

Predicción de Churn en Telecomunicaciones

1. Problema de Negocio (ficticio)

Una empresa de telecomunicaciones necesita identificar qué clientes tienen mayor probabilidad de abandonar sus servicios (churn) para implementar estrategias de retención proactivas. El objetivo es:

1. **Identificar factores de riesgo** que llevan al abandono de clientes
2. **Cuantificar el impacto individual** de cada variable en la decisión de churn
3. **Priorizar acciones de retención** basadas en probabilidades de abandono

2. Desafíos Técnicos

- Dataset con 85+ variables de comportamiento del cliente
- Propuesta de un entorno de trabajo: Docker + Airflow
- Construcción de un DAG que represente el ciclo de vida típico de construcción y aplicación de un modelo de analítica avanzada
- Propuestas de mejoras: uso de PostgreSQL, MLFlow, Grafana y Prometheus

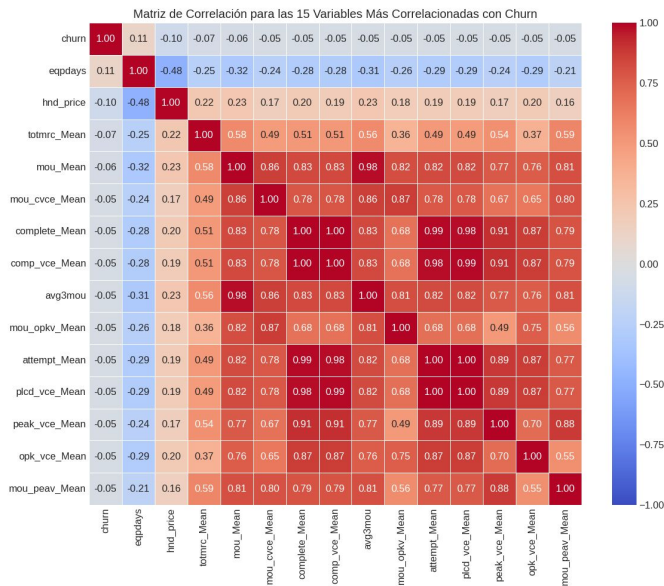
3. Desarrollo de la prueba

1. EDA sobre el dataset proporcionado:

- Exploración básica de los datos (estadísticas, nulos, distribución de la variable objetivo)
- Análisis univariado de variables numéricas y categóricas
- Análisis bivariado y correlaciones
- Análisis de características importantes para el churn
- Análisis de segmentación entre algunas variables clave

Link al notebook:

https://github.com/Tessie295/dag_system/tree/main



Conclusiones generales EDA

1. **Abandono multifactorial:** El churn parece ser un fenómeno complejo determinado por múltiples factores, no por variables individuales aisladas.

2. **Importancia de la antigüedad del equipo:** Los días de antigüedad del equipo ("**eqpdays**") tienen la correlación positiva más alta con el abandono, sugiriendo que los clientes con dispositivos más antiguos podrían ser un segmento a priorizar para retención.

3. **Valor del dispositivo como factor de retención:** Los clientes con dispositivos más caros ("**hnd_price**") tienden a ser más leales, posiblemente debido a un mayor compromiso financiero o mejor experiencia de usuario.

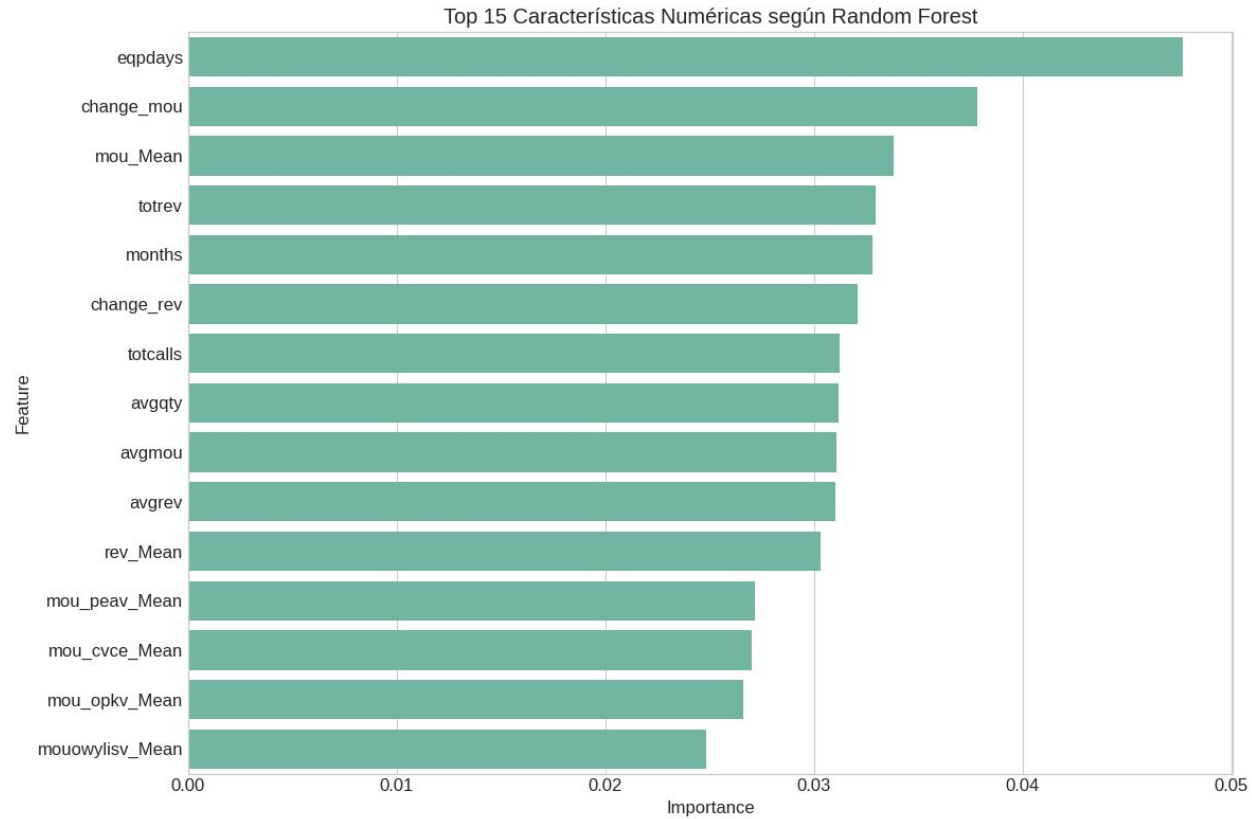
4. **Segmentos de alto riesgo:** Aunque no hay una separación clara, podríamos identificar segmentos de alto riesgo como:

- Clientes con equipos antiguos
- Clientes con dispositivos de bajo costo
- Algunos casos atípicos con patrones de uso inusuales

