TRAXIÓN

Detección de Conductores con Alta Probabilidad de Retirarse

Defensa de prueba técnica
Autor: Mtra. Gabriela Durán Meza

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

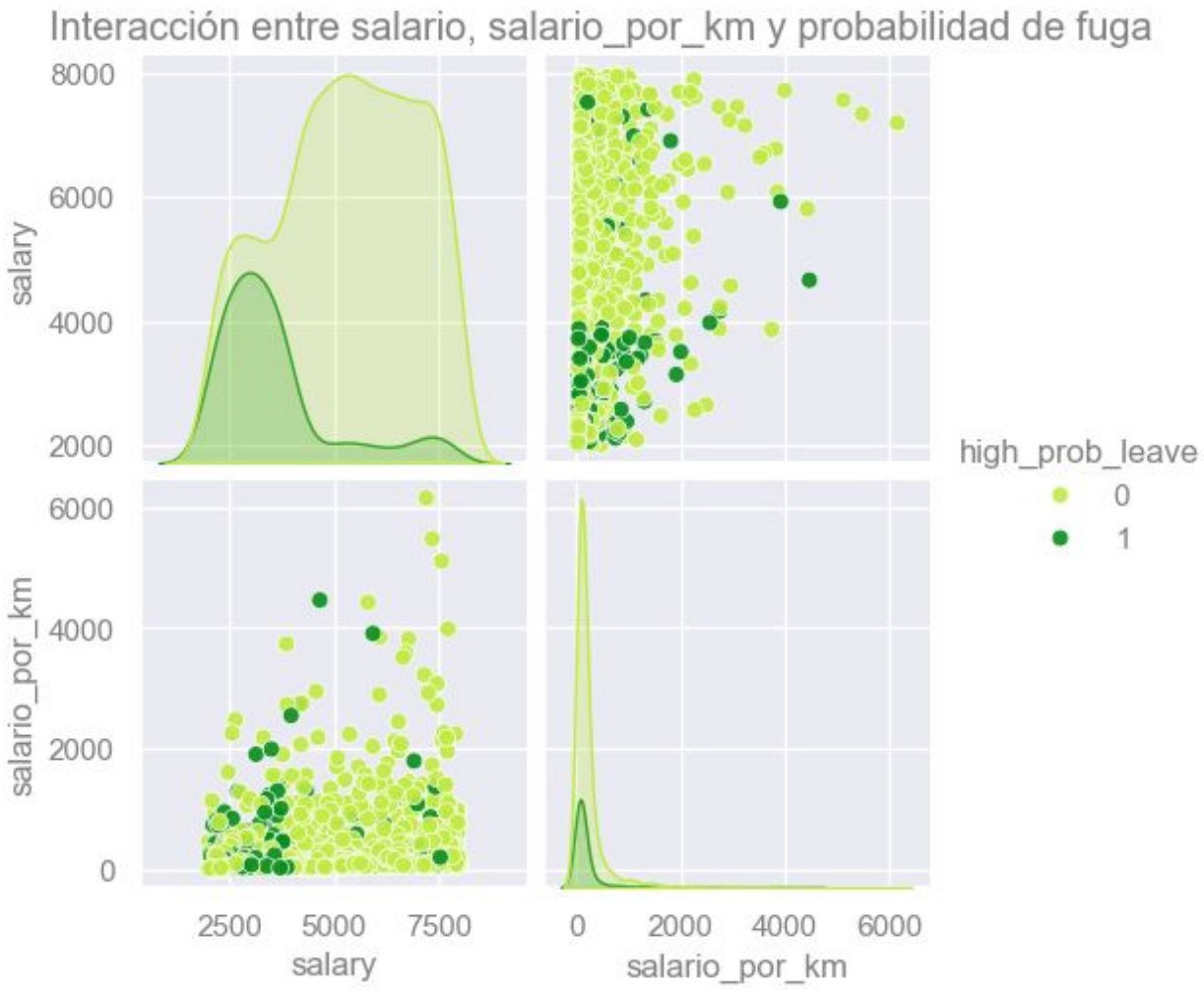
Generación de Datos Sintéticos

- 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 Dataframe						GenAI	Feature engineering			
driver_id	tag	age	salary	risk_zone	high_prob_le	message	salario_por_k	experiencia_	salario_por_d	is_solo_provi
3571958	operaciones	48	7390	baja	0	Hola, me gustaria que revis	449.326288	0.20833333	246.325122	0
3880410	recursos hum	30	5444	alta	0	Con un salario de 5444 pes	826.633723	0.1	34.2387784	0
3777069	operaciones	57	4433	baja	0	Me siento frustrado porque	83.717508	0.35087719	66.1631916	0
3842752	recursos hum	46	3082	alta	1	Me siento frustrado porque	66.7794414	0.19565217	128.411316	0
3831490	finanzas	22	5556	baja	0	Con el salario de 5556 pesc	746.663027	0.09090909	67.7552713	0

TRAXIÓN

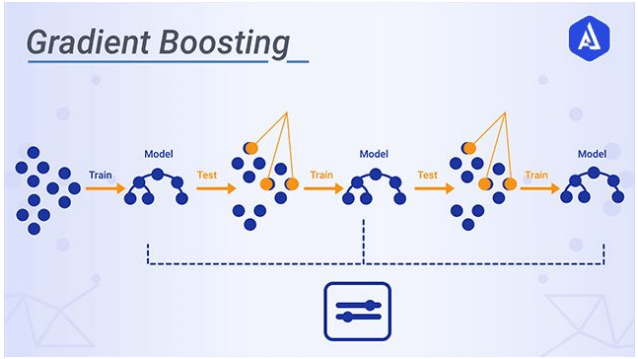
Análisis Exploratorio de Datos (EDA)



- La mayoría de los casos de alta probabilidad de fuga se concentra en zonas de bajo salario y bajo ingreso por kilómetro recorrido. No es sólo cuánto se gana sino cuánto se gana en proporción al esfuerzo.

