



RETO PROYECTO FINAL

Big Data & Inteligencia Artificial

Jose Maria Pinilla Melgar

Contenido

1. Introducción del caso de uso.....	2
1.1 Contexto del caso de uso	2
1.2 Problema de negocio	2
1.3 Objetivo del reto.....	2
1.4 Dataset	2
1.5 Objetivos específicos	2
1.6 Alcance del proyecto.....	3
1.7 Beneficios esperados para el negocio	3
2. Entendimiento de los datos (EDA)	3
2.1 Subida de datos a S3.....	3
2.2 Creación de dominio en SageMaker AI.....	4
2.3 Creación de un usuario.....	7
2.4 Data Wrangler.....	8
2.4.1 Informe EDA completo	9
3. Preparación de los datos.....	14
4. Modelado.....	16
5. Evaluación y conclusión	18
6. Despliegue	19

1. Introducción del caso de uso

1.1 Contexto del caso de uso

Una empresa de mensajería de comercio electrónico realiza entregas de última milla y ya dispone de un software de optimización de rutas que calcula rutas “cortas”. Sin embargo, la empresa quiere extraer más valor de los datos operativos para entender qué factores están influyendo en los tiempos de entrega y cómo mejorar la planificación diaria de sus operaciones.

1.2 Problema de negocio

Aunque la empresa optimiza rutas por distancia, el tiempo real de entrega puede variar significativamente por factores que no dependen solo de la ruta más corta. Esta variabilidad genera:

- Retrasos y peor experiencia del cliente
- Mayor coste operativo (horas, combustible, reintentos)
- Dificultad para planificar turnos y asignación de recursos (conductores/vehículos)
- Mayor riesgo operativo si existen patrones asociados a incidentes

1.3 Objetivo del reto

Desarrollar un sistema analítico basado en datos que permita entender, predecir y reducir los tiempos de entrega de última milla.

1.4 Dataset

<https://www.kaggle.com/datasets/pingpang/lastmiledeliverytimes?resource=download>

1.5 Objetivos específicos

1. Analizar los factores que mas influyen en el tiempo de entrega (EDA + conclusiones).
2. Entrenar modelos de Machine Learning para predecir el tiempo de reparto a partir del resto de variables.
3. Comparar al menos 2 modelos y seleccionar el mejor según métricas.
4. Proponer recomendaciones operativas basadas en los hallazgos.
5. Preparar el despliegue del modelo para simular predicciones.

1.6 Alcance del proyecto

- Incluye: análisis exploratorio, preparación de datos ingeniería de características, modelado predictivo y despliegue.

1.7 Beneficios esperados para el negocio

- Mejor estimación de tiempos de entrega.
- Identificación de palancas de mejora: carga de paquetes, canal de cumplimiento, experiencia, tipo de vehículo, etc.
- Reducción de costes por planificación más eficiente.
- Mejora de KPIs logísticos: puntualidad, productividad por ruta, menor variabilidad.

2. Entendimiento de los datos (EDA)

2.1 Subida de datos a S3

- Creación de Bucket

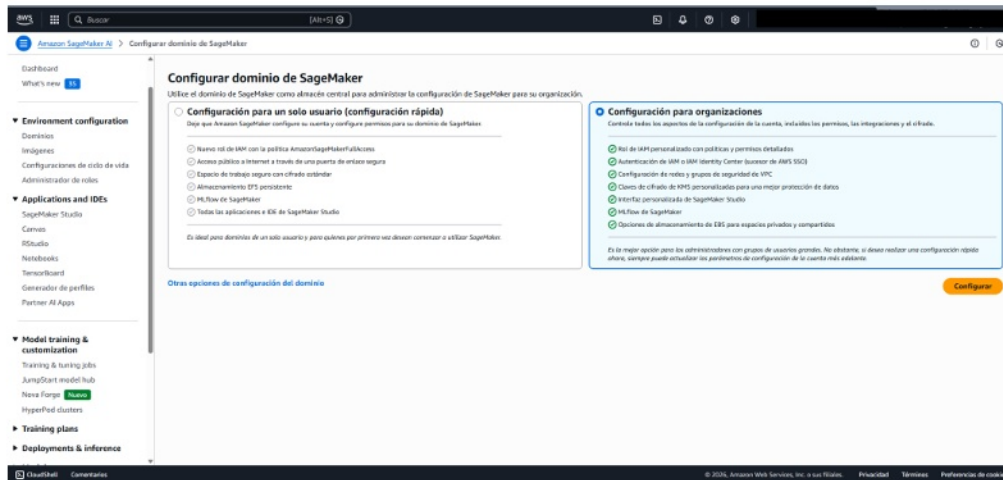
The screenshot shows the 'Crear bucket' (Create bucket) page in the AWS console. The page is titled 'Crear bucket' and includes a sub-header 'Información'. Below this, there are several sections for configuring the bucket. The 'Configuración general' (General configuration) section is active, showing the 'Región de AWS' (AWS Region) as 'us-east-1' and the 'Tipo de bucket' (Bucket type) as 'Uso general' (General purpose). The 'Nombre del bucket' (Bucket name) is 'delivery-project'. The 'Propiedad de objetos' (Object properties) section is also visible, showing 'ACL de deshabilitación (recomendado)' (Disable ACL (recommended)) and 'ACL habilitada' (ACL enabled).

- Carga de datos en el Bucket

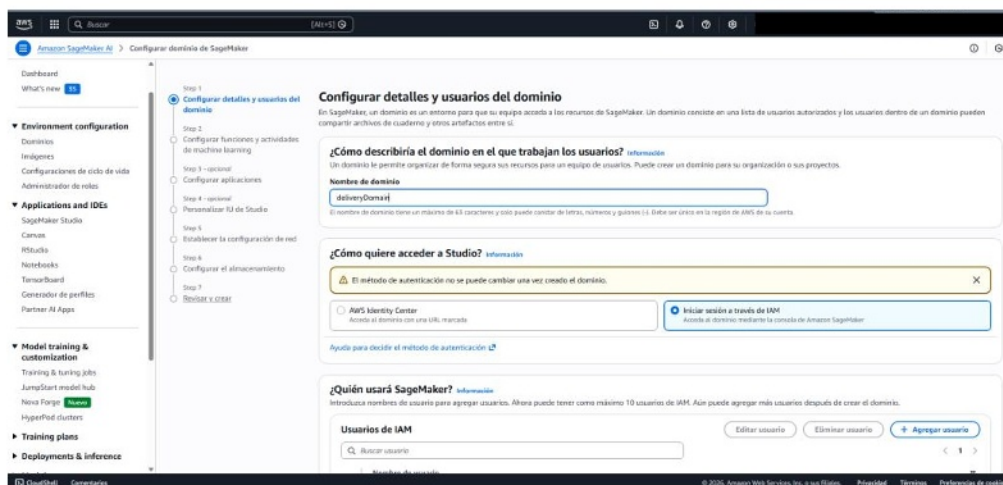
The screenshot shows the 'Cargar: estado' (Upload: status) page in the AWS console. The page is titled 'Cargar: estado' and includes a sub-header 'Resumen' (Summary). The 'Resumen' section shows the 'Destino' (Destination) as 's3://delivery-project' and the 'Realizado correctamente' (Successfully completed) status as '1 archivo, 33.1 KB (100.00%)'. The 'Archivos y carpetas' (Files and folders) section is active, showing a table with columns for 'Nombre' (Name), 'Carpeta' (Folder), 'Tipo' (Type), 'Tamaño' (Size), 'Estado' (Status), and 'Error' (Error). The table contains one row with the file 'logMileDeliveryTimes080305.csv' and a status of 'Realizado correctamente' (Successfully completed).