3. Desarrollo de la prueba

2. Selección de características

- a. Análisis de outliers
- b. Selección de características:
 - i. Preparación de datos para análisis de correlación
 - ii. Eliminación de características numéricas altamente correlacionadas
 - iii. Selección basada en ANOVA F-value (solo variables numéricas)
 - iv. Selección basada en Información Mutua (solo variables numéricas)
 - v. Selección basada en Random Forest (solo variables numéricas)
 - vi. Análisis de características categóricas y selección con Random Forest
 - vii. Lista final de características numéricas y categóricas seleccionadas



3. Desarrollo de la prueba

3. Preparación de los datos para el modelado

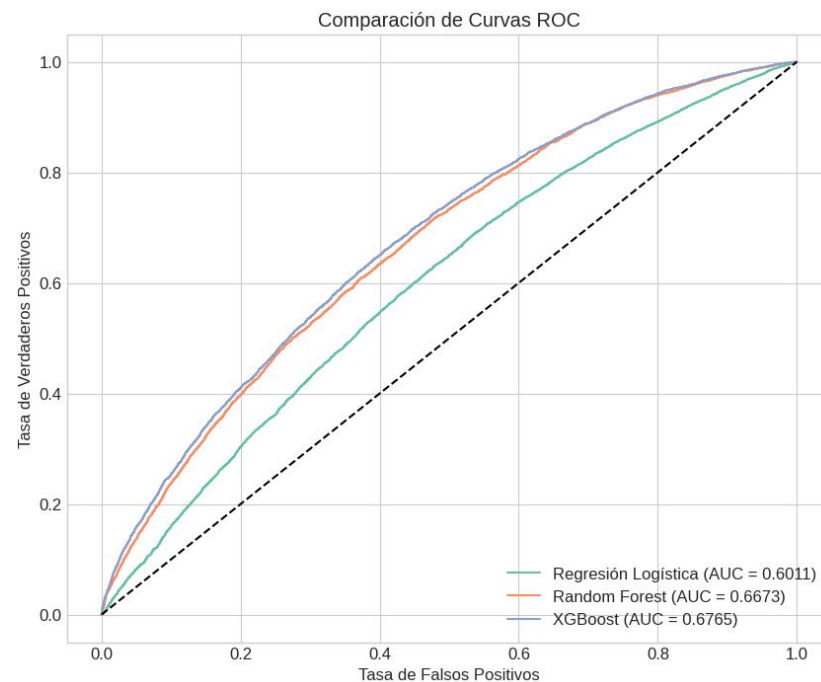
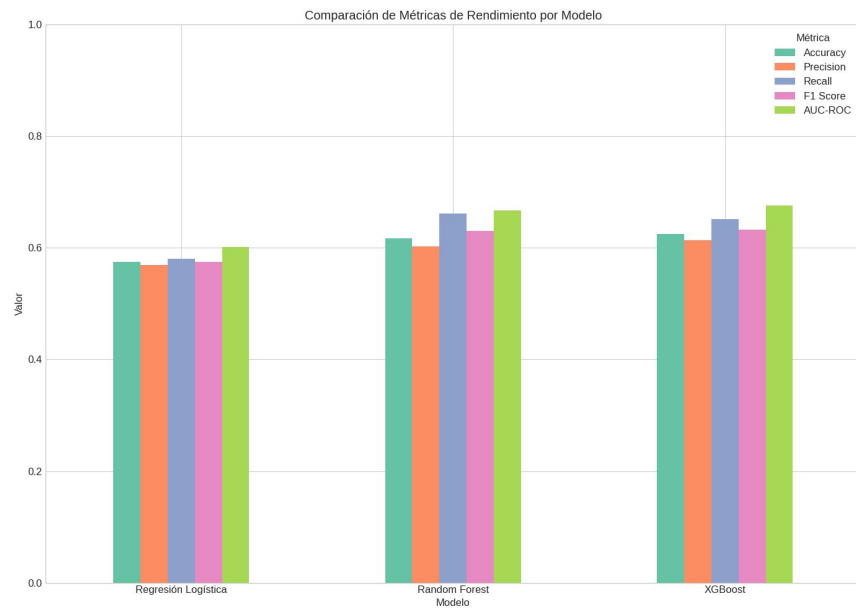
- a. Manejo de valores faltantes (mediana numéricas, moda categóricas)
- b. Manejo de outliers (percentiles capping)
- c. Codificación de variables categóricas
- d. Escalado de características (teniendo en cuenta la selección anterior)
- e. División en conjuntos de entrenamiento y prueba

3. Desarrollo de la prueba

4. Modelado

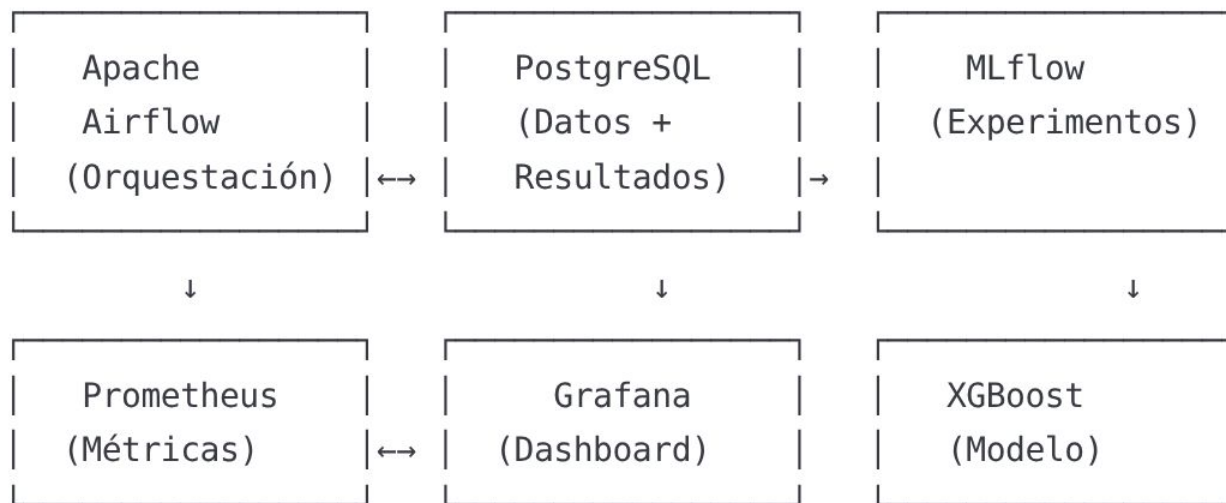
- a. Modelo de Regresión Logística
- b. Modelo de Random Forest
 - i. Entrenamiento con validación cruzada para ver rendimiento
 - ii. Optimización de parámetros con GridSearchCV
- c. Modelo de XGBoost
 - i. Entrenamiento con validación cruzada para ver rendimiento
 - ii. Optimización de parámetros con RandomizedSearchCV

