

Despliegue Completo de un Algoritmo de Machine Learning en AWS SageMaker.

1. Entendimiento del negocio y diseño del Caso de Uso

Nuestro objetivo será desarrollar un modelo de Im que pueda predecir si el equipo azul ganara la partida de este videojuego (League of Legends), utilizando estadísticas de los primeros 10 minutos de cada partida.

Es de vital importancia para los intereses de un equipo competitivo saber si puedes ganar una partida dependiendo de distintas estadísticas que puedas evaluar. En nuestro caso y en la mayoría de los que me he encontrado en Internet, falta calidad en los dataset. Hay muchos detalles en las partidas actualmente que deberían de ser evaluados para poder predecir el análisis de una manera más exacta, pero este era el más acorde para desempeñar esta entrega.

¿Para qué nos puede servir y por qué hemos elegido este tipo de dataset?

Básicamente te permitirá ajustar estrategias en futuras partidas, mejorar sistemas de búsqueda en el juego, dar información valiosa a jugadores.

Si conseguimos predecir después de ver solo los 10 primeros minutos de una partida estaremos logrando generar unas ventajas respecto los demás altísimas.

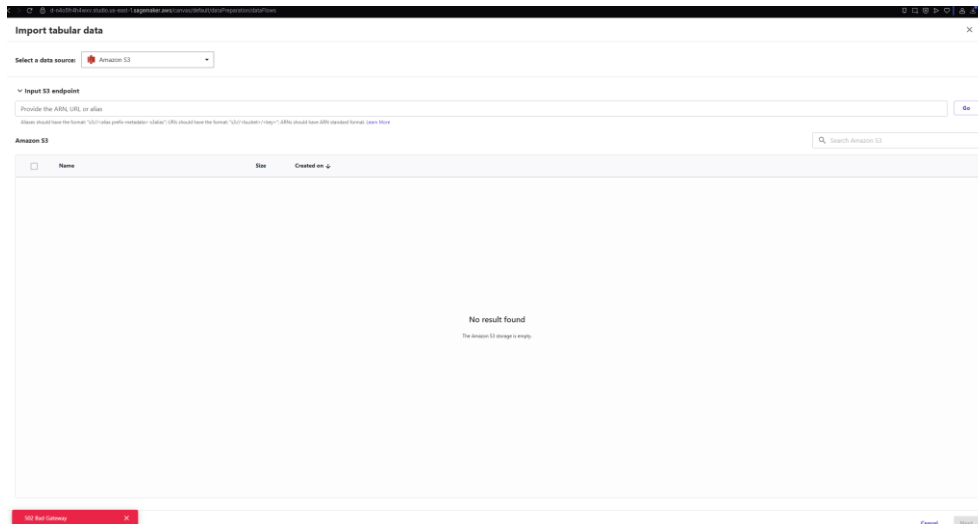
El dataset contiene datos de partidas que se representa cada una en una fila y en cada columna datos como el oro,visión,kills etc.

Utilizaremos como target la columna de blueWins. Esta indica si el equipo azul gana o pierde (1 o 0).

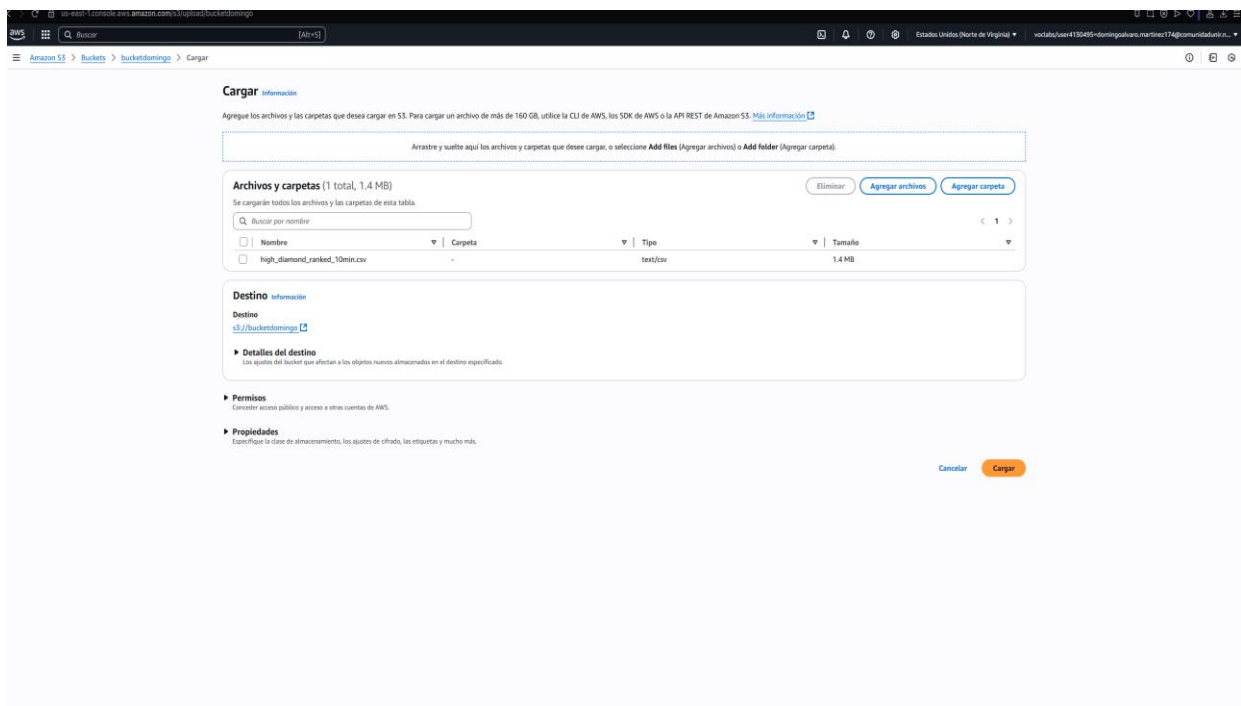
En cuanto a los casos de uso como he comentado anteriormente, cualquier equipo de competitivo podría usar la herramienta para realizar el análisis predictivo y sacar conclusión de sus partidas en los primeros 10 minutos. Si conseguimos que el modelo obtenga buenos resultados la herramienta aportará diferencias competitivas que no podrán ser contrarrestadas.

2. EDA (Exploratory Data Analysis) para obtener una comprensión más profunda.

Empezaremos abriendo el laboratorio para realizar las cargas del dataset desde Amazon s3. En el primer laboratorio de los dos que tuve extras, nos salía este error en distintas ocasiones:



Hay veces que si no tienes permisos de administrador en el acceso a s3 y demás carpetas locales te dará 502 Bad Gateway. En teoría se puede activar desde el administrador del laboratorio, pero no me hizo falta ya que el otro laboratorio que tenía disponible no salía este error y pude cargar perfectamente desde s3.



The image consists of two screenshots from the AWS Management Console.

The top screenshot shows the 'Upload' page for the 'lgbdatasetprojecto' bucket. A green notification bar at the top states: 'Se ha realizado la carga correctamente. Para obtener más información, consulte la tabla Archivos y carpetas.' Below this, the 'Cargar: estado' section shows a message: 'Después de salir de esta página, la siguiente información ya no estará disponible.' The 'Resumen' section indicates the destination is 's3://lgbdatasetprojecto', with 'Realizado correctamente' (1 archivo, 1.4 MB (100.00%)) and 'Con errores' (0 archivos, 0 B (0%)). The 'Archivos y carpetas' tab is active, showing a table with one entry: 'high_diamond_ranked_10min.csv' (1.4 MB, Realizado correctamente).

