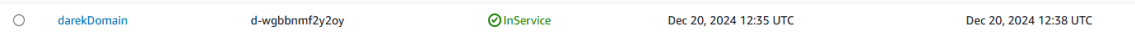


Step 1: set up your enviornment.

- Create and configurate a dominion using amazon SageMaker

In this case mine is already prepared:



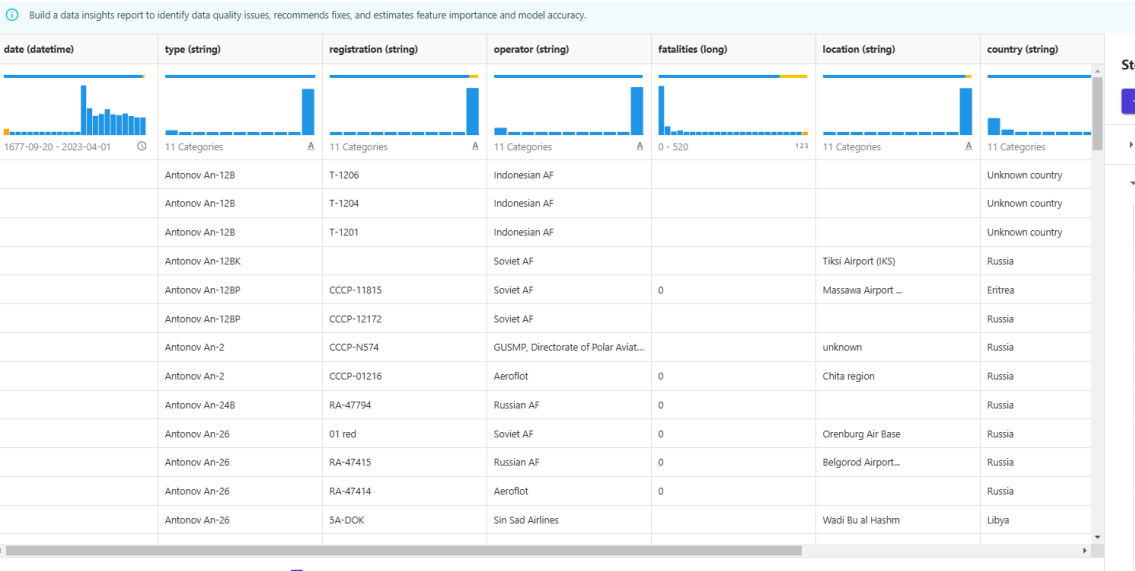
Launch the jupyter studio through the user previously created in your domain

- access Data wrangler through the left sidebar, open it through Amazon SagemakerCanvas and import your csv.

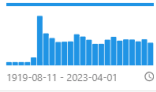
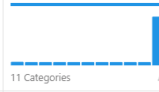

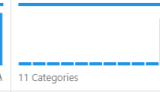

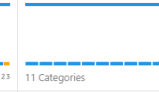

We'll be using this dataset: <https://www.kaggle.com/datasets/warcoder/civil-aviation-accidents>

Step 2: exploratory analysis

you'll see a preview of your dataset, worth mentioning amazon canvas already identifies by itself the datatypes of each row.



Click and select “get data insights” to perform an in depth analysis of the set in search for empty values or anything that can harm the analysis, ive transformed the dataset to delete those empty values.

date (datetime)	type (string)	registration (string)	operator (string)	fatalities (long)	location (string)	country (string)
						
1919-08-11 00:00:00	Felixstowe Fury	N123	RAF	1	near Felixstowe RNAS	U.K.
1920-02-23 00:00:00	Handley Page O/7	G-EANV	Handley Page Transport	0	Acadia Siding, C...	South Africa
1920-02-25 00:00:00	Handley Page O/400	G-EAMC	Handley Page Transport	0	near El Shereik	Sudan
1920-06-30 00:00:00	Handley Page O/400	G-EAKE	Handley Page Transport	0	ÄdöstanÅv	Sweden
1920-12-14 00:00:00	Handley Page O/400	G-EAMA	Handley Page Transport	4	Golders Green	U.K.
1921-03-02 00:00:00	Handley Page O/7	G-IAAC	HP Indo-Burmese Transport	0	Meerut	India
1921-08-26 00:00:00	Farman F.60 Goliath	O-BLAN	SNETA	2	near Calais (The Engl...	France
1921-09-27 00:00:00	Farman F.60 Goliath	O-BRUN	SNETA	0	Evere Airfield	Belgium
1921-09-27 00:00:00	Farman F.60 Goliath	O-BLEU	SNETA	0	Evere Airfield	Belgium
1922-01-22 00:00:00	Handley Page O/10	G-EATN	Handley Page Transport	0	near Senlis	France
1922-04-07 00:00:00	Farman F.60 Goliath	F-GEAD	Grands Express AÃ©riens	5	Thieuloy-Saint-A...	France

Now the insights show the prediction potential and chart stats of the dataset correctly.

Dataset statistics

Key	Value	Feature type	Count
Number of features	9	numeric	1
Number of rows	23917	categorical	2
Missing	3.88%	text	4
Valid	96.1%	datetime	1
Duplicate rows	0.623%	binary	0
		unknown	0

High Priority Warnings

1 high severity warnings were detected. See the list below.

Target leakage High

The feature date predicts the target extremely well on it's own. A feature this predictive often indicates an error called target leakage. The cause is typically data that is not available at time of prediction. For example, a duplicate of the target column in the dataset can result in target leakage.

Alternatively, if the machine learning task is 'easy', then a single feature can have legitimately high prediction power. If you think that a single feature is very highly predictive, you don't need to do anything further. However, if you think there's target leakage, we recommended that you remove the highly predictive column from the dataset using the Drop column transform under Manage columns.

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 l
- 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
date	0.499	datetime	97.6%	2.37%	1	0
type	0.162	text	100%	0%	0	0
registration	0.143	text	100%	0.0167%	0	0
operator	0.143	text	96%	4.01%	0	0

As an additional step in this case, im performing an histogram to see how many fatalities have occurred in aviation accidents per year.