Comparación de modelos



3. Desarrollo de la prueba

5. Arquitectura/pipeline



Creación de los contenedores de servicios en Docker

CONTAINER ID	IMAGE	COMMAND	CREATED	STATUS	PORTS
2a29b94588b3	dag_telecom-mlflow	"bash -c '\n mkdir -..."	3 minutes ago	Up 3 minutes	0.0.0.0:5000->5000/tcp, [::]:5000->5000/tc
p_dag_telecom-mlflow-1					
2f31da4561d7	grafana/grafana:9.5.1	"/run.sh"	3 minutes ago	Up 3 minutes	0.0.0.0:3000->3000/tcp, [::]:3000->3000/tc
p_dag_telecom-grafana-1					
62df82758528	apache/airflow:2.7.1	"/usr/bin/dumb-init ..."	3 minutes ago	Up 3 minutes (healthy)	8080/tcp
p_dag_telecom-airflow-triggerer-1					
fbfcb4f70f4f	apache/airflow:2.7.1	"/usr/bin/dumb-init ..."	3 minutes ago	Up 3 minutes (healthy)	0.0.0.0:8080->8080/tcp, [::]:8080->8080/tc
p_dag_telecom-airflow-webserver-1					
c28f1781f2e6	apache/airflow:2.7.1	"/usr/bin/dumb-init ..."	3 minutes ago	Up 3 minutes (healthy)	8080/tcp
p_dag_telecom-airflow-worker-1					
48e733cf55bc	postgres:13	"docker-entrypoint.s..."	3 minutes ago	Up 3 minutes (healthy)	0.0.0.0:5435->5432/tcp, [::]:5435->5432/tc
p_dag_telecom-postgres-ml-1					
e3deefea00fd	apache/airflow:2.7.1	"/usr/bin/dumb-init ..."	3 minutes ago	Up 3 minutes (healthy)	8080/tcp
p_dag_telecom-airflow-scheduler-1					
b79b00baffca	prom/prometheus:v2.43.0	"/bin/prometheus --..."	3 minutes ago	Up 3 minutes (healthy)	9090/tcp
p_dag_telecom-prometheus-1					
4cc41be4ec6a	postgres:13	"docker-entrypoint.s..."	3 minutes ago	Up 3 minutes (healthy)	5432/tcp
p_dag_telecom-postgres-1					
b46af602aa35	redis:latest	"docker-entrypoint.s..."	3 minutes ago	Up 3 minutes (healthy)	6379/tcp
p_dag_telecom-redis-1					

Servicio	URL	Credenciales
Airflow	http://localhost:8080	airflow/airflow
MLflow	http://localhost:5000	No requiere
Grafana	http://localhost:3000	admin/admin
Prometheus	http://localhost:9090	No requiere

Configuración y puesta en marcha del DAG en Airflow

localhost:8080/connection/list/

Airflow DAGs Cluster Activity Datasets Security Browse Admin Docs

18:05 (+02:00)

Added Row

List Connection

Search

+ Actions

Record Count: 1

	Conn Id	Conn Type	Description	Host	Port	Is Encrypted	Is Extra Encrypted
<input type="checkbox"/>	postgres_default	postgres		postgres-ml	5432	False	False



Explicación pipeline - DAG

1. Setup PostgreSQL (setup_postgres_tables)

- Creación de tablas en PostgreSQL
- Establecimiento de permisos

2. Preparación de Datos (prepare_data)

- Carga el dataset CSV
- Limpieza y formateado datos
- Manejo de valores faltantes y outliers
- Codificación de variables categóricas
- Selección de características importantes
- Escalado de datos y división en train/holdout
- Guarda registros en MLflow

3. Entrenamiento del Modelo (train_model)

- Configuración hiperparámetros de XGBoost
- Validación cruzada
- Entrenamiento del modelo final
- Cálculo de importancia de características
- Guardado del modelo entrenado
- Registro métricas en MLflow

4. Evaluación del Modelo (evaluate_model)

- Evaluación el modelo en conjunto holdout
- Cálculo de métricas de rendimiento
- Generación de visualizaciones (ROC, Confusion Matrix)
- Creación explicaciones SHAP
- Análisis de clientes específicos
- Guardado de resultados en PostgreSQL

Métricas y Evaluación

Métricas del Modelo

- Accuracy: Precisión general del modelo
- Precision: Proporción de verdaderos positivos
- Recall: Capacidad de detectar churns
- F1-Score: Media armónica de precision y recall
- AUC-ROC: Área bajo la curva ROC

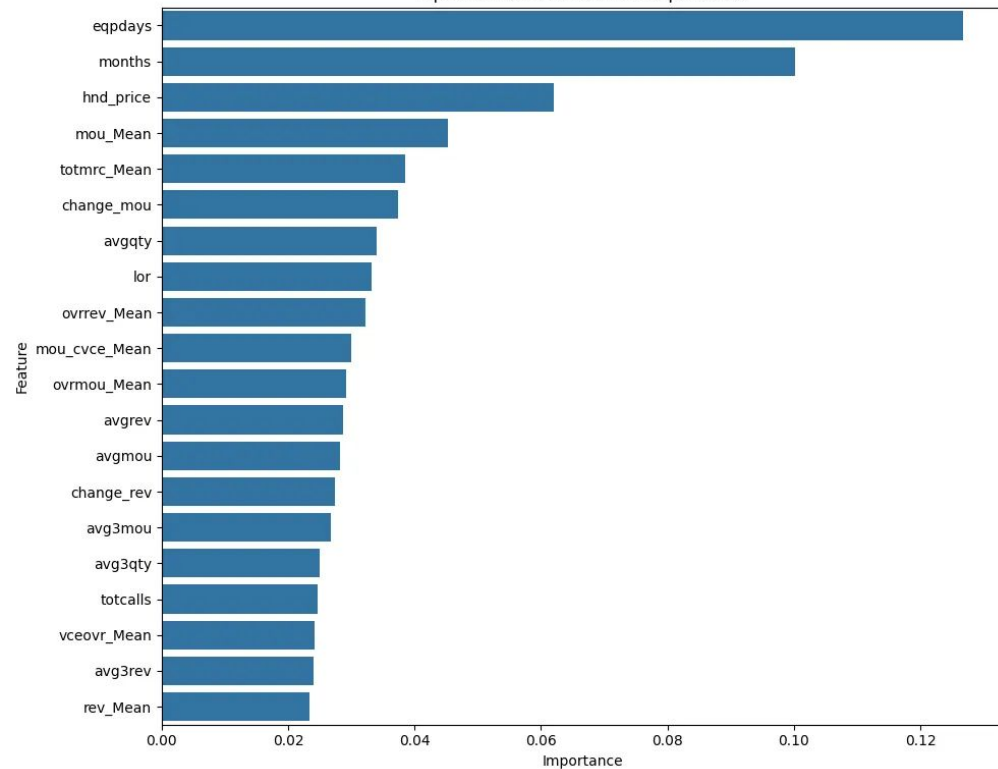
Explicabilidad SHAP

- Análisis de importancia de características
- Explicaciones a nivel individual
- Visualizaciones de factores de riesgo

Métrica	Valor
Accuracy	62.38%
Precision	61.29%
Recall	65.40%
F1-Score	63.28%
AUC-ROC	67.89%

Métricas y Evaluación

Top 20 Características Más Importantes



Matriz de Confusión

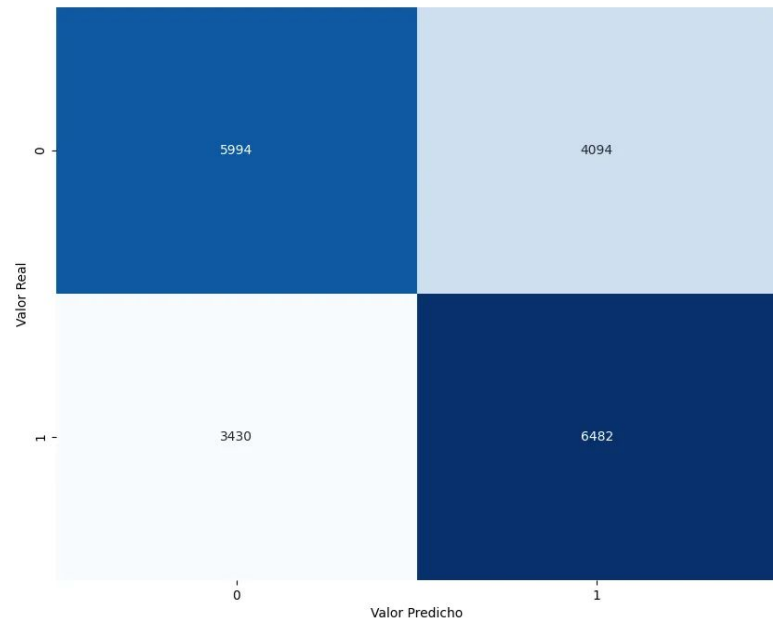


Gráfico SHAP

- Eje horizontal: Representa el impacto en la predicción de churn, donde:
 - Valores negativos (izquierda): Reducen la probabilidad de abandono
 - Valores positivos (derecha): Aumentan la probabilidad de abandono
 - La línea vertical en 0.0: Representa el punto neutro
- Eje vertical: Lista las características ordenadas por su importancia global en el modelo, con las más influyentes en la parte superior.
 - Puntos de colores: Cada punto representa un cliente en el conjunto de datos.
 - Azul: Valores bajos de esa característica
 - Rojo: Valores altos de esa característica

Análisis SHAP

Cliente de ALTO Riesgo (Probabilidad: >0.9)

Factores que INCREMENTAN el riesgo:

1. **mou_Mean**: +1.2145 (Uso muy bajo de minutos)
2. **change_mou**: +0.2944 (Sin cambio en el patrón de uso)
3. **eqpdays**: +0.2286 (Equipo relativamente reciente)
4. **rev_Mean**: +0.1971 (Ingresos bajos)
5. **avgqty**: +0.1951 (Alto consumo en cantidad)

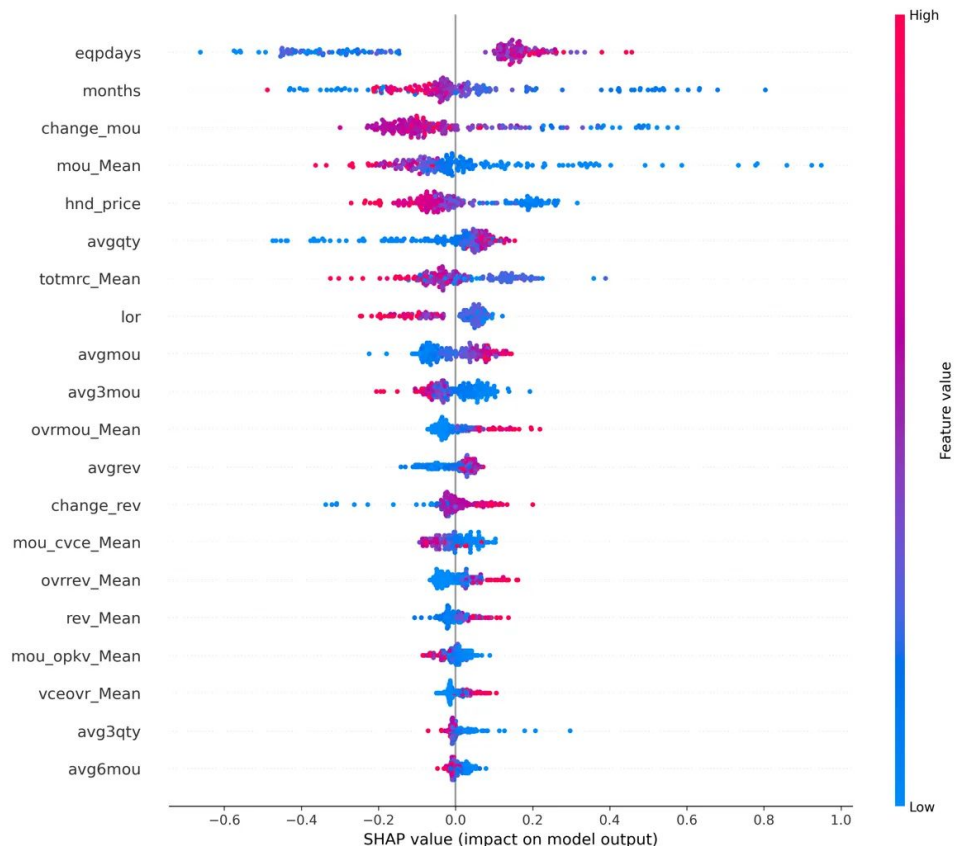
Factores que REDUCEN el riesgo:

1. **hnd_price**: -0.0399 (Dispositivo de precio medio)
2. **totmou**: -0.0329 (Total de minutos moderado)

Cliente de BAJO Riesgo (Probabilidad: <0.1)

Factores que PROTEGEN del churn:

1. **months**: -0.9502 (Cliente muy antiguo - 3.38 años)
2. **avgqty**: -0.4246 (Consumo moderado)
3. **totmrc_Mean**: -0.3098 (Facturas moderadas)
4. **rev_Mean**: -0.3075 (Ingresos consistentes)
5. **avg3rev**: -0.3003 (Ingresos promedio estables)



Estrategias teóricas para reducir el churn

Ciclo de renovación de equipos: La variable más importante (eqpdays - antigüedad del dispositivo) sugiere implementar programas proactivos de renovación de equipos antes de que los clientes lleguen al punto crítico donde consideran el cambio.

Segmentación por patrón de uso: Los patrones en mou_Mean y change_mou (Minutes Of Usage) indican la necesidad de monitorear cambios en los patrones de uso para identificar señales tempranas de abandono.