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:
  - Implementación de 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.

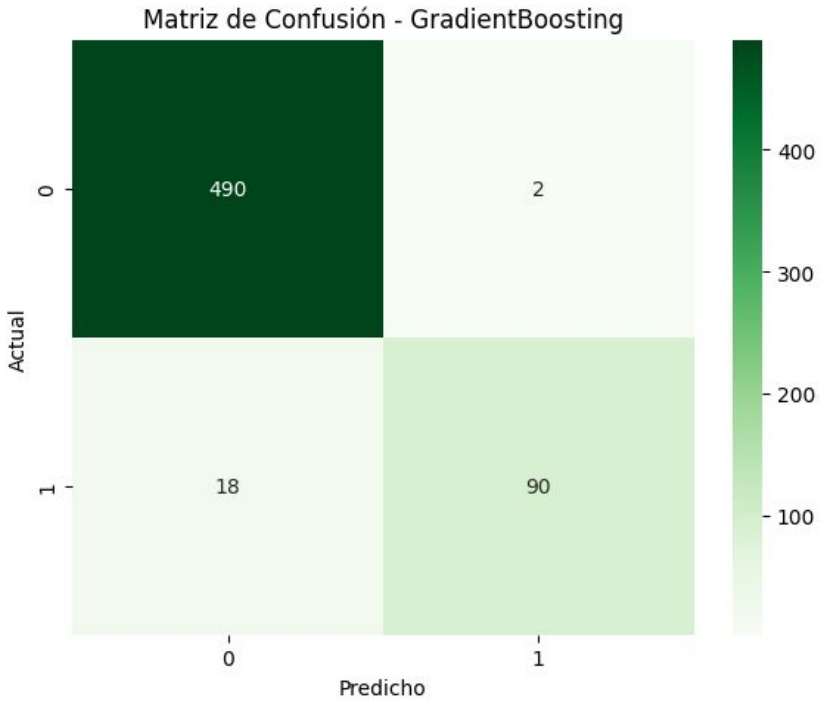
- Reporte de Clasificación :

- **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.

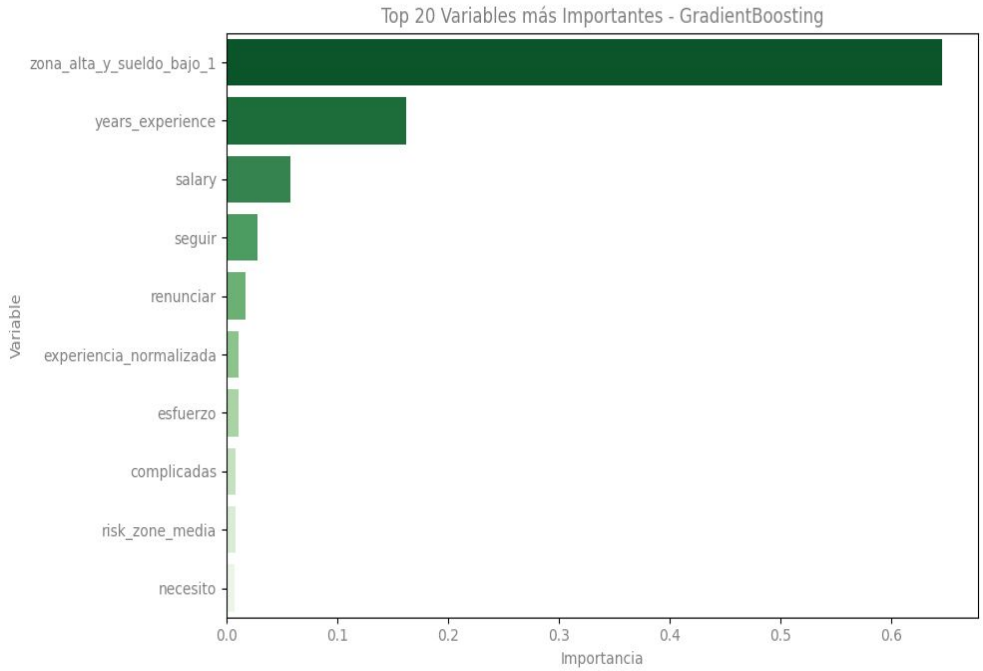
- Se identifican oportunidades para aplicar técnicas de balanceo (e.g., sobremuestreo, submuestreo, ajuste de pesos).



	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**
  Refleja que condiciones exigentes con poca compensación económica impulsan el abandono.
- **Conductores con alta experiencia**
  A mayor antigüedad, mayor riesgo de churn.
- **Sueldo insuficiente**
  Los mensajes y datos numéricos confirman que los bajos salarios siguen siendo un factor clave.
- **Mensajes que indican frustración**
  Palabras clave como *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.

TRAXIÓN

Conclusiones y próximos pasos

- Conclusiones:
  - El modelo de Gradient Boosting ofrece un desempeño robusto en un entorno de datos desbalanceados
  - El proceso de selección y validación (con GridSearchCV) respalda la elección técnica
- Próximos Pasos:
  - Implementar técnicas de balanceo para mejorar el rendimiento en la clase minoritaria
  - Continuar evaluando y ajustando el modelo a medida que se disponga de nuevos datos o feedback
  - Preparar documentación para la ejecución offline y revisión en equipo