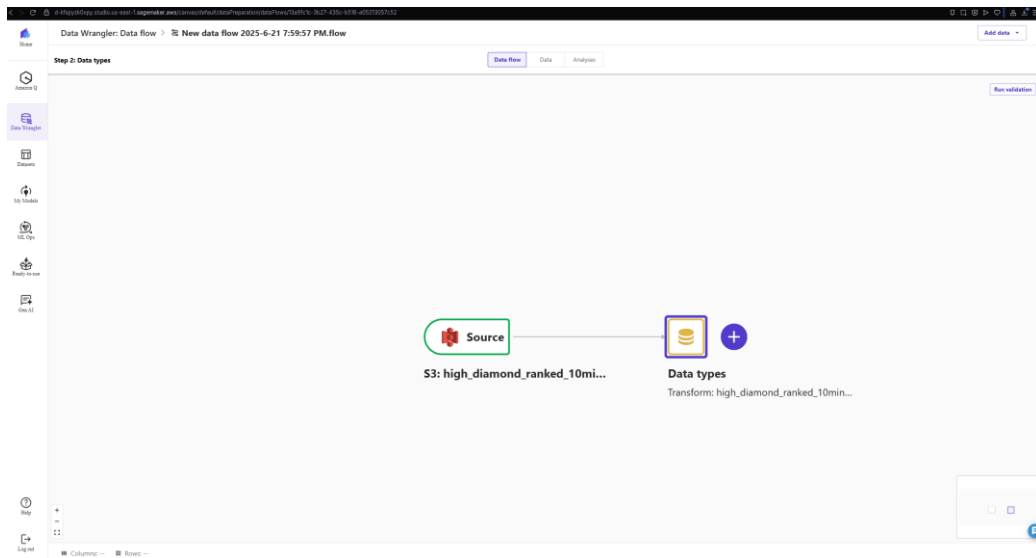
The bottom screenshot shows the 'Import tabular data' configuration page. The 'Select a data source' dropdown is set to 'Amazon S3'. The 'Input S3 endpoint' section has a text input field with the placeholder 'Provide the ARN, URL, or alias' and a 'Go' button. Below this, the 'Amazon S3 / lgbdatasetprojecto' section shows a search bar and a table of files. The table has columns: 'Name', 'Size', and 'Last updated'. One file is listed: 'high_diamond_ranked_10min.csv' (1 MB, 06/21/2025 7:57 PM). At the bottom right, there are 'Cancel' and 'Next' buttons.

Dataset perfectamente cargado desde Amazon s3.

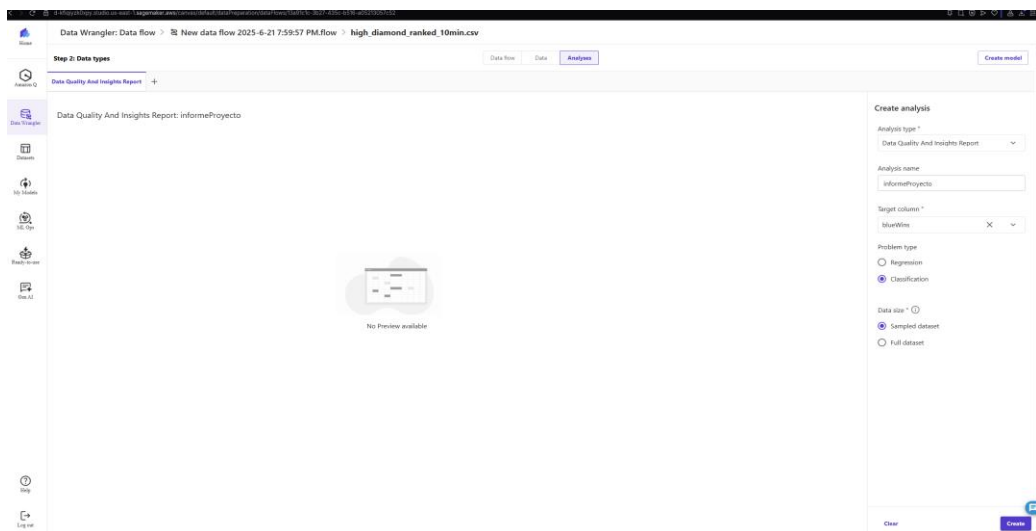
A continuación, seguiremos con la parte de generar los informes.

Utilizaremos Data Wrangler para este proceso.

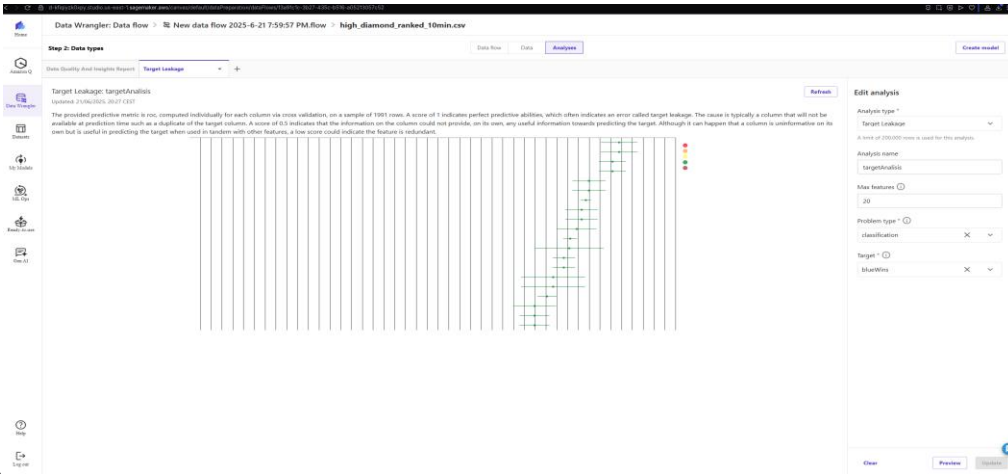
Ya tenemos preparado todo para genera los informes. Dándole al + saldrá la opción.



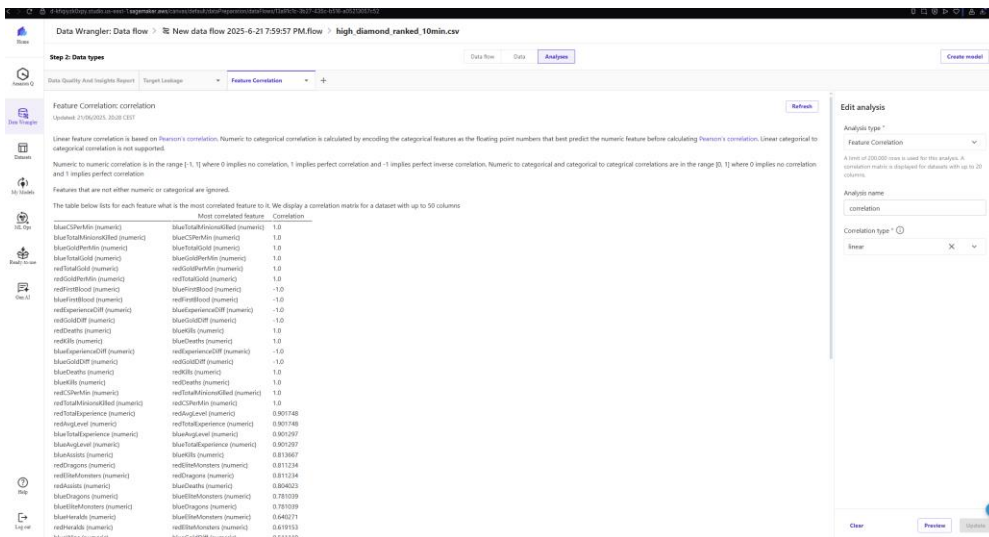
Data Quality para empezar.