Optimización de planes: La relación entre totmrc_Mean (Cargo Mensual Recurrente Total) y el churn indica que se debe revisar la estructura de precios y ofrecer planes más ajustados a las necesidades reales para evitar cobros excesivos.

Retención de clientes por valor: El impacto del precio del terminal (hnd_price - precio del dispositivo) sugiere desarrollar programas de fidelización centrados en dispositivos premium para los clientes de mayor valor.

Atención a señales de alarma: Monitorear excesos en el uso (ovrmou_Mean, ovrrrev_Mean - excedentes del plan) puede permitir intervenciones preventivas antes de que el cliente decida abandonar.

Logs de ejecución en Airflow

The screenshot displays the Airflow web interface for a DAG named 'churn_prediction_advanced'. The task 'prepare_data' is selected, and its execution logs are visible. The interface includes a top navigation bar with buttons for 'Clear task', 'Mark state as...', and 'Filter Tasks'. Below the navigation bar, there are tabs for 'Details', 'Graph', 'Gantt', 'Code', and 'Logs', with 'Logs' being the active tab. On the left side, a vertical bar shows the duration of the task execution, with a green bar indicating the progress. Below this bar, a list of tasks is shown, with 'prepare_data' highlighted. The main area displays the logs for the 'prepare_data' task, which include information about the task's execution, such as the time, the file source, and the output of the task. The logs show that the task was executed successfully on 2025-05-19 at 16:12:16 UTC. The logs also show the output of the task, which includes a list of variables and their values. The variables include 'kidie17', 'credited', 'eqpdays', 'aplicando tratamiento de outliers', 'codificación de variables categóricas', 'variables categóricas identificadas', 'new_cell', 'crlscod', 'asl_flag', 'prizm_social_one', 'area', 'dualband', 'refurb_new', 'hnd_webcap', 'ownrent', 'dwlltype', 'marital', 'infobase', 'Hstatin', 'dwllsize', 'ethnic', 'kid0_2', 'kid3_5', 'kid6_10', 'kid11_15', 'kid16_17', 'credited', 'Dimensiones antes de la codificación', 'Dimensiones después de la codificación', 'base.py:73', 'db_operations.py:24', 'data_preparation.py:93', 'data_preparation.py:243', 'data_preparation.py:265', 'data_preparation.py:274', and 'logging_mixin.py:151'. The logs also show a warning message from mlflow.sklearn: 'Model was missing function: predict. Not logging python_function flavor!'.

Duration: 00:00:31

setup_postgres_tables

prepare_data

train_model

evaluate_model

send_summary

churn_prediction_advanced / 2025-05-18, 02:00:00 / prepare_data

Clear task Mark state as... Filter Tasks

Details Graph Gantt Code Logs

(by attempts)

1

All Levels All File Sources

Wrap Download See More

```
[2025-05-19, 16:12:16 UTC] {data_preparation.py:156} INFO - kidie17: Rellenado con la moda: U
[2025-05-19, 16:12:16 UTC] {data_preparation.py:156} INFO - credited: Rellenado con la moda: Y
[2025-05-19, 16:12:16 UTC] {data_preparation.py:159} INFO - eqpdays: Rellenado con la mediana: 342.0
[2025-05-19, 16:12:16 UTC] {data_preparation.py:166} INFO - Aplicando tratamiento de outliers...
[2025-05-19, 16:12:17 UTC] {data_preparation.py:213} INFO - Codificación de variables categóricas...
[2025-05-19, 16:12:17 UTC] {data_preparation.py:220} INFO - Variables categóricas identificadas: 21
[2025-05-19, 16:12:17 UTC] {data_preparation.py:224} INFO - new_cell: 3 valores únicos
[2025-05-19, 16:12:17 UTC] {data_preparation.py:224} INFO - crlscod: 54 valores únicos
[2025-05-19, 16:12:17 UTC] {data_preparation.py:224} INFO - asl_flag: 2 valores únicos
[2025-05-19, 16:12:17 UTC] {data_preparation.py:224} INFO - prizm_social_one: 5 valores únicos
[2025-05-19, 16:12:17 UTC] {data_preparation.py:224} INFO - area: 19 valores únicos
[2025-05-19, 16:12:17 UTC] {data_preparation.py:224} INFO - dualband: 4 valores únicos
[2025-05-19, 16:12:17 UTC] {data_preparation.py:224} INFO - refurb_new: 2 valores únicos
[2025-05-19, 16:12:17 UTC] {data_preparation.py:224} INFO - hnd_webcap: 3 valores únicos
[2025-05-19, 16:12:17 UTC] {data_preparation.py:224} INFO - ownrent: 2 valores únicos
[2025-05-19, 16:12:17 UTC] {data_preparation.py:224} INFO - dwlltype: 2 valores únicos
[2025-05-19, 16:12:17 UTC] {data_preparation.py:224} INFO - marital: 5 valores únicos
[2025-05-19, 16:12:17 UTC] {data_preparation.py:224} INFO - infobase: 2 valores únicos
[2025-05-19, 16:12:17 UTC] {data_preparation.py:224} INFO - Hstatin: 6 valores únicos
[2025-05-19, 16:12:17 UTC] {data_preparation.py:224} INFO - dwllsize: 15 valores únicos
[2025-05-19, 16:12:17 UTC] {data_preparation.py:224} INFO - ethnic: 17 valores únicos
[2025-05-19, 16:12:17 UTC] {data_preparation.py:224} INFO - kid0_2: 2 valores únicos
[2025-05-19, 16:12:17 UTC] {data_preparation.py:224} INFO - kid3_5: 2 valores únicos
[2025-05-19, 16:12:17 UTC] {data_preparation.py:224} INFO - kid6_10: 2 valores únicos
[2025-05-19, 16:12:17 UTC] {data_preparation.py:224} INFO - kid11_15: 2 valores únicos
[2025-05-19, 16:12:17 UTC] {data_preparation.py:224} INFO - kid16_17: 2 valores únicos
[2025-05-19, 16:12:17 UTC] {data_preparation.py:224} INFO - credited: 2 valores únicos
[2025-05-19, 16:12:17 UTC] {data_preparation.py:233} INFO - Dimensiones antes de la codificación: (100000, 100)
[2025-05-19, 16:12:17 UTC] {data_preparation.py:234} INFO - Dimensiones después de la codificación: (100000, 211)
[2025-05-19, 16:12:17 UTC] {base.py:73} INFO - Using connection ID 'postgres_default' for task execution.
[2025-05-19, 16:12:24 UTC] {db_operations.py:24} INFO - Datos guardados en tabla churn_raw_data de PostgreSQL
[2025-05-19, 16:12:24 UTC] {data_preparation.py:93} INFO - Muestra de datos guardada en PostgreSQL (tabla: churn_raw_data)
[2025-05-19, 16:12:25 UTC] {data_preparation.py:243} INFO - Selección de características...
[2025-05-19, 16:12:31 UTC] {data_preparation.py:265} INFO - Se seleccionaron 30 características de 209 disponibles
[2025-05-19, 16:12:31 UTC] {data_preparation.py:274} INFO - Escalado de características...
[2025-05-19, 16:12:31 UTC] {logging_mixin.py:151} WARNING - 2025/05/19 16:12:31 WARNING mlflow.sklearn: Model was missing function: predict. Not logging python_function flavor!
```