2.2 Creación de dominio en SageMaker AI

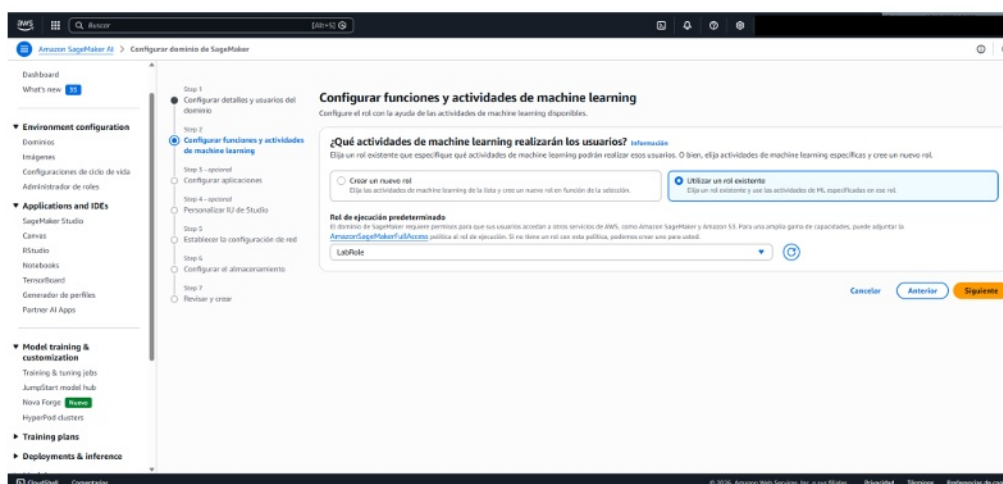
- Empezamos con la creación del dominio de SageMaker, en la primera opción escogeremos la configuración para organizaciones:



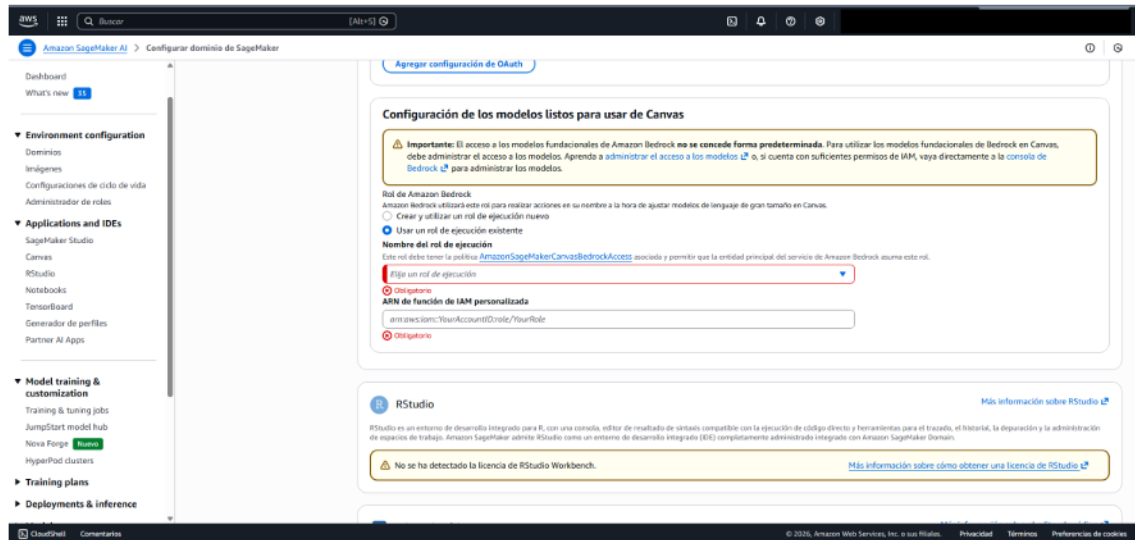
- Ahora escribiremos un nombre para el dominio:



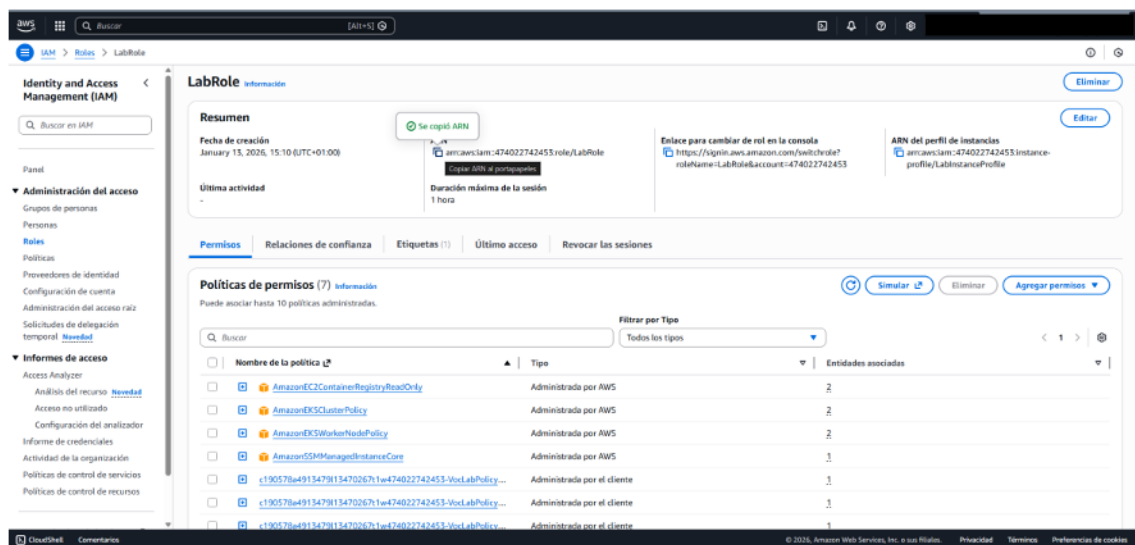
- En el siguiente paso elegiremos nuestro rol de labRole con la opción de rol existente:



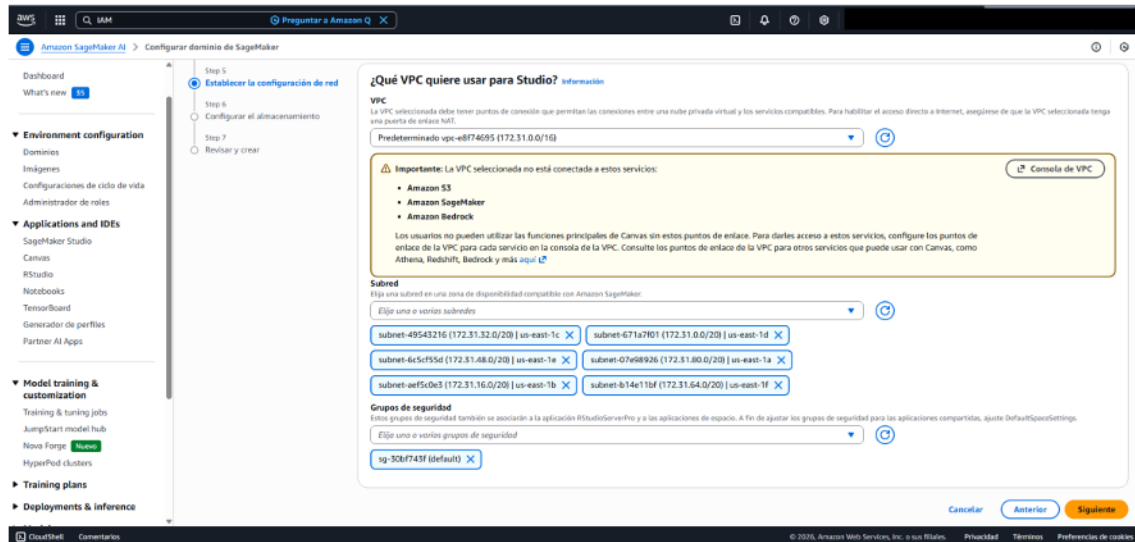
- Ahora nos dirigimos al apartado de Canvas y buscamos la opción de Configuración de modelos listos para Canvas, elegimos usar un rol existente y nos pedirá nuestro arn:



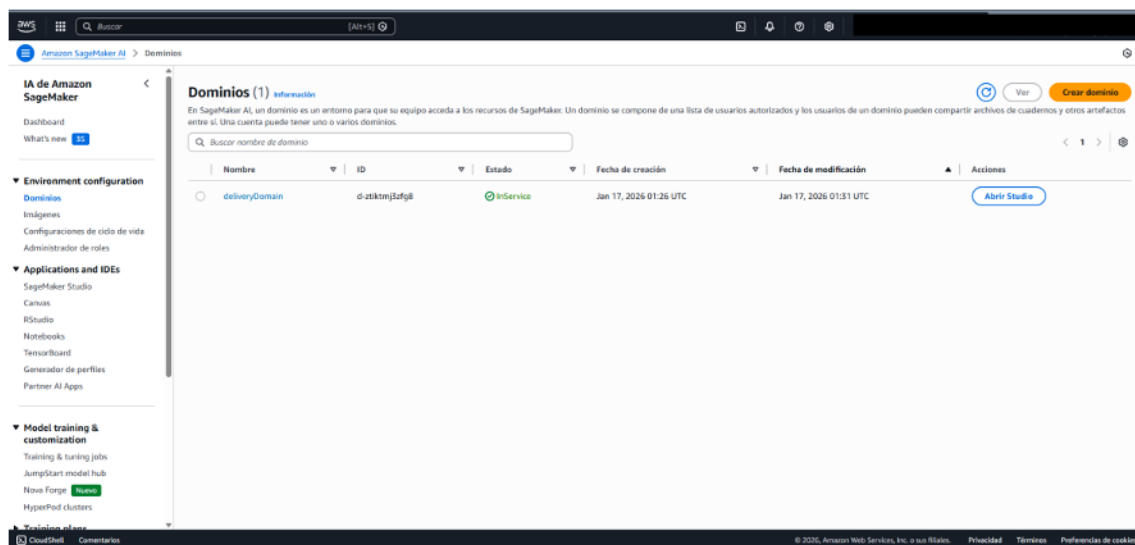
- Para rellenar el arn del paso anterior tenemos que acceder al servicio de AWS llamado IAM donde buscaremos el role de labRole y nos saldrá toda la información de dicho role incluido el arn:



- Después de introducir el arn le daremos a siguiente y nos llevara al apartado de la configuración de red donde elegiremos la WPC, todas las subredes que nos deje y el grupo de seguridad:

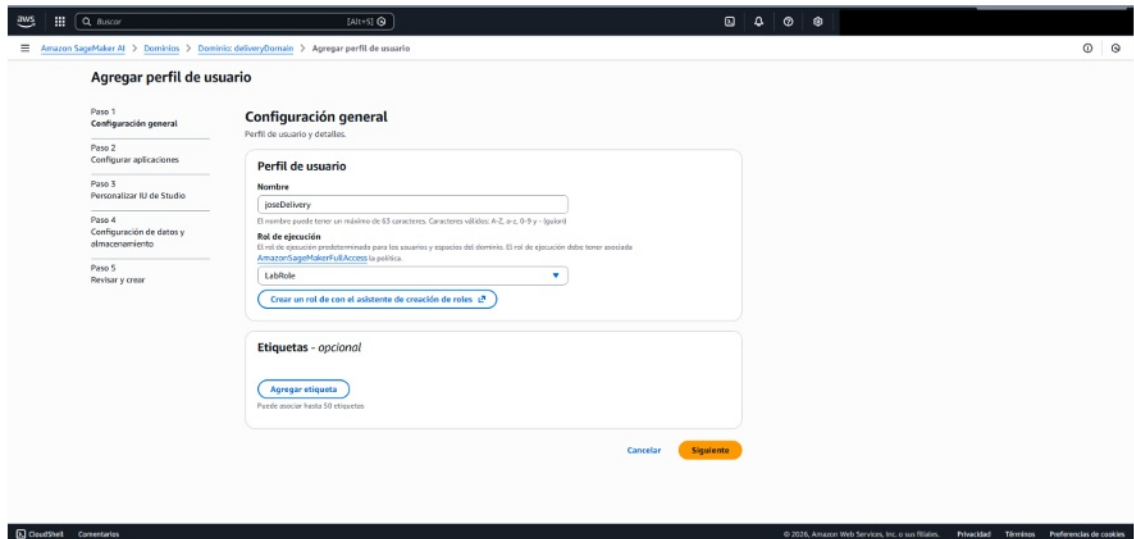


- Para finalizar nos aparecerá el resumen de toda la configuración, le daremos a enviar y esperaremos unos minutos a que el estado del dominio se encuentre en InService:

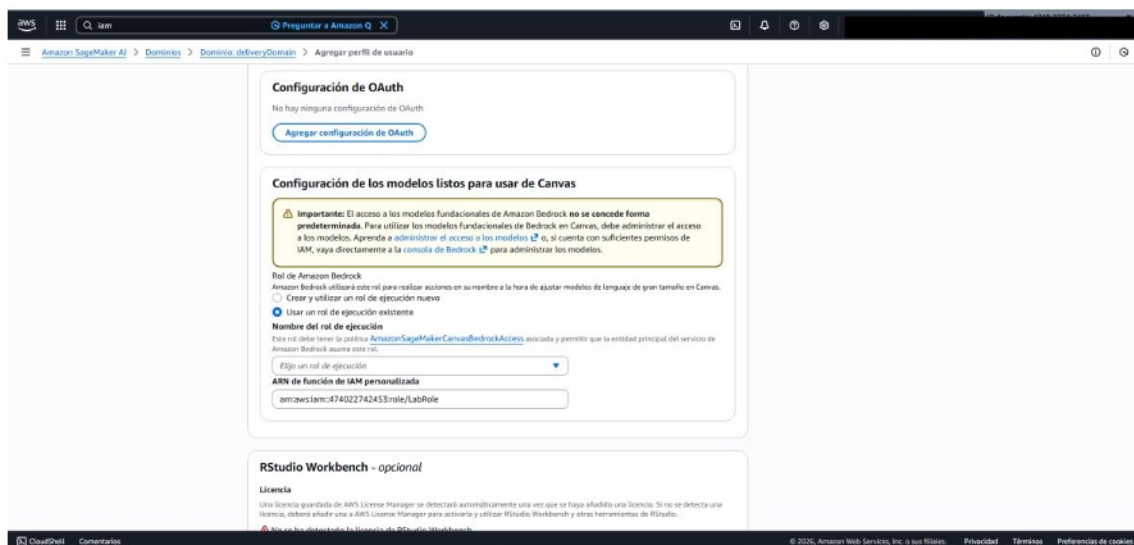


2.3 Creación de un usuario

- Una vez tengamos el dominio tenemos que asignarle un usuario que crearemos a continuación:



- Después de asignar nombre y role nos pedirá exactamente el mismo paso que en el dominio donde en la sección de Canvas volveremos a introducir el arn:



- El resto de los pasos los dejamos tal y como están predeterminadamente y al final le damos a enviar.

2.4 Data Wrangler

- Dentro de Canvas en la columna de la izquierda nos aparecerá Data Wrangler donde empezaremos a realizar el EDA, lo primero que nos pedirá será la importación de los datos:

Import tabular data ×

Select a data source: Amazon S3

Input S3 endpoint

Provide the ARN, URL, or alias

Aliases should have the format: "s3://<alias-prefix>/<alias>". URLs should have the format: "s3://<bucket>/<key>". ARNs should have ARN standard format. [Learn More](#)

Amazon S3 / delivery-project Search Amazon S3

<input checked="" type="checkbox"/>	Name	Size	Last updated ↓
<input checked="" type="checkbox"/>	lastMileDeliveryTimesWithCats.csv	33 kB	01/16/2026 8:48 PM

Cancel Next

- Como primer análisis a generar, crearemos el mas importante que seria el informe EDA al completo:

Data Wrangler: Data flow > New data flow 2026-1-17 4:57:24 PM.flow > lastMileDeliveryTimesWithCats.csv

Data Quality And Insights Report

Data Quality And Insights Report: informe_eda_completo

No Preview available

Create analysis

Analysis type *

Data Quality And Insights Report

Analysis name

informe_eda_completo

Target column *

DeliveryTime

Problem type

☒ Regression

☐ Classification

Data size *

☒ Sampled dataset

☐ Full dataset

Clear Create

2.4.1 Informe EDA completo

1. Descripción general del Dataset

El conjunto de datos analizados contiene un total de 1000 registros correspondientes a entregas a domicilio realizadas por una empresa de mensajería de comercio electrónico.

El Dataset incluye 8 variables, de las cuales:

- 5 son numéricas
- 2 son binarias
- 1 es la variable objetivo

No existen valores nulos ni filas duplicadas, por lo que los datos presentan una alta calidad inicial y son adecuados para su análisis y modelado.

2. Variable objetivo: DeliveryTime

La variable objetivo DeliveryTime representa el tiempo total de entrega en horas.

Estadísticas principales:

- Valor mínimo: 0,6 horas
- Valor máximo: 11,4 horas
- Media: 5,77 horas
- Mediana: 5,8 horas
- Distribución aproximadamente normal

El histograma muestra que la mayoría de las entregas se concentran entre 4 y 8 horas, existiendo algunos casos extremos de entregas muy rápidas o muy lentas.

Esto indica que la operación logística tiene una duración media estable, pero con cierta variabilidad operativa que conviene optimizar.

Target column



3. Calidad de los datos

El informe de calidad indica:

- 100% de valores válidos
- 0% de valores perdidos
- 0% de filas duplicadas
- No se detectan advertencias de severidad alta

Esto confirma que los datos están correctamente registrados y no requieren limpieza adicional para su análisis.

Summary

Dataset statistics

Key	Value	
Number of features	8	
Number of rows	1000	
Missing	0%	
Valid	100%	
Duplicate rows	0%	

4. Importancia predictiva de las variables

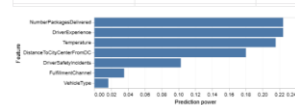
El sistema ha evaluado la capacidad predictiva individual de cada variable sobre el tiempo de entrega.

Feature summary

See a summary of the features ordered by the prediction power. Prediction power is measured by stratified splitting the data into 80%/20% training and validation folds. We fit a model for each feature separately on the training fold after applying minimal feature pre-processing and measure prediction performance on the validation data.

- The scores are normalized to the range (0,1).
- Higher prediction power scores, toward 1, indicate columns that are more useful for predicting the target on their own.
- Lower scores, toward 0 point to columns that contain little useful information for predicting the target on their own. Although it can happen that a column is uninformative on its own but is useful in predicting the target when used in tandem with other features, a low score usually indicates the feature is redundant.
- A score of 1 implies perfect predictive abilities, which often indicates an error called target leakage. The cause is typically a column that will not be available at prediction time such as a duplicate of the target.

Feature	Prediction power	Type	Valid	Missing	High severity warnings	Medium severity warnings	
NumberPackagesDelivered	0.223	numeric	100%	0%	0	0	
DriverExperience	0.223	numeric	100%	0%	0	0	
Temperature	0.216	numeric	100%	0%	0	0	
DistanceToCityCenterFromDC	0.18	numeric	100%	0%	0	0	
DriverSafetyIncidents	0.103	numeric	100%	0%	0	0	
FulfillmentChannel	0.0354	binary	100%	0%	0	0	



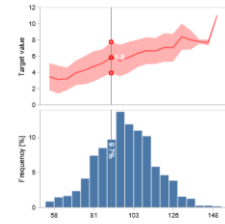
Esto indica que los factores operativos tienen un impacto mayor que los factores estructurales del vehículo.

5. Análisis de las principales variables

Feature details

NumberPackagesDelivered - numeric

key	value	
Feature name	NumberPackagesDelivered	
Type	numeric	
Prediction power	0.225	
Valid	100%	
Missing	0%	
Outliers	0%	
Min	53	
Max	153	
Mean	97.7	



Histogram of NumberPackagesDelivered with the corresponding target distribution. The lower plot provides the feature distribution and the upper - the corresponding target average with a confidence band of one standard deviation.

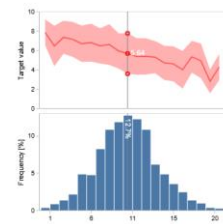
El gráfico muestra una relación claramente creciente entre el número de paquetes y el tiempo de entrega.

A mayor volumen de entregas en una ruta, mayor es la duración total del reparto.

Esto indica que la carga de trabajo diaria del conductor es uno de los principales factores que afectan al tiempo de entrega.

DriverExperience - numeric

key	value	
Feature name	DriverExperience	
Type	numeric	
Prediction power	0.225	
Valid	100%	
Missing	0%	
Outliers	0%	
Min	0.2	
Max	21.1	
Mean	10.1	



Histogram of DriverExperience with the corresponding target distribution. The lower plot provides the feature distribution and the upper - the corresponding target average with a confidence band of one standard deviation.

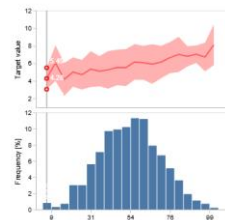
Se observa una relación inversa entre experiencia y tiempo de entrega:

los conductores más experimentados tienden a completar las rutas en menos tiempo.

Esto sugiere que la experiencia mejora la eficiencia operativa y la toma de decisiones en ruta.

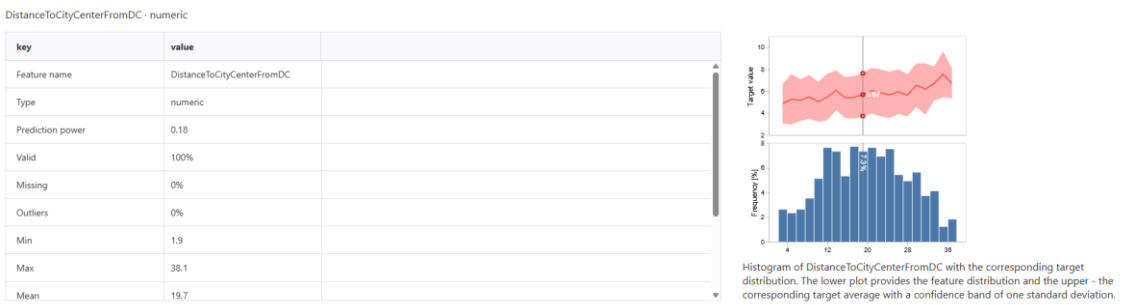
Temperature - numeric

key	value	
Feature name	Temperature	
Type	numeric	
Prediction power	0.216	
Valid	100%	
Missing	0%	
Outliers	0%	
Min	3.7	
Max	104	
Mean	53.2	



Histogram of Temperature with the corresponding target distribution. The lower plot provides the feature distribution and the upper - the corresponding target average with a confidence band of one standard deviation.

El gráfico muestra una ligera tendencia al aumento del tiempo de entrega con temperaturas más extremas, lo que sugiere que las condiciones climáticas pueden afectar al rendimiento de las rutas.



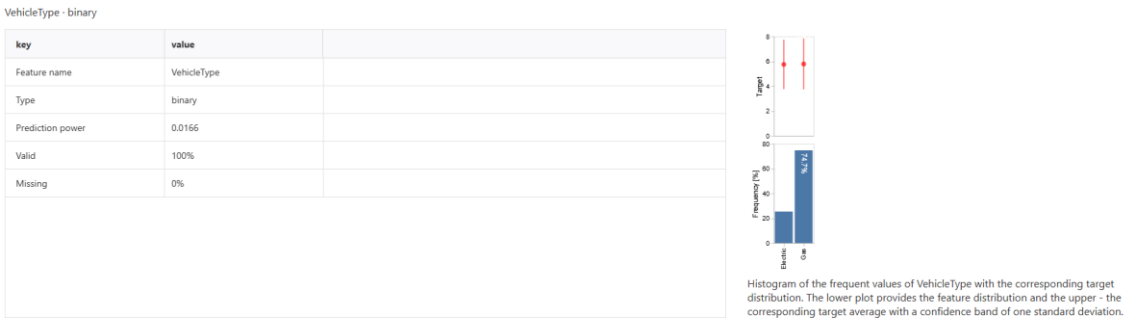
Se observa una relación directa entre la distancia al centro urbano y el tiempo de entrega.

Cuanto más alejado se encuentra el centro de distribución, mayor es la duración total de la ruta.

Esto refleja el impacto de la dispersión geográfica en la eficiencia logística.



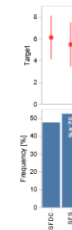
Los conductores con mayor número de incidentes presentan tiempos de entrega más elevados, lo que puede estar relacionado con comportamientos menos eficientes o con mayor número de interrupciones operativas.



El tipo de vehículo presenta una influencia limitada en el tiempo de entrega. Aunque existen ligeras diferencias, su impacto es menor en comparación con factores humanos y operativos.

FulfillmentChannel - binary

key	value	
Feature name	FulfillmentChannel	
Type	binary	
Prediction power	0.0354	
Valid	100%	
Missing	0%	



Histogram of the frequent values of FulfillmentChannel with the corresponding target distribution. The lower plot provides the feature distribution and the upper - the corresponding target average with a confidence band of one standard deviation.

El canal de cumplimiento presenta una influencia reducida sobre el tiempo de entrega, aunque se observan ligeras diferencias entre SFD y SFS, posiblemente relacionadas con la proximidad del inventario al cliente final.

6. Detección de anomalías

El sistema ha identificado algunas entregas atípicas mediante el algoritmo Isolation Forest, principalmente asociadas a combinaciones poco habituales de:

- Alto número de paquetes
- Distancias cortas
- Conductores muy novatos
- Temperaturas extremas

Estas observaciones pueden representar casos excepcionales que conviene analizar de forma individual.

7. Conclusión del análisis

A partir del análisis exploratorio se concluye que:

- El número de paquetes entregados es el principal factor que determina el tiempo de entrega.
- La experiencia del conductor reduce significativamente la duración de las rutas.
- La distancia al centro urbano incrementa directamente el tiempo total de reparto.
- Las condiciones climáticas pueden afectar al rendimiento operativo.
- El tipo de vehículo tiene un impacto limitado frente a los factores humanos y logísticos.

8. Relación con el objetivo de negocio

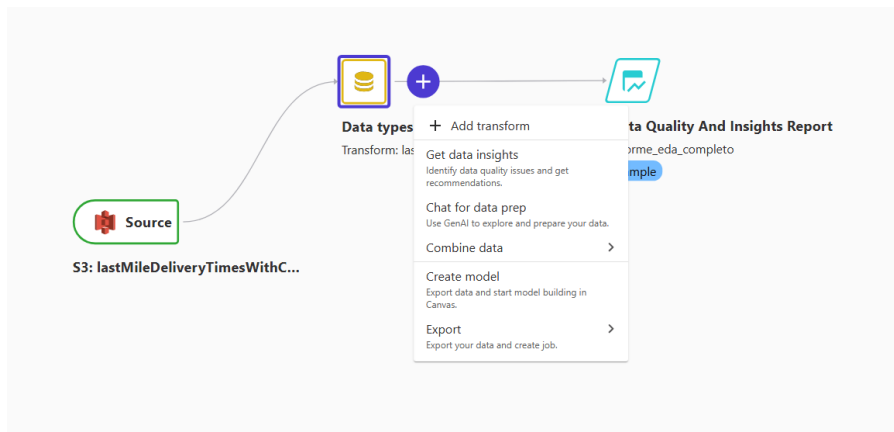
El análisis confirma que la empresa puede optimizar sus operaciones de última milla actuando sobre:

- La asignación de paquetes por conductor
- La planificación de rutas según experiencia
- La selección de centros de distribución más próximos
- La planificación según condiciones climáticas

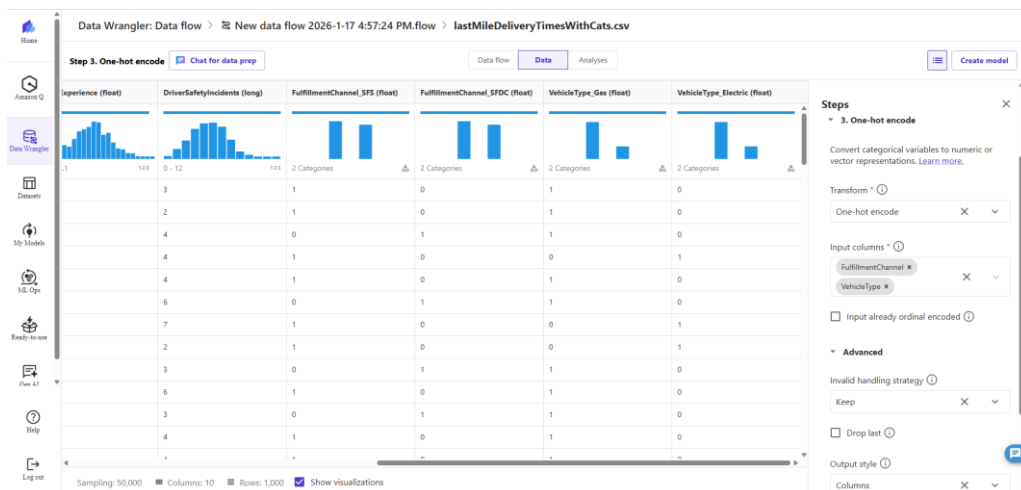
Estos resultados permiten construir un modelo predictivo que ayude a estimar con mayor precisión los tiempos de entrega y mejorar la planificación operativa diaria.

3. Preparación de los datos

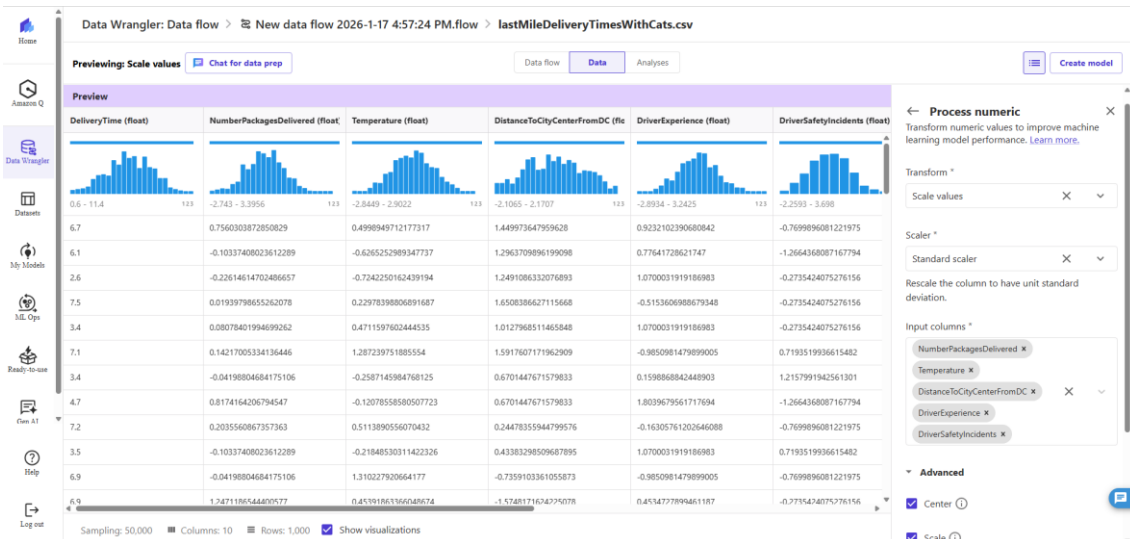
- Volvemos al esquema principal y en Data types vamos a añadir una transformación



1. La primera transformación que vamos a aplicar a nuestros datos va a ser una conversión de tipo de datos, para ello utilizaremos Encode categorical y lo que haremos será convertir los valores de VehicleType y FulfillmentChannel a números.

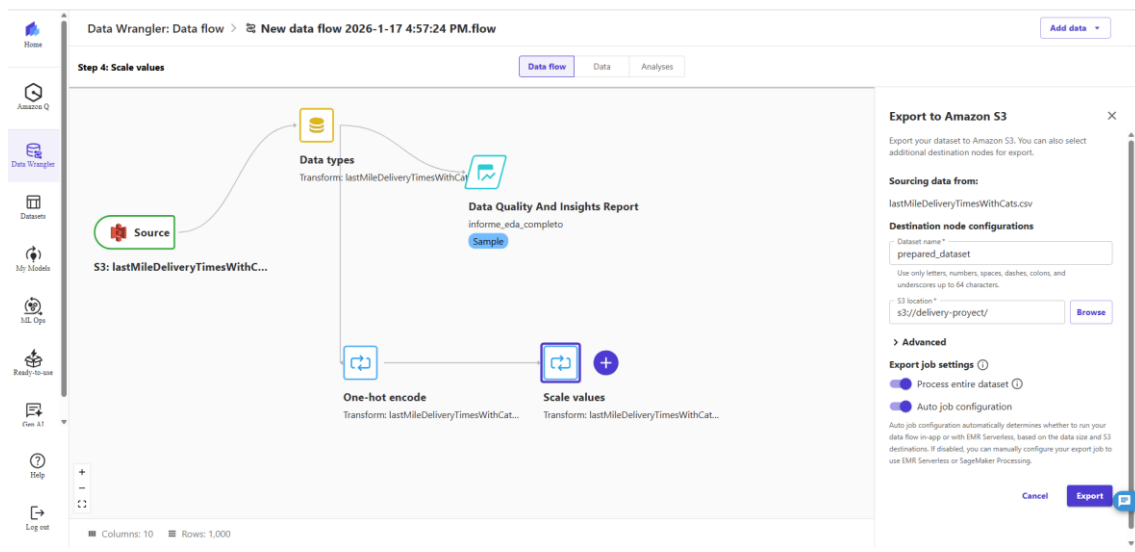


- Las variables categóricas FulfillmentChannel y VehicleType fueron transformadas mediante One-Hot Encoding para evitar introducir relaciones ordinales inexistentes entre categorías y permitir que el modelo las procese correctamente.
-
2. Como segunda transformación vamos a aplicar una normalización mediante Standard Scaling sobre las variables numéricas para evitar sesgos derivados de diferencias de escala y mejorar la estabilidad y rendimiento del modelo predictivo.



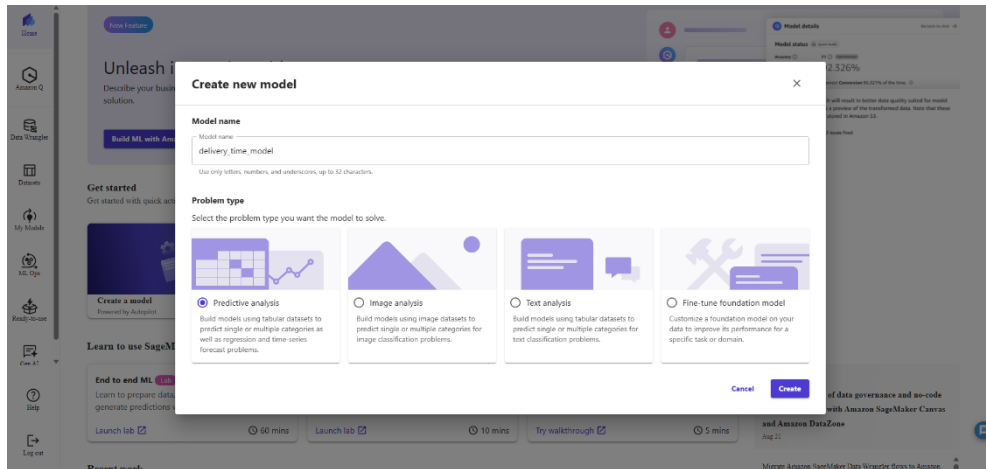
- Las variables numéricas fueron normalizadas mediante Standard Scaling (centrado y escalado) para garantizar que todas las características tengan una escala homogénea, evitando que variables con mayor rango dominen el entrenamiento del modelo.

3. Para finalizar exportaremos a nuestro Bucket de s3 el Dataset preparado:

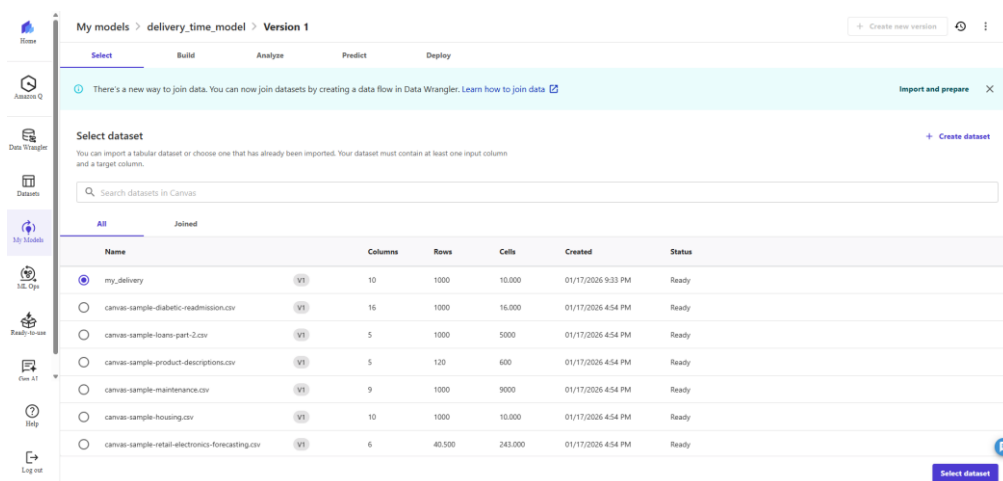


4. Modelado

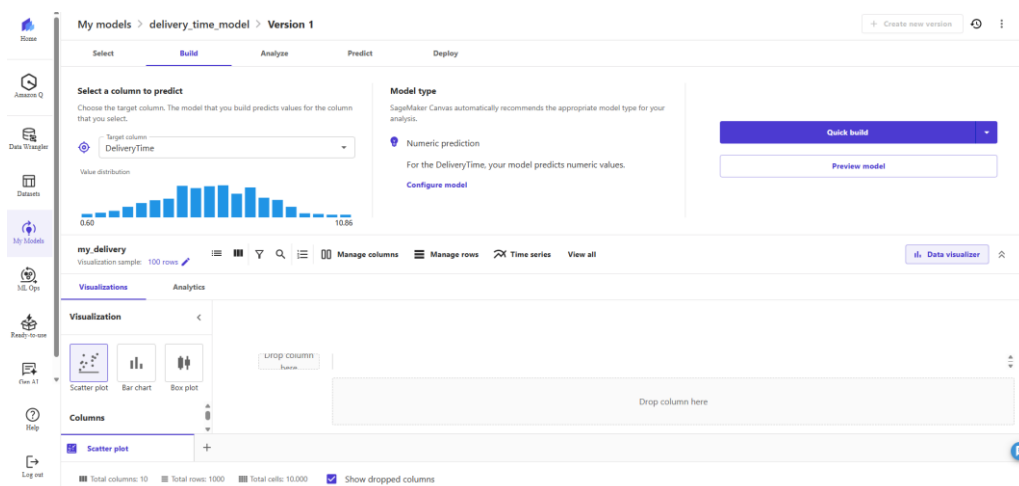
- Iniciamos el modelado con la creación de un modelo de tipo Predictive análisis:



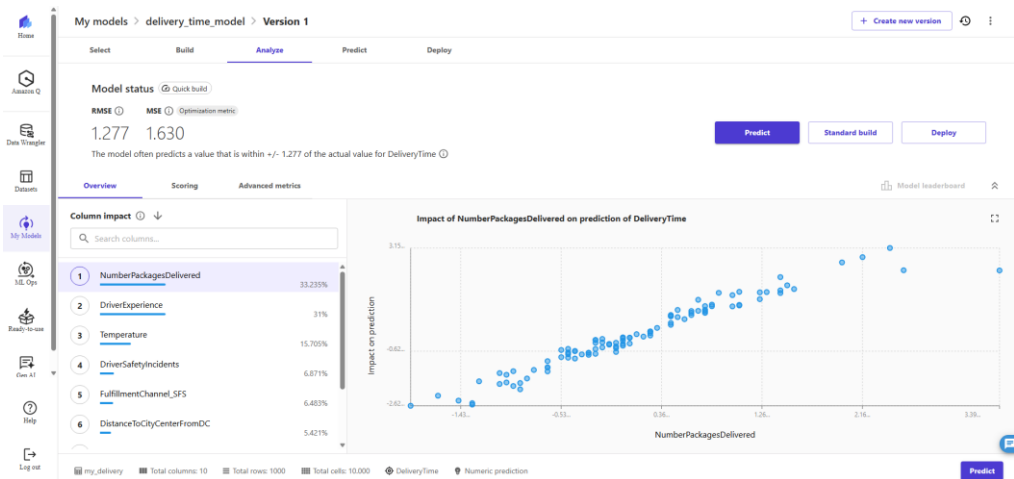
- Creamos un Dataset con el csv creado en el data Wrangler para acceder desde Canva:



- Seleccionamos la columna de DeliveryTime y realizaremos el Quick build que entrenara el modelo con una configuración automática:



- Una vez realizado el modelo entrenado automáticamente puedo decir que es capaz de predecir el tiempo de entrega con un error medio aproximado de 1,27 horas, lo que representa un rendimiento aceptable para un primer modelo de referencia:



- También a continuación voy a realizar una tabla con las variables y su importancia en el modelo:

Variables	Importancia en el modelo
NumberPackagesDelivered	33.2%
DriverExperience	31%
Temperature	15.7%
DriverSafetyIncidents	6.8%
FulfillmentChannel_SFS	6.4%
DistanceToCityCenterFromDC	5.4%
VehicleType_Gas	0.5%
VehicleType_Electric	0.3%
FulfillmentChannel_SFDC	0.3%

- Ahora vamos a hacer un Standard build para realizar un comparativa con el modelo ya realizado:



- Como en el anterior modelo voy a plasmar las variables mas importantes en una tabla:

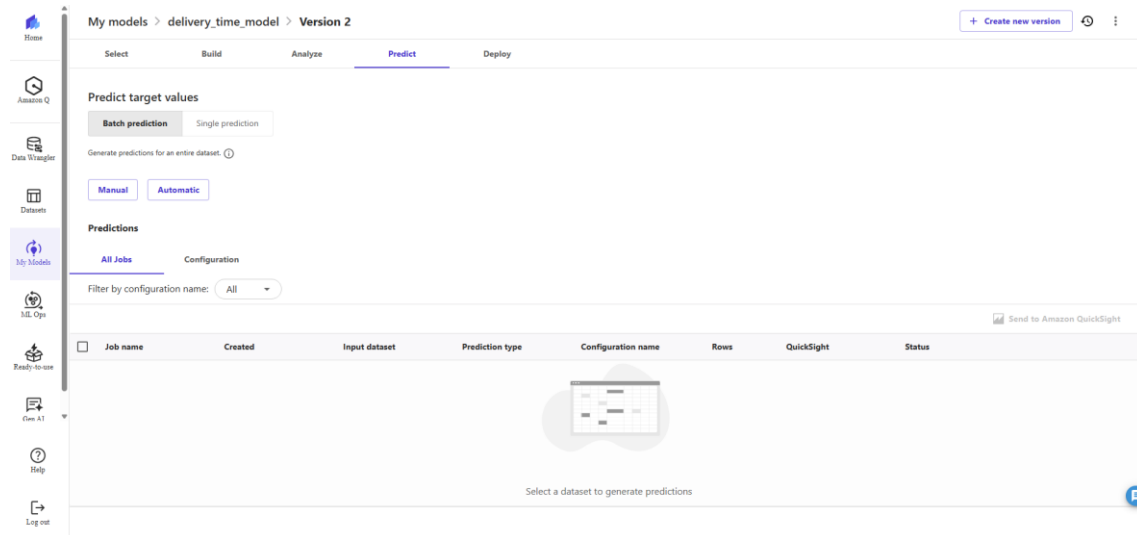
Variables	Importancia en el modelo
NumberPackagesDelivered	36.15%
DriverExperience	32.07%
Temperature	11.63%
DriverSafetyIncidents	7.76%
FulfillmentChannel_SFS	4.20%
DistanceToCityCenterFromDC	4.02%
FulfillmentChannel_SFDC	3.6%
VehicleType_Electric	0.2%
VehicleType_Gas	0.2%

5. Evaluación y conclusión

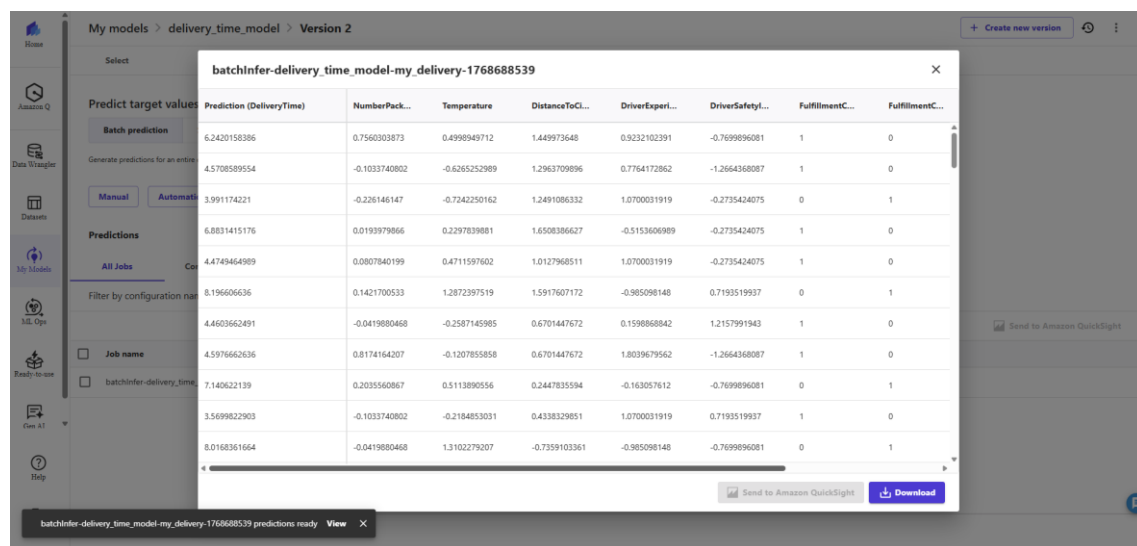
- Después de comprobar los datos de los dos modelos podemos sacar las siguientes conclusiones:
 - El error medio ha sido reducido de 1.277 horas a 1.115 horas
 - Mejora de precisión en un 12.7%
 - Un aumento significativo de la importancia de la variable FulfillmentChannel_SFDC
 - Contra mas paquetes más tiempo de entrega
 - La experiencia mejora el rendimiento
 - La temperatura extrema afecta negativamente
 - Los incidentes reducen la eficiencia
 - El canal de cumplimiento impacta significativamente en los tiempos

6. Despliegue

- Para finalizar realizaremos el despliegue de la Versión 2, es decir, del modelo hecho por el formato estándar, como no poseemos permisos para un deploy normal realizaremos un Batch prediction:



- Una vez creado el job y con el estado en ready lo previsualizamos y si queremos lo descargamos:



- Visualización del Bucket creado para el proyecto con el Dataset base, el Dataset con todas las transformaciones y la predicción con el deploy final