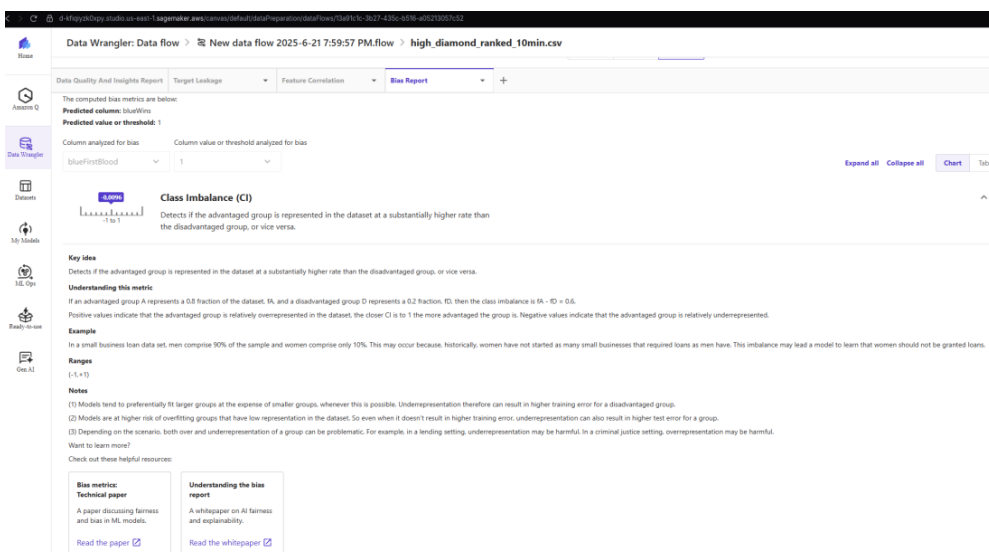
Seguimos con target leak



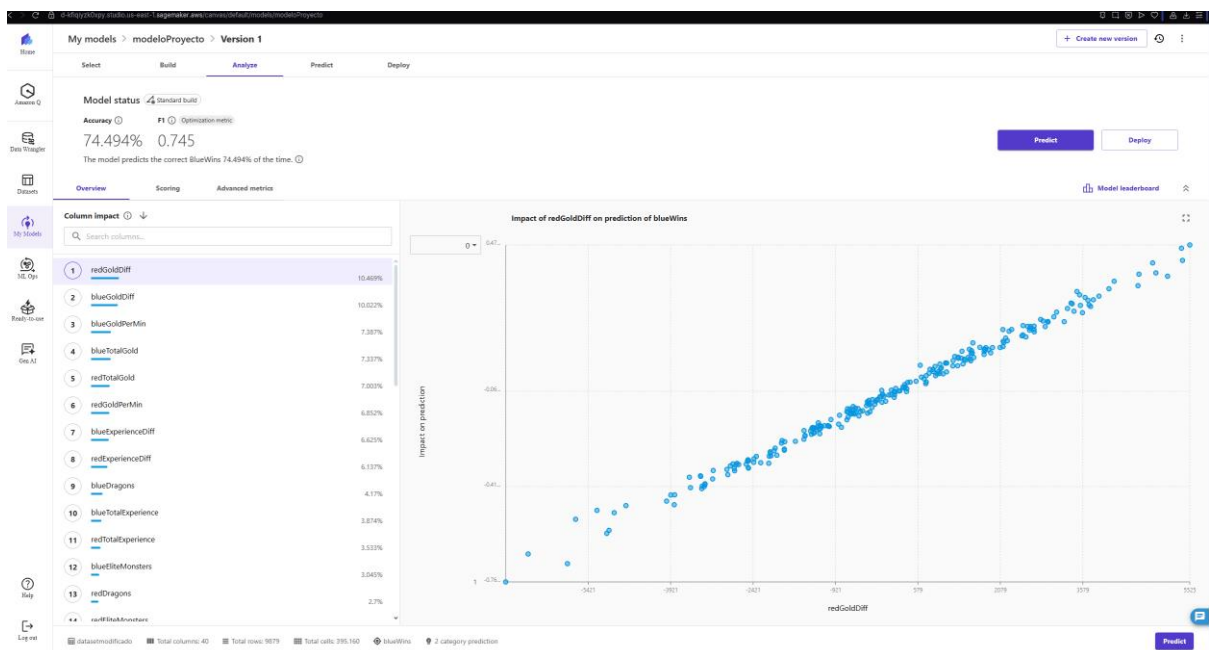
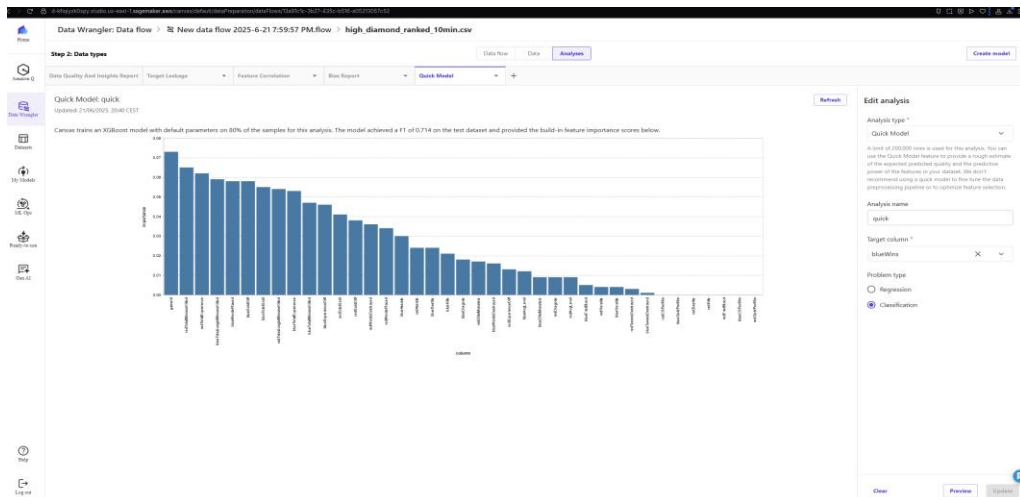
Seguimos con correlación de características.



Informe BIAS.



Terminamos con Quick Model.

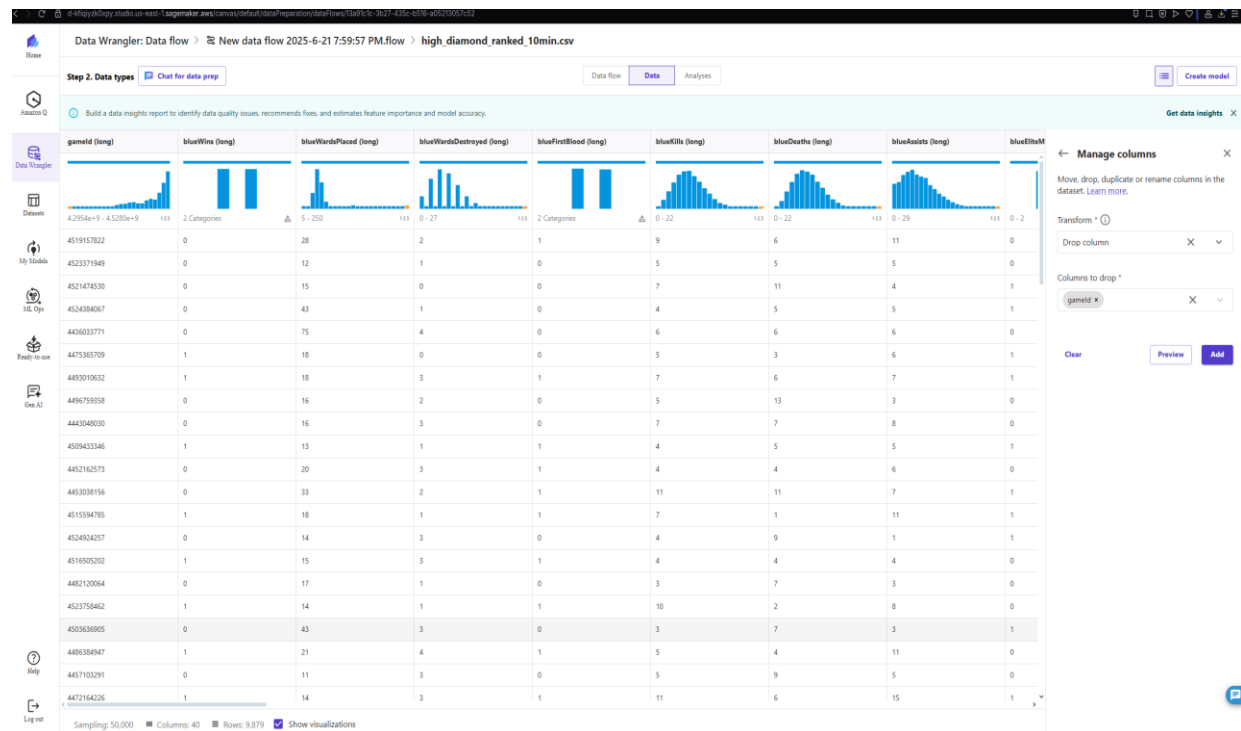


Añado esta captura para ver la correlación de la diferencia de oro en el impacto en predecir blueWins. A medida que la diferencia de oro roja es un fuerte predictor de blueWins.