Ejecución completada con éxito:

The screenshot displays the Apache Airflow web interface. On the left, a task list shows the following tasks: `setup_postgres_tables`, `prepare_data`, `train_model`, `evaluate_model`, and `send_summary`. The `send_summary` task is highlighted in blue, indicating it is the selected task. Above the task list, a progress bar shows the duration of the tasks: `00:06:35` for the first task, `00:03:17` for the second, and `00:00:00` for the third. The main panel shows the details of the `send_summary` task, which is part of the `churn_prediction_advanced` DAG. The task is in the 'Run' state, and the execution date is `2025-05-18, 02:00:00`. The task ID is `send_summary`. The task is completed successfully, and the logs show the following output:

```
[2025-05-19, 16:18:36 UTC] [task_command.py:415] INFO - Running <TaskInstance: churn_prediction_advanced.send_summary manual__2025-05-19T16:12:01.302389+00:00 [running]> on host c28f1781f2e6
[2025-05-19, 16:18:37 UTC] [taskinstance.py:1660] INFO - Exporting env vars: AIRFLOW_CTX_DAG_EMAIL='data_science@example.com' AIRFLOW_CTX_DAG_OWNER='data_science_team' AIRFLOW_CTX_DAG_ID='churn_prediction_advanced' AIRFLOW
[2025-05-19, 16:18:37 UTC] [reporting.py:85] INFO -
=====
RESUMEN DE EJECUCIÓN: PREDICCIÓN DE CHURN
=====

Fecha de ejecución: 2025-05-19 16:18:37

MÉTRICAS DEL MODELO:
- Accuracy: 0.6238
- Precision: 0.6129
- Recall: 0.6540
- F1 Score: 0.6328
- AUC-ROC: 0.6789

UBICACIÓN DE RESULTADOS:
- Modelo guardado en: /opt/***/data/churn_model.pkl
- Métricas detalladas en: /opt/***/data/model_metrics.json
- Visualizaciones en: /opt/***/data

Para ver más detalles, consulte el dashboard de MLflow o
los logs de ejecución en Airflow.

[2025-05-19, 16:18:37 UTC] [reporting.py:126] INFO - Simulando envío de email de notificación...
[2025-05-19, 16:18:37 UTC] [reporting.py:127] INFO - Para implementar realmente, configurar SMTP en Airflow
[2025-05-19, 16:18:37 UTC] [reporting.py:142] INFO - Simulando envío de notificación a Slack...
[2025-05-19, 16:18:37 UTC] [reporting.py:143] INFO - Para implementar realmente, configurar webhook de Slack en Airflow
[2025-05-19, 16:18:37 UTC] [reporting.py:159] INFO - Archivo de integración creado en /opt/***/data/integration_data.json
[2025-05-19, 16:18:37 UTC] [python.py:194] INFO - Done. Returned value was: {'status': 'success', 'message': 'Resumen generado correctamente', 'summary_path': '/opt/***/data/execution_summary.txt'}
[2025-05-19, 16:18:37 UTC] [taskinstance.py:1398] INFO - Marking task as SUCCESS. dag_id=churn_prediction_advanced, task_id=send_summary, execution_date=20250519T161201, start_date=20250519T161836, end_date=20250519T161837
[2025-05-19, 16:18:37 UTC] [local_task_job_runner.py:228] INFO - Task exited with return code 0
[2025-05-19, 16:18:37 UTC] [taskinstance.py:2776] INFO - 0 downstream tasks scheduled from follow-on schedule check
```

Almacenamiento en Postgres

1. Datos crudos y procesados:

- churn_raw_data: Muestra de los datos originales
- churn_processed_data: Datos después de la preparación y transformación

2. Métricas y resultados del modelo:

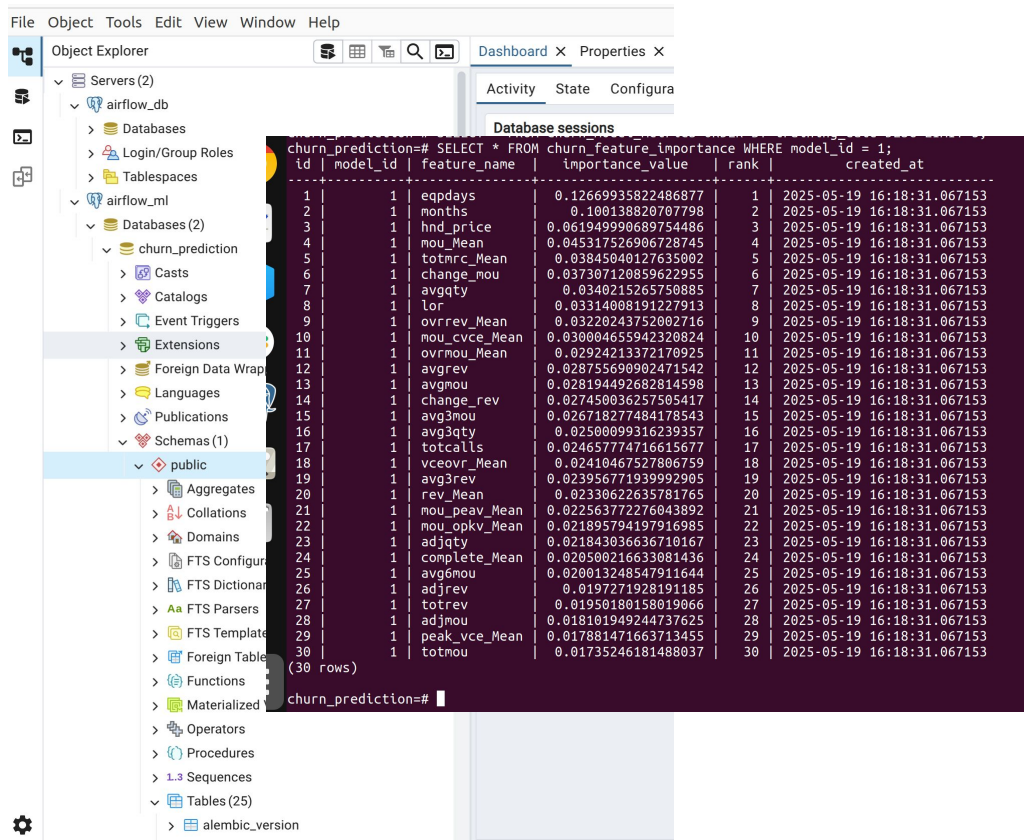
- churn_model_metrics: Métricas de evaluación (accuracy, precision, recall, F1, AUC-ROC)
- churn_feature_importance: Importancia de las características del modelo
- churn_feature_importance_temp: Tabla temporal para la importancia de características

3. Predicciones y explicaciones:

- churn_predictions: Predicciones individuales para cada cliente con sus probabilidades
- churn_shap_values: Valores SHAP para explicar predicciones individuales

4. Datos de monitoreo:

- churn_model_monitoring: Eventos de monitoreo del modelo
- churn_action_plans: Planes de acción basados en predicciones



The screenshot displays a PostgreSQL database interface. On the left, the 'Object Explorer' pane shows a tree structure of the database, including 'Servers (2)', 'airflow_db', 'Databases', 'Login/Group Roles', 'Tablespaces', 'airflow_ml', 'Databases (2)', 'churn_prediction', 'Casts', 'Catalogs', 'Event Triggers', 'Extensions', 'Foreign Data Wraps', 'Languages', 'Publications', 'Schemas (1)', 'public', 'Aggregates', 'Collations', 'Domains', 'FTS Configur', 'FTS Dictionar', 'FTS Parsers', 'FTS Template', 'Foreign Table', 'Functions', 'Materialized', 'Operators', 'Procedures', 'Sequences', and 'Tables (25)'. The 'churn_prediction' table is selected under the 'public' schema.

The main window shows a query result for the 'churn_prediction' table. The query is: `SELECT * FROM churn_prediction WHERE model_id = 1;`. The result is a table with 30 rows and 6 columns: `id`, `model_id`, `feature_name`, `importance_value`, `rank`, and `created_at`. The first few rows are:

id	model_id	feature_name	importance_value	rank	created_at
1	1	eqpdays	0.12669935822486877	1	2025-05-19 16:18:31.067153
2	1	months	0.100138820707798	2	2025-05-19 16:18:31.067153
3	1	hnd_price	0.061949990689754486	3	2025-05-19 16:18:31.067153
4	1	mou_Mean	0.045317526906728745	4	2025-05-19 16:18:31.067153
5	1	totmrc_Mean	0.03845040127635002	5	2025-05-19 16:18:31.067153
6	1	change_mou	0.037307120859622955	6	2025-05-19 16:18:31.067153
7	1	avgqty	0.0340215265750885	7	2025-05-19 16:18:31.067153
8	1	lor	0.03314008191227913	8	2025-05-19 16:18:31.067153
9	1	ovrrev_Mean	0.03220243752002716	9	2025-05-19 16:18:31.067153
10	1	mou_cvce_Mean	0.030004655942320824	10	2025-05-19 16:18:31.067153
11	1	ovrmou_Mean	0.02924213372170925	11	2025-05-19 16:18:31.067153
12	1	avgrev	0.028755690902471542	12	2025-05-19 16:18:31.067153
13	1	avgmou	0.028194492682814598	13	2025-05-19 16:18:31.067153
14	1	change_rev	0.027450036257505417	14	2025-05-19 16:18:31.067153
15	1	avg3mou	0.026718277484178543	15	2025-05-19 16:18:31.067153
16	1	avg3qty	0.02500099316239357	16	2025-05-19 16:18:31.067153
17	1	totcalls	0.024657774716615677	17	2025-05-19 16:18:31.067153
18	1	vceovr_Mean	0.02410467527806759	18	2025-05-19 16:18:31.067153
19	1	avg3rev	0.023956771939992905	19	2025-05-19 16:18:31.067153
20	1	rev_Mean	0.02330622635781765	20	2025-05-19 16:18:31.067153
21	1	mou_peav_Mean	0.022563772276043892	21	2025-05-19 16:18:31.067153
22	1	mou_opkv_Mean	0.021895794197916985	22	2025-05-19 16:18:31.067153
23	1	adjqty	0.021843036636710167	23	2025-05-19 16:18:31.067153
24	1	complete_Mean	0.020500216633081436	24	2025-05-19 16:18:31.067153
25	1	avg6mou	0.020013248547911644	25	2025-05-19 16:18:31.067153
26	1	adjrev	0.0197271928191185	26	2025-05-19 16:18:31.067153
27	1	totrev	0.01950180158019066	27	2025-05-19 16:18:31.067153
28	1	adjmou	0.018101949244737625	28	2025-05-19 16:18:31.067153
29	1	peak_vce_Mean	0.017881471663713455	29	2025-05-19 16:18:31.067153
30	1	totmou	0.01735246181488037	30	2025-05-19 16:18:31.067153

The bottom of the window shows the query: `churn_prediction=#`.

Almacenamiento en Local

Datos procesados:

- processed_data.pkl: Datos de entrenamiento, escalador y nombres de características
- holdout_data.pkl: Datos de validación para evaluación final

Modelo entrenado:

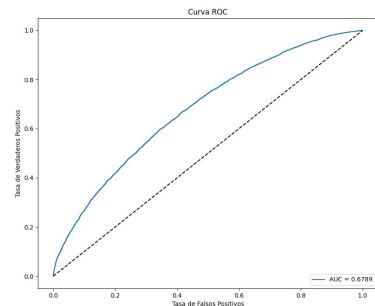
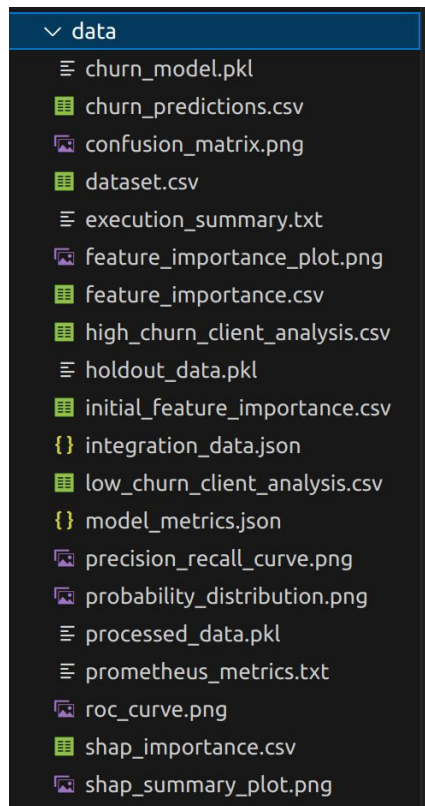
- churn_model.pkl: Modelo XGBoost serializado con sus metadatos

Métricas y visualizaciones:

- model_metrics.json: Métricas de evaluación en formato JSON
- feature_importance.csv: Importancia de características
- confusion_matrix.png: Matriz de confusión visualizada
- roc_curve.png: Curva ROC
- precision_recall_curve.png: Curva Precision-Recall
- shap_summary_plot.png: Visualización de valores SHAP

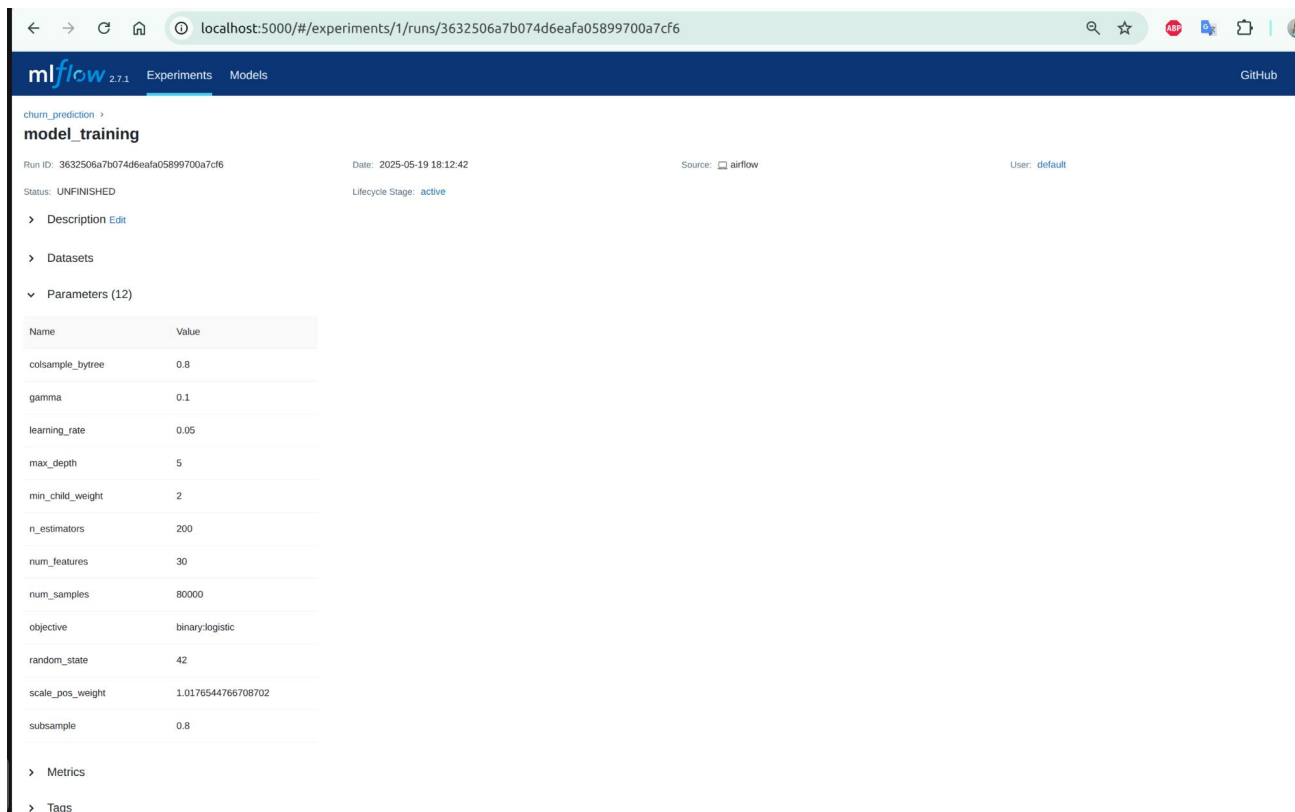
Datos para Prometheus:

- prometheus_metrics.txt: Métricas en formato para Prometheus



```
ecom > data > {} model_metrics.json > ...  
{  
  "accuracy": 0.6238,  
  "precision": 0.6128971255673222,  
  "recall": 0.653954802259887,  
  "f1": 0.6327606403748536,  
  "auc_roc": 0.678921005642677,  
  "evaluation_date": "2025-05-21T14:14:14Z"  
}
```


Visibilidad de métricas del entrenamiento en MLFlow



localhost:5000/#/experiments/1/runs/3632506a7b074d6eafa05899700a7cf6

mlflow 2.7.1 Experiments Models GitHub

churn_prediction >
model_training

Run ID: 3632506a7b074d6eafa05899700a7cf6 Date: 2025-05-19 18:12:42 Source: airflow User: default

Status: UNFINISHED Lifecycle Stage: active

> Description [Edit](#)

> Datasets

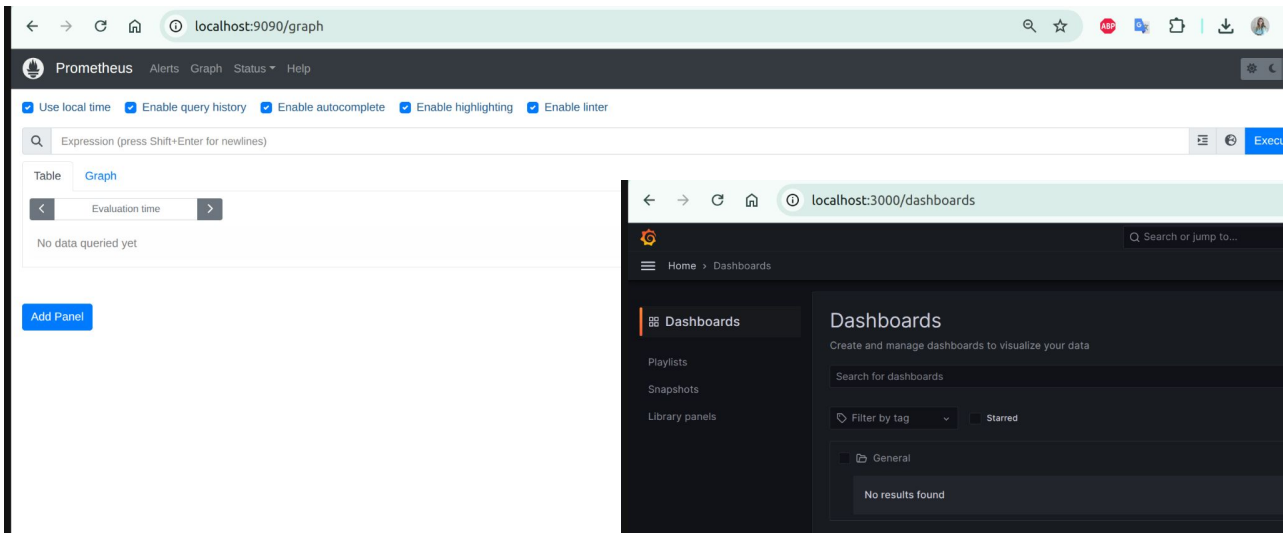
▼ Parameters (12)

Name	Value
colsample_bytree	0.8
gamma	0.1
learning_rate	0.05
max_depth	5
min_child_weight	2
n_estimators	200
num_features	30
num_samples	80000
objective	binary:logistic
random_state	42
scale_pos_weight	1.0176544766708702
subsample	0.8

> Metrics

> Tags

Integración con herramientas de monitorización



¡Gracias!
