Amazon SageMaker Autopilot Data Exploration Report

This report contains insights about the dataset you provided as input to the AutoML job. This data report was generated by **automl-fraudcase-03-00-14-16** AutoLM job. To check for any issues with your data and possible improvements that can be made to it, consult the sections below for guidance. You can use information about the predictive power of each feature in the **Data Sample** section and from the correlation matrix in the **Cross Column Statistics** section to help select a subset of the data that is most significant for making predictions.

Note: SageMaker Autopilot data reports are subject to change and updates. It is not recommended to parse the report using automated tools, as they may be impacted by such changes.

Dataset Summary

Dataset Properties

Rows	Columns	Duplicate rows	Target column	Missing target values	Invalid target values	Detected problem type
227846	29	0.53%	Class	0.00%	0.00%	BinaryClassification

Detected Column Types

	Numeric	Categorical	Text	Datetime	Sequence
Column Count	28	0	0	0	0
Percentage	100.00%	0.00%	0.00%	0.00%	0.00%

Report Contents