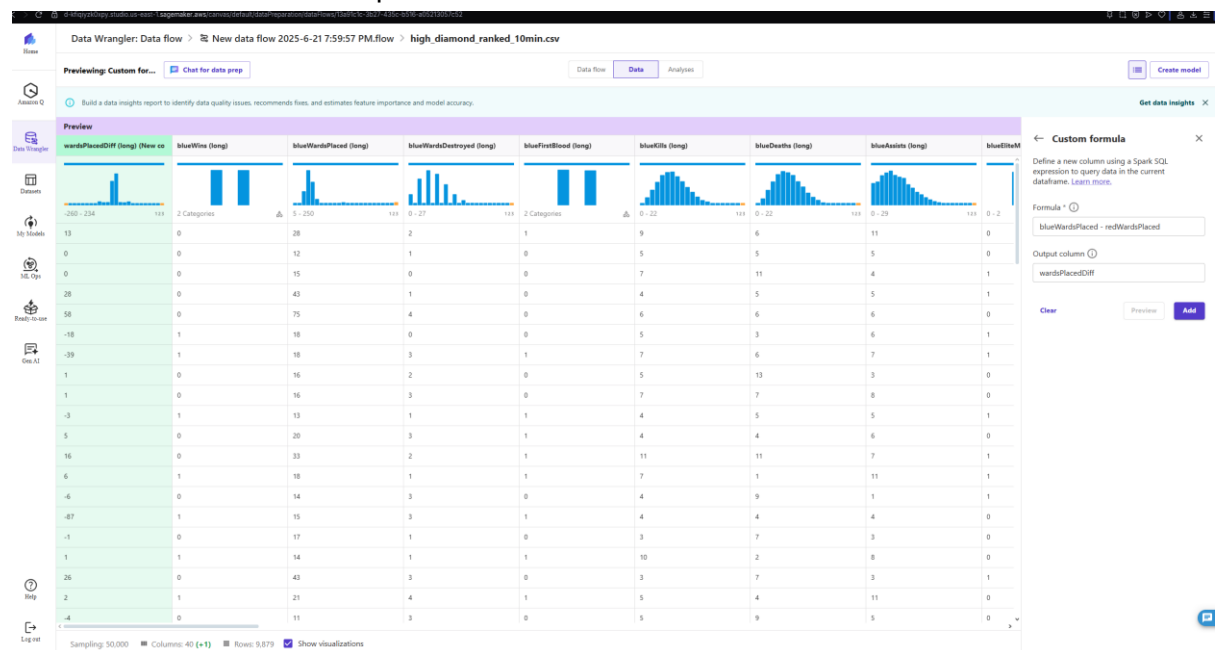
3. Ingeniería de características

Vamos a hacer un par de transformaciones al dataset para facilitar su modelado.

Aplicaremos primero un drop de una columna sin valor realmente como gameld.

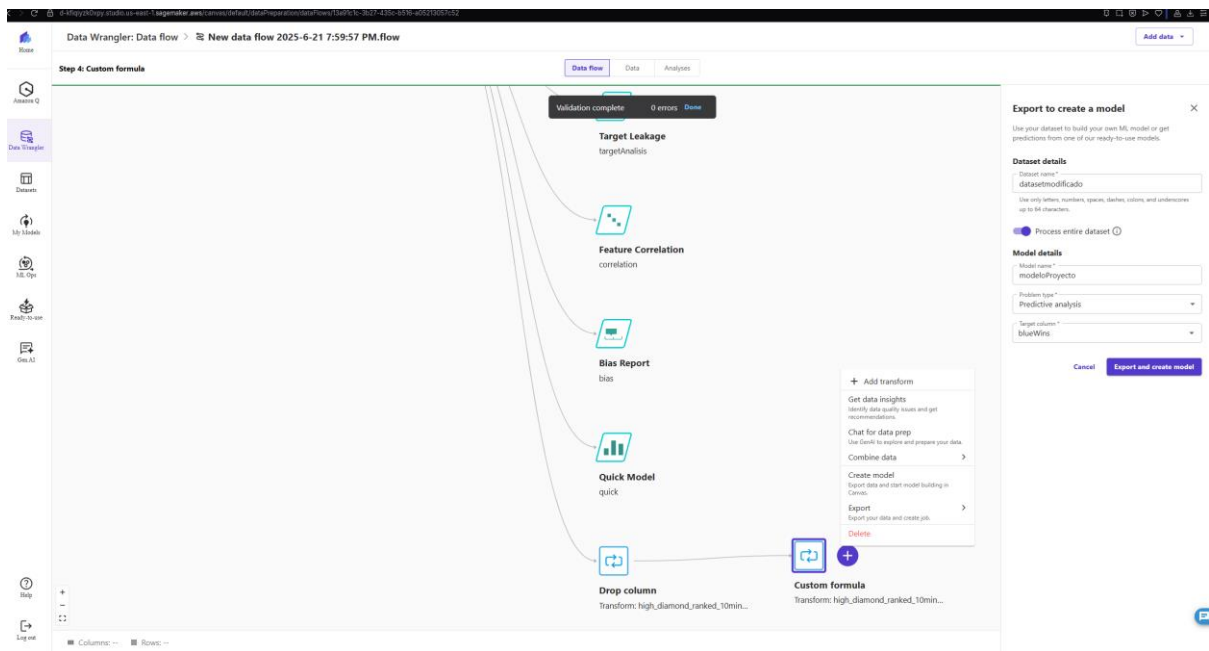


Y ahora seguimos con la creación de una columna adicional que aportara valor añadiendo información importante acerca de la visión.

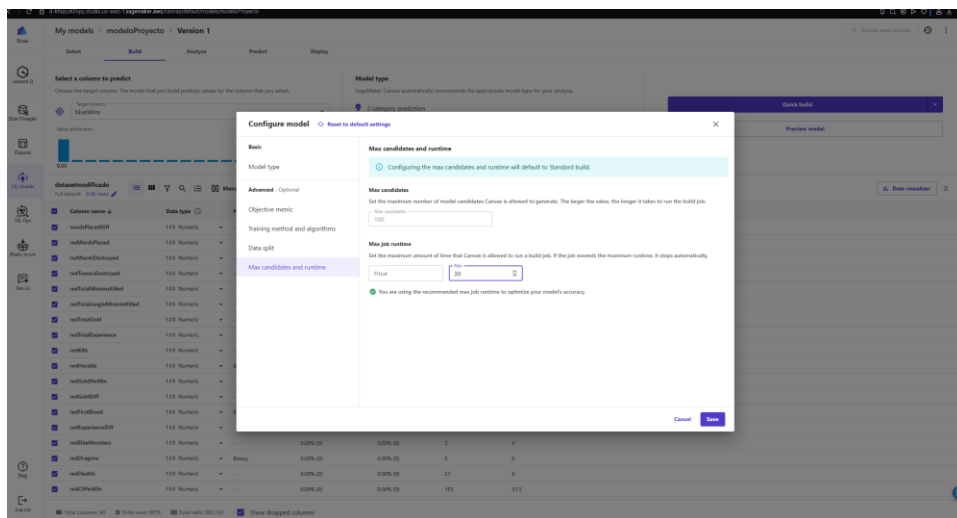


4. Entrenamiento del Modelo en AWS SageMaker.

Empezamos entrenando el modelo a raíz del dataset resultante de hacer la ingeniería de características.



Cambiamos en la pestaña de max candidates and runtime a 30 min y se nos activa el botón de standard build .



Column name	Data type	Feature type	Missing	Mismatched	Unique	Mode
wordCountDiff	123 Numeric	-	0.00%	0.00%	284	0
wordCountProd	123 Numeric	-	0.00%	0.00%	151	15
wordCountDestroyed	123 Numeric	-	0.00%	0.00%	25	2
wordCountDestroyed	123 Numeric	-	0.00%	0.00%	3	0
wordCountDestroyed	123 Numeric	-	0.00%	0.00%	103	215
wordCountDestroyed	123 Numeric	-	0.00%	0.00%	75	52
wordCountDestroyed	123 Numeric	-	0.00%	0.00%	4732	18,076
wordCountDestroyed	123 Numeric	-	0.00%	0.00%	4732	17,860
wordCountDestroyed	123 Numeric	-	0.00%	0.00%	20	5
wordCountDestroyed	123 Numeric	Binary	0.00%	0.00%	2	0
wordCountDestroyed	123 Numeric	-	0.00%	0.00%	4732	18,074
wordCountDestroyed	123 Numeric	-	0.00%	0.00%	8947	405
wordCountDestroyed	123 Numeric	Binary	0.00%	0.00%	2	0
wordCountDestroyed	123 Numeric	-	0.00%	0.00%	1096	40
wordCountDestroyed	123 Numeric	-	0.00%	0.00%	3	0
wordCountDestroyed	123 Numeric	Binary	0.00%	0.00%	2	0
wordCountDestroyed	123 Numeric	-	0.00%	0.00%	21	6
wordCountDestroyed	123 Numeric	-	0.00%	0.00%	103	215

5. Evaluación y Registro del Modelo

Aquí mostramos una captura de BIAS en sustitución a lo de Clarify.

Data Wrangler: Data flow > **New data flow 2025-6-21 7:59:57 PM.flow** > **high_diamond_ranked_10min.csv**

Bias Report

The computed bias metrics are below:
Predicted column: blueWins
Predicted value or threshold: 1

Column analyzed for bias: blueFirstBlood
 Column value or threshold analyzed for bias: 1

Class Imbalance (CI)
 Detects if the advantaged group is represented in the dataset at a substantially higher rate than the disadvantaged group, or vice versa.

Key idea
 Detects if the advantaged group is represented in the dataset at a substantially higher rate than the disadvantaged group, or vice versa.

Understanding this metric
 If an advantaged group A represents a 0.8 fraction of the dataset, 8A, and a disadvantaged group D represents a 0.2 fraction, 8D, then the class imbalance is 8A - 8D = 0.6.
 Positive values indicate that the advantaged group is relatively overrepresented in the dataset, the closer CI is to 1 the more advantaged the group is. Negative values indicate that the advantaged group is relatively underrepresented.

Example
 In a small business loan data set, men comprise 90% of the sample and women comprise only 10%. This may occur because, historically, women have not started as many small businesses that required loans as men have. This imbalance may lead a model to learn that women should not be granted loans.

Ranges
 [-1, +1]

Notes
 (1) Models tend to preferentially fit larger groups at the expense of smaller groups, whenever this is possible. Underrepresentation therefore can result in higher training error for a disadvantaged group.
 (2) Models are at higher risk of overfitting groups that have low representation in the dataset. So even when it doesn't result in higher training error, underrepresentation can also result in higher test error for a group.
 (3) Depending on the scenario, both over and underrepresentation of a group can be problematic. For example, in a lending setting, underrepresentation may be harmful. In a criminal justice setting, overrepresentation may be harmful.
 Want to learn more?
 Check out these helpful resources:

Bias metrics:
 Technical paper
 A paper discussing fairness and bias in ML models.
[Read the paper](#)

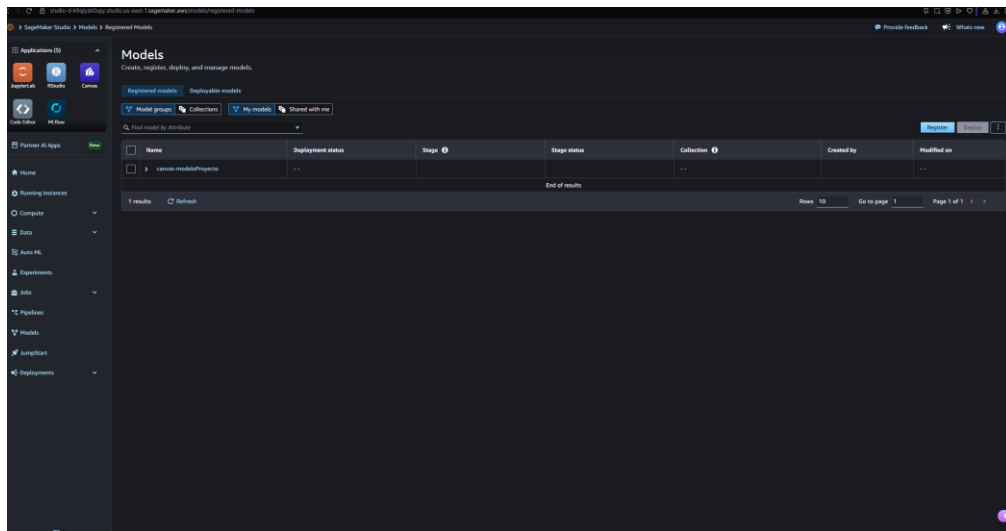
Understanding the bias report
 A whitepaper on AI fairness and explainability.
[Read the whitepaper](#)

Y una vez ya tenemos las dos versiones disponibles registramos la mejor.

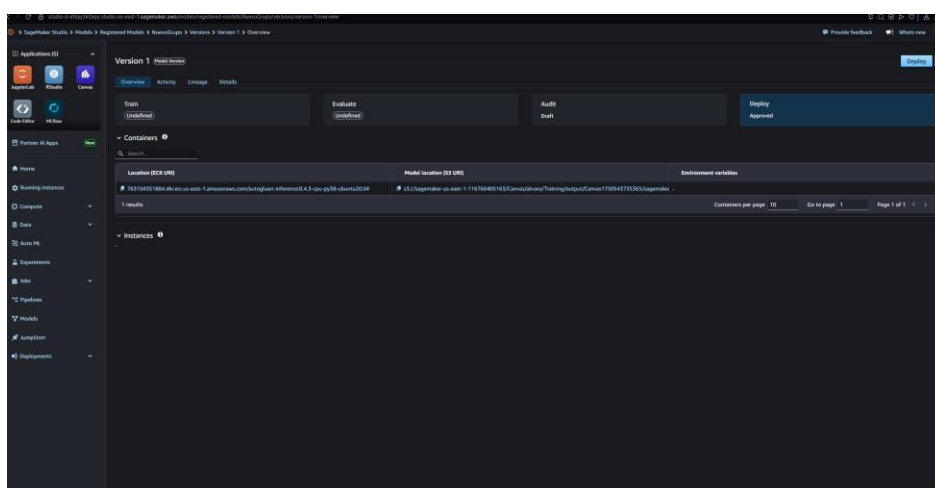
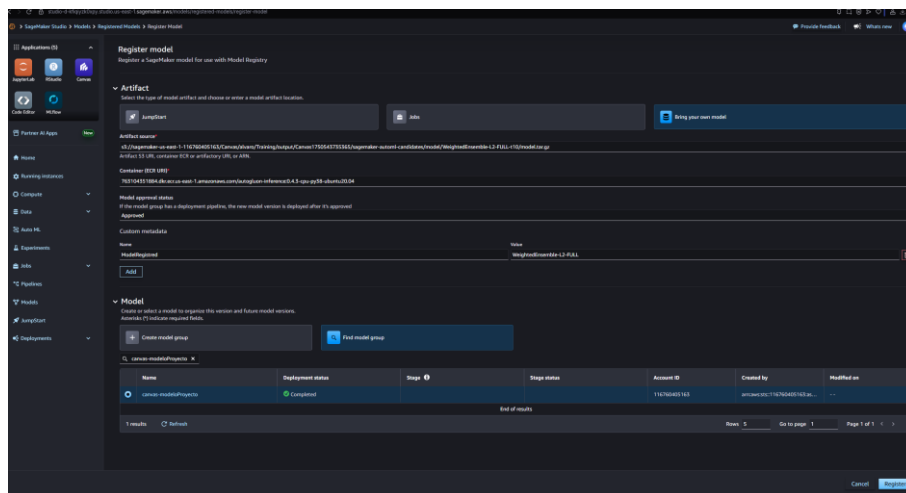
Version	Status	Build type	Created	Dataset	Accuracy	F1	Precision	Recall	AUC-ROC	Model Registry
001	Ready	Standard	06/22/2025...	dataset...	74.14%	0.744	73.784%	73.011%	0.817	Not Registered
002	Ready	Standard	06/22/2025...	dataset...	74.49%	0.795	74.885%	74.247%	0.817	Registered

Ademas de hacer esto desde aquí, hemos utilizado la herramienta de SageMaker Model Registry para registrar versiones para hacer el registro del modelo.

Lo primero es ir a SageMaker Studio.



Una vez aquí hace falta rellenar los campos con sus valores de s3 correspondientes a tu modelo.



Registro finalizado.

6. Despliegue del Modelo en AWS SageMaker

Vamos a realizar primero la parte de la inferencia. Añadimos el dataset inferencia sin la columna objetivo para realizar bien este apartado.

dataset_inferencia : Create Tabular dataset

Select a data source: Local upload

Upload files to import

1 file ready to import
dataset_inferencia.csv

Import preview

Previewing first 100 rows

Close preview

Create dataset

If your data has special character delimiters, import your data into a Data Wrangler data flow and use the advanced import settings to specify a custom delimiter. [Learn More](#)

dataset_inferencia.csv

Use first row as header

Delete

blueWardPits...	blueWardDes...	blueFirstBlood	blueKills	blueDeaths	blueAssists	blueEpicMon...	blueDragons	blueHerolds
20	2	1	9	6	11	0	0	0
12	1	0	5	5	5	0	0	0
15	0	0	7	11	4	1	1	0
43	1	0	4	5	5	1	0	1
75	4	0	6	6	6	0	0	0
18	0	0	5	3	6	1	1	0

Select dataset for predictions

To make predictions on a dataset, select it or import it. The dataset that you select must have the same number of feature columns as the training dataset.

Search datasets in Canvas

Name	Columns	Rows	Cells	Created	Status
dataset_inferencia	39	9879	395,281	06/22/2025 12:24 PM	Ready
New dataset 2025-6-22 2:30:01 AM	41	9879	405,039	06/22/2025 2:32 AM	Ready
datasetFinal	40	9879	395,160	06/22/2025 12:02 AM	Ready
canvas sample housing.csv	10	1000	10,000	06/21/2025 7:55 PM	Incompatible
canvas sample retail-electronics-forecasting.csv	6	40,000	240,000	06/21/2025 7:55 PM	Incompatible
canvas sample loans-part-0.csv	5	1000	5000	06/21/2025 7:55 PM	Incompatible
canvas sample shipping-logs.csv	12	1000	12,000	06/21/2025 7:55 PM	Incompatible
canvas sample starbucks-daily-15k.csv	2	2000	4000	06/21/2025 7:55 PM	Incompatible
canvas sample stockprices.csv	9	1000	9000	06/21/2025 7:55 PM	Incompatible
canvas sample diabetic-medication.csv	16	1000	16,000	06/21/2025 7:55 PM	Incompatible
canvas sample product-descriptions.csv	5	100	600	06/21/2025 7:55 PM	Incompatible
canvas sample loans-part-1.csv	19	1000	19,000	06/21/2025 7:55 PM	Incompatible

Y, por último, realizamos la inferencia y vemos el dataset resultado que hemos cargado en el apartado de datasets.

Datasets > datasetprediccioninferencia

Create a data flow Update dataset Create a model

Dataset details

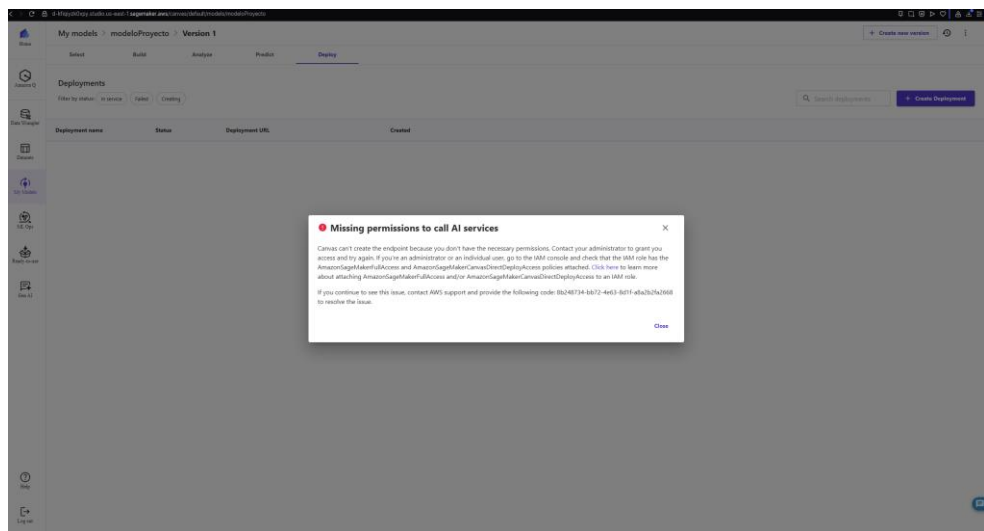
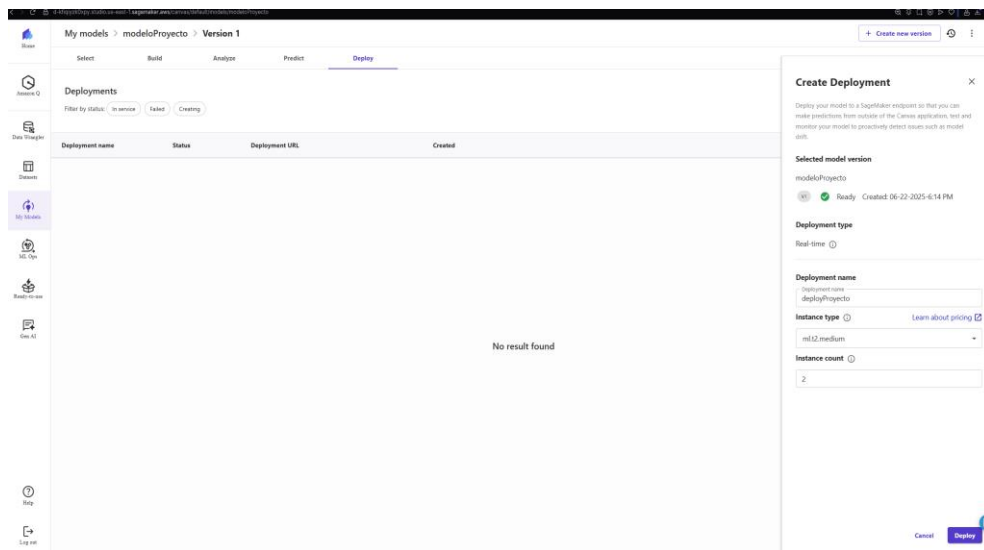
datasetprediccioninferencia

Previewing up to the first 100 rows of datasetprediccioninferencia

blueWin	probability	blueWardPlaced	blueWardDes...	blueFirstBlood	blueKills	blueDeaths	blueAssists	blueEpicMonsters	blueDragons	blueHerolds
1	0.5941355287	28	2	1	9	6	11	0	0	0
0	0.8302741051	12	1	0	5	5	5	0	0	0
0	0.6509468054	15	0	0	7	11	4	1	1	0
0	0.6373567913	43	1	0	4	5	5	1	0	1
0	0.6487622728	75	4	0	6	6	6	0	0	0
1	0.6217468001	18	0	0	5	3	6	1	1	0
1	0.8504031195	16	3	1	7	6	7	1	1	0
0	0.8524041176	16	2	0	5	12	3	0	0	0
0	0.7750232412	16	3	0	7	7	8	0	0	0
0	0.6118074973	15	1	1	4	5	5	1	1	0
0	0.638027608	20	3	1	4	4	6	0	0	0
0	0.7051540613	33	2	1	11	11	7	1	0	1
1	0.916285426	16	1	1	7	1	10	1	1	0
0	0.82382152	14	3	0	4	9	1	1	0	1
0	0.5198069977	15	3	1	4	4	4	0	0	0
0	0.783672568	17	1	0	3	7	3	0	0	0
1	0.9443677007	14	1	1	10	2	8	0	0	0

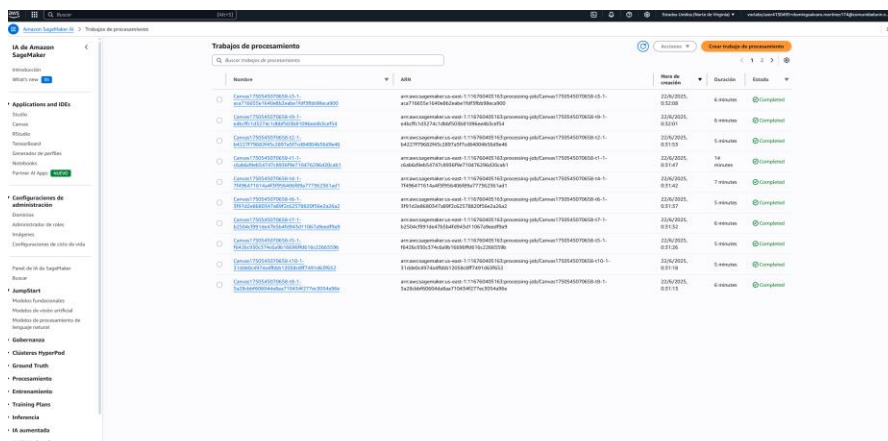
Dataset type: Tabular Total dataset cells (columns x rows): 40539 (41 x 9879) Data source: Local

Y para la parte del deploy obtenemos un error ya que no podemos realizarlo aun así mostramos la captura que nos muestra cuando le damos al deploy.

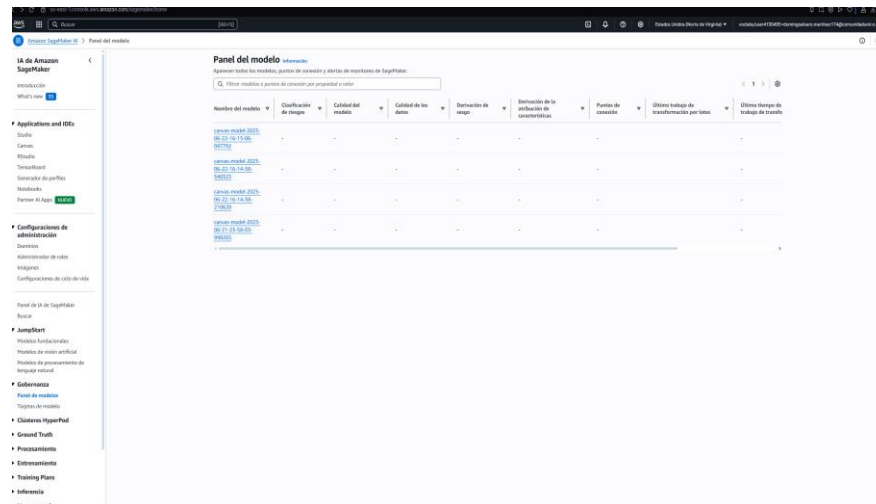


7. Implementar estrategias de monitorización del modelo en producción

Vamos a empezar enseñando algunas capturas de la zona de SageMaker estudio para ver el modelo y sus distintas versiones etc.

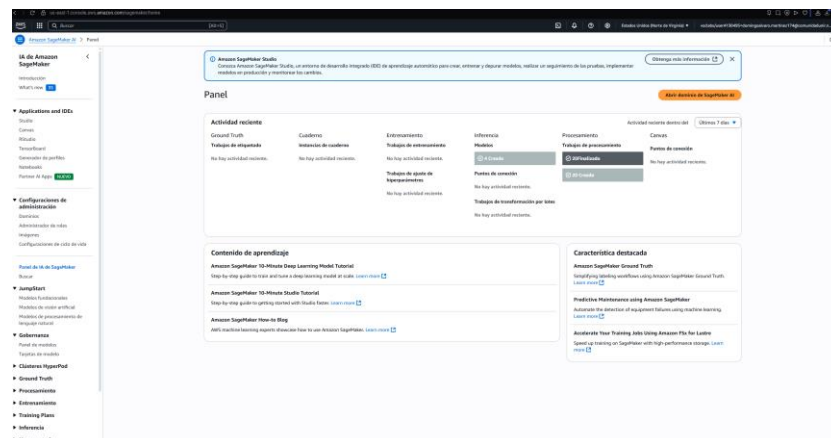


Aquí monitoreamos los trabajos de procesamiento de datos.



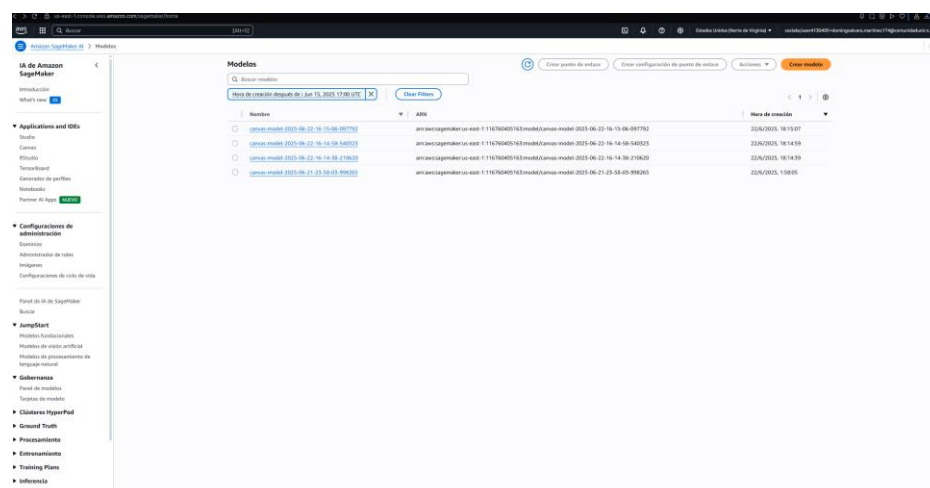
Nombre del modelo	Clasificación de riesgos	Calidad del modelo	Calidad de los datos	Derivación de tiempo	Derivación de la calidad de las estadísticas	Puntos de conexión	Último trabajo de transformación por lotes	Último tiempo de trabajo de trabajo
carson-model-2025-06-22-16-15-08-087792	-	-	-	-	-	-	-	-
carson-model-2025-06-22-16-15-08-087792	-	-	-	-	-	-	-	-
carson-model-2025-06-22-16-15-08-087792	-	-	-	-	-	-	-	-
carson-model-2025-06-22-16-15-08-087792	-	-	-	-	-	-	-	-
carson-model-2025-06-22-16-15-08-087792	-	-	-	-	-	-	-	-
carson-model-2025-06-22-16-15-08-087792	-	-	-	-	-	-	-	-

En este caso veríamos el panel donde se listan los modelos.



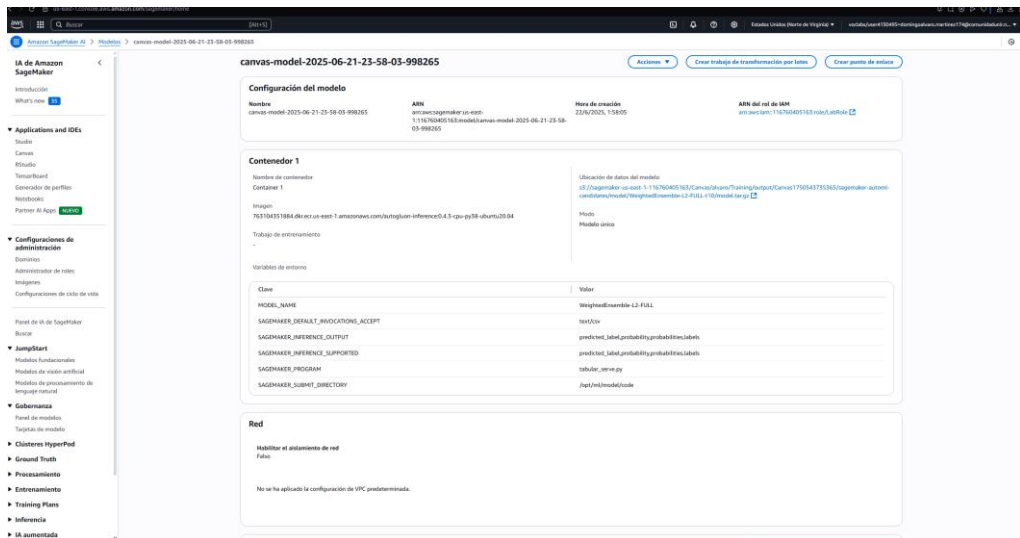
Actividad reciente	Contenido de aprendizaje	Características destacadas
Actividad reciente Ground Truth Trabajo de etiquetado No hay actividad reciente. Cualidades Indicadores de cualidades No hay actividad reciente. Entrenamientos Trabajo de entrenamiento No hay actividad reciente. Trabajo de ajuste de hiperparámetros No hay actividad reciente. Inferencia Modelos Puntos de conexión Trabajo de transformación por lotes No hay actividad reciente. Procesamiento Trabajo de procesamiento Puntos de conexión Trabajo de transformación por lotes No hay actividad reciente. Cancas Puntos de conexión No hay actividad reciente.	Contenido de aprendizaje Amazon SageMaker Studio Step-by-step guide to create and train a deep learning model in SageMaker Studio. Amazon SageMaker 10-Minute Studio Tutorial Step-by-step guide to getting started with SageMaker Studio. Amazon SageMaker How-to Blog AWS Machine Learning experts describe how to use Amazon SageMaker.	Características destacadas Amazon SageMaker Ground Truth Simplify labeling workflows using Amazon SageMaker Ground Truth. Predictive Maintenance using Amazon SageMaker Automate the detection of equipment failure using machine learning. Accelerate Your Training with SageMaker HyperPod for Lambda Speed up training on SageMaker with high-performance storage.

Este sería el panel de SageMakerla.



Nombre	ARN	Hora de creación
carson-model-2025-06-22-16-15-08-087792	arn:aws:sagemaker:us-east-1:116760405763:model:carson-model-2025-06-22-16-15-08-087792	22/6/2025, 16:15:07
carson-model-2025-06-22-16-15-08-087792	arn:aws:sagemaker:us-east-1:116760405763:model:carson-model-2025-06-22-16-15-08-087792	22/6/2025, 16:14:59
carson-model-2025-06-22-16-15-08-087792	arn:aws:sagemaker:us-east-1:116760405763:model:carson-model-2025-06-22-16-15-08-087792	22/6/2025, 16:14:59
carson-model-2025-06-22-16-15-08-087792	arn:aws:sagemaker:us-east-1:116760405763:model:carson-model-2025-06-22-16-15-08-087792	22/6/2025, 16:14:59

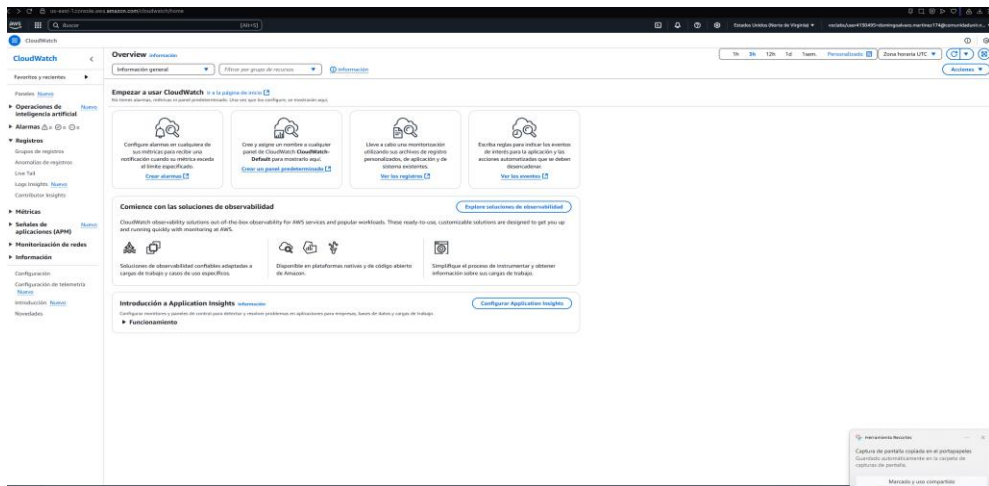
Aquí los modelos creados.



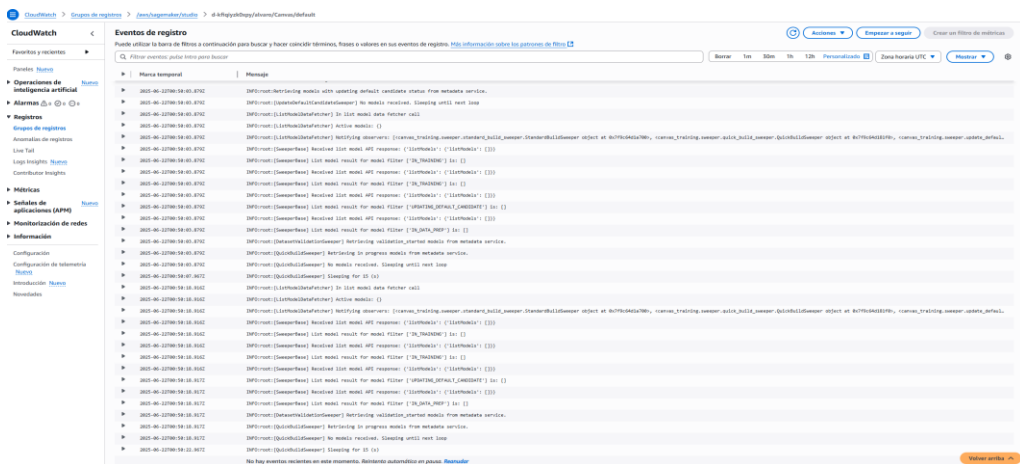
Y esta la vista de uno de los modelos del panel.

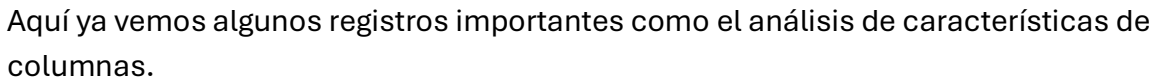
En cuanto a las métricas de CloudWatch adjuntaremos algunas capturas de pantalla.

Empezaremos entrando al menú de CloudWatch.



Aquí ya entramos a la parte de eventos de registro.





Tenemos un modelo con unos resultados bastante buenos que mejorarían a niveles insospechados si tuviéramos tanto más tiempo y recursos para hacerlo como por supuesto un dataset con columnas de más relevancia que las que tiene este mismo. Hoy, no son muchos los equipos que tienen analistas de datos con IA sino que simplemente usan analistas de graficas o datos de manera básica. Esta en crecimiento y cada día más equipos lo implementan, pero aún falta desarrollar más herramientas.