un ratio bajo de pago respecto al tiempo/distancia de cada viaje está asociado a una intención de renuncia más elevada.
- **Combinación de Zona Alta y Sueldo Bajo**
  - Este factor conjunto intensifica el riesgo de renuncia.

TRAXIÓN

Modelo de Churn de Conductores



Alto desempeño en escenarios con datos heterogéneos y puede manejar mejor el desbalanceo.

Resistente a datos heterogéneos (numéricos, categóricos o textos codificados).

Modelos evaluados:

- Random Forest vs. Gradient Boosting

Metodología utilizada:

- GridSearchCV para optimización de hiperparámetros
- Validación cruzada para robustecer la evaluación

Métricas evaluadas:

- Precisión, recall, F1-score y matriz de confusión

TRAXIÓN

Resultados y evaluación del modelo

Gradient Boosting fue seleccionado por su desempeño global superior.

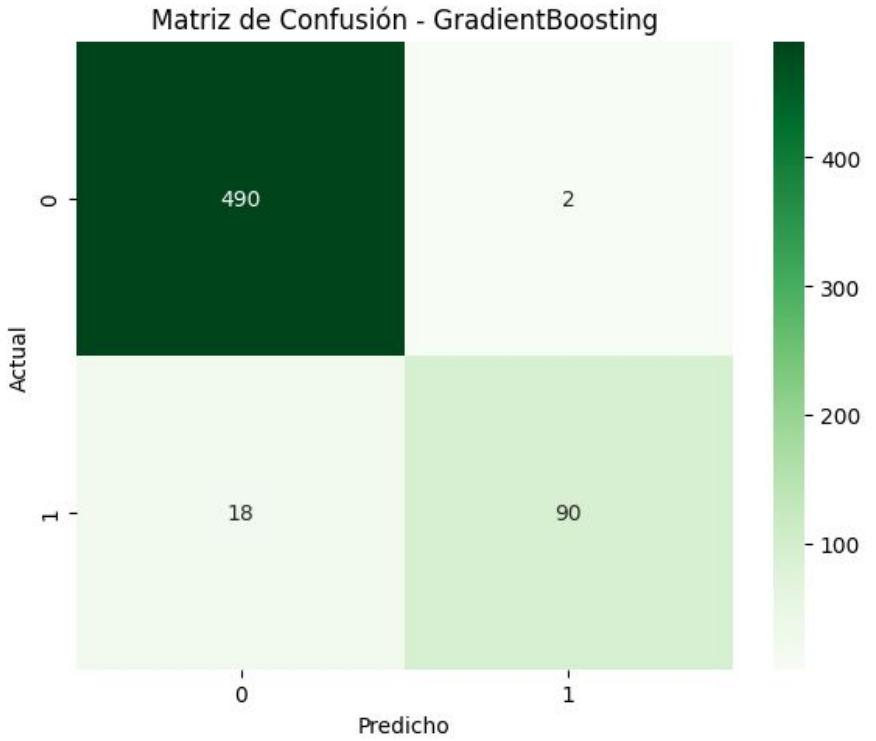
- Reporte de Clasificación :

- **Precisión general del 96.6%**, con excelente capacidad para distinguir entre conductores que se quedarán y quienes podrían abandonar la empresa.
- **Recall de 83.3% para clase de fuga (1)**: el modelo identifica correctamente a 8 de cada 10 conductores en riesgo de irse.

- Observaciones:

- El recall perfecto en la clase mayoritaria es consecuencia del desbalanceo.

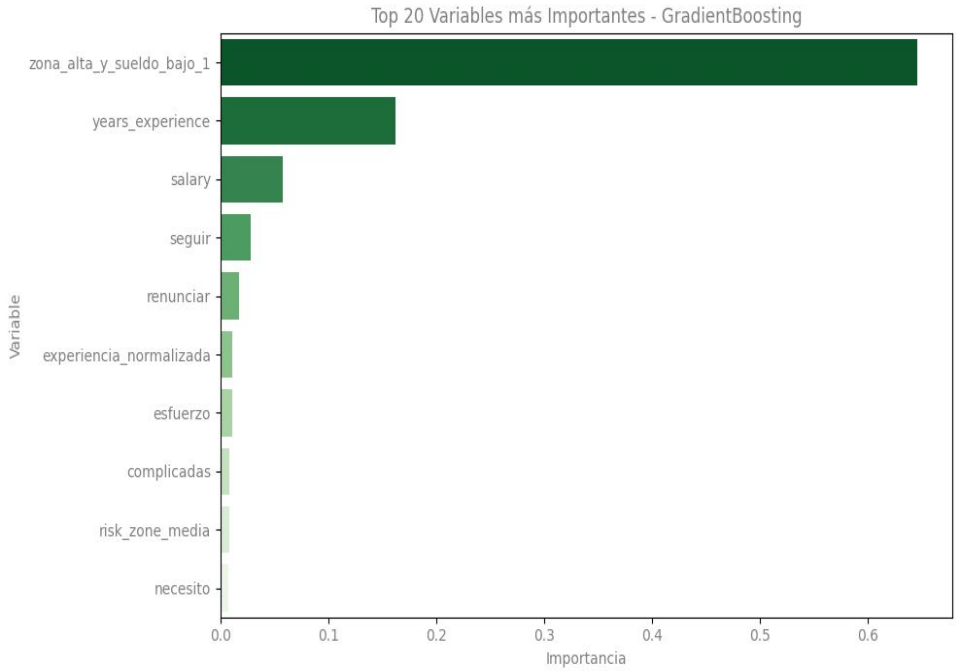
- Se identifican oportunidades para aplicar técnicas de balanceo (e.g., sobremuestreo, submuestreo, ajuste de pesos).



	precision	recall	f1-score	support
0	0.964567	0.995935	0.98	492
1	0.978261	0.833333	0.9	108
accuracy	0.966667	0.966667	0.966667	0.966667
macro avg	0.971414	0.914634	0.94	600
weighted avg	0.967032	0.966667	0.9656	600

TRAXIÓN

Resultados y evaluación del modelo



Insights del Modelo de Churn de Conductores

- **Zonas de riesgo con bajo salario:** Condiciones exigentes con poca compensación económica impulsan el abandono.
- **Conductores con alta experiencia:** A mayor antigüedad, mayor riesgo de churn.
- **Sueldo insuficiente:** Los bajos salarios siguen siendo un factor clave.
- **Mensajes que indican frustración:** *renunciar*, *seguir*, *esfuerzo* reflejan desgaste emocional y percepción de injusticia.

💡 **Recomendaciones**

- Ajustar salarios en zonas complicadas.
- Retener talento con experiencia.
- Usar el contenido de mensajes como sistema de alerta temprana.

TRAXIÓN

Consideraciones y limitaciones

- Aspectos Técnicos:
  - La implementación off-line del modelo de lenguaje que genera el texto realista basado en los datos sintéticos, no fue posible por cuestiones de incompatibilidad de versiones del ambiente local.
  - El desbalanceo es un área de mejora. Es necesario aplicar posibles técnicas (como sobremuestreo, submuestreo o ajuste de pesos de clases) para mejorar la detección de la clase minoritaria.
- Propuestas de Mejora
  :
  - Un ajuste en el modelo de lenguaje puede generar clases más balanceadas.
  - Definir la variable objetivo con base en datos históricos, usando datos reales.
  - El sistema puede funcionar como un "early warning system", donde la combinación de mensajes recientes con otros factores (bajo ingreso, alta experiencia, muchas quejas) activa alertas para el área de RH u operaciones.

TRAXIÓN

Conclusiones y próximos pasos

1. Valor del Mensaje en Lenguaje Natural

- El análisis de sentimiento y la extracción de palabras clave refuerzan las predicciones del modelo al detectar emociones como frustración, desmotivación o resignación.
- La solución propuesta, pese a estar basada en datos sintéticos, está estructurada para escalar fácilmente a datos reales una vez se integren, por la coherencia lógica y relacional entre variables.

2. Importancia del Feature Engineering

- Las variables derivadas como **`years_experience`** o **`income_per_km`** resultaron ser más predictivas que las variables base.

- Estas variables capturan relaciones no lineales clave, como la combinación de esfuerzo y recompensa, y mejoran la interpretabilidad del modelo.

3. Insights Geográficos

- Zonas geográficas específicas muestran patrones sistemáticos de insatisfacción, lo que sugiere que podría aplicarse una estrategia focalizada de intervención (e.g.. incentivos regionales o mejoras logísticas).