Amazon SageMaker Autopilot Data Exploration

This report provides insights about the dataset you provided as input to the AutoML job. It was automatically generated by the AutoML training job: automl-dm-1624877542.

As part of the AutoML job, the input dataset was randomly split into two pieces, one for training and one for validation. The training dataset was randomly sampled, and metrics were computed for each of the columns. This notebook provides these metrics so that you can:

1. **Understand how the job analyzed features to select the candidate pipelines.**
2. **Modify and improve the generated AutoML pipelines using knowledge that you have about the dataset.**

We read `7110` rows from the training dataset. The dataset has `2` columns and the column named `sentiment` is used as the target column. This is identified as a `MulticlassClassification` problem. Here are 3 examples of labels: `['-1', '1', '0']`.

Suggested Action Items - Look for sections like this for recommended actions that you can take.

Contents

Dataset Sample

The following table is a random sample of 10 rows from the training dataset.

Suggested Action Items - Verify the input headers correctly align with the columns of the dataset sample. If they are incorrect, update the header names of your input dataset in Amazon Simple Storage Service (Amazon S3).

	sentiment	review_body
0	1	This skirt is stunning but i really wish it wa...
1	0	This is a very cute shirt with great details. ...
2	1	I'm really looking forward to wearing this sko...
3	-1	Ag jeans have been a main stay for me. they ar...
4	1	The fabric is thin though. you better wash it...
5	-1	I'm a rather small person--5'2" about 100 lbs...
6	0	This top fits like shown in the pictures. howe...
7	1	Super cute and comfy perfect for fall and win...
8	0	This dress has been one that i just adored onl...
9	0	Loved the design and print of this blouse. ho...

Column Analysis

The AutoML job analyzed the `2` input columns to infer each data type and select the feature processing pipelines for each training algorithm. For more details on the specific AutoML pipeline candidates, see Amazon

SageMaker Autopilot Candidate Definition Notebook.ipynb.

Percent of Missing Values

Within the data sample, the following columns contained missing values, such as: `nan`, white spaces, or empty fields.

SageMaker Autopilot will attempt to fill in missing values using various techniques. For example, missing values can be replaced with a new 'unknown' category for `Categorical` features and missing `Numerical` values can be replaced with the **mean** or **median** of the column.

We found **0 of the 2** of the columns contained missing values.

 Suggested Action Items - Investigate the governance of the training dataset. Do you expect this many missing values? Are you able to fill in the missing values with real data? - Use domain knowledge to define an appropriate default value for the feature. Either: - Replace all missing values with the new default value in your dataset in Amazon S3. - Add a step to the feature pre-processing pipeline to fill missing values, for example with a [sklearn.impute.SimpleImputer](https://scikit-learn.org/stable/modules/generated/sklearn.impute.SimpleImputer.html).

	% of Missing Values

Count Statistics

For `String` features, it is important to count the number of unique values to determine whether to treat a feature as `Categorical` or `Text` and then processes the feature according to its type.

For example, SageMaker Autopilot counts the number of unique entries and the number of unique words. The following string column would have **3** total entries, **2** unique entries, and **3** unique words.

	String Column
0	"red blue"
1	"red blue"
2	"red blue yellow"

If the feature is `Categorical`, SageMaker Autopilot can look at the total number of unique entries and transform it using techniques such as one-hot encoding. If the field contains a `Text` string, we look at the number of unique words, or the vocabulary size, in the string. We can use the unique words to then compute text-based features, such as Term Frequency-Inverse Document Frequency (tf-idf).

Note: If the number of unique values is too high, we risk data transformations expanding the dataset to too many features. In that case, SageMaker Autopilot will attempt to reduce the dimensionality of the post-processed data, such as by capping the number vocabulary words for tf-idf, applying Principle Component Analysis (PCA), or other dimensionality reduction techniques.

The table below shows **2 of the 2** columns ranked by the number of unique entries.

 Suggested Action Items - Verify the number of unique values of a feature is expected with respect to domain knowledge. If it differs, one explanation could be multiple encodings of a value. For example `US` and `U.S.` will count as two different words. You could correct the error at the data source or pre-process your dataset in your S3 bucket. - If the number of unique values seems too high for Categorical variables, investigate if using domain knowledge to group the feature to a new feature with a smaller set of possible values improves performance.

	Number of Unique Entries	Number of Unique Words (if Text)
sentiment	3	n/a
review_body	7108	17986

Descriptive Statistics

For each of the numerical input features, several descriptive statistics are computed from the data sample.

SageMaker Autopilot may treat numerical features as `Categorical` if the number of unique entries is sufficiently low. For `Numerical` features, we may apply numerical transformations such as normalization, log and quantile transforms, and binning to manage outlier values and difference in feature scales.

We found 1 of the 2 columns contained at least one numerical value. The table below shows the 1 columns which have the largest percentage of numerical values.

Suggested Action Items - Investigate the origin of the data field. Are some values non-finite (e.g. infinity, nan)? Are they missing or is it an error in data input? - Missing and extreme values may indicate a bug in the data collection process. Verify the numerical descriptions align with expectations. For example, use domain knowledge to check that the range of values for a feature meets with expectations.

	% of Numerical Values	Mean	Median	Min	Max
sentiment	100.0%	0.0	0.0	-1.0	1.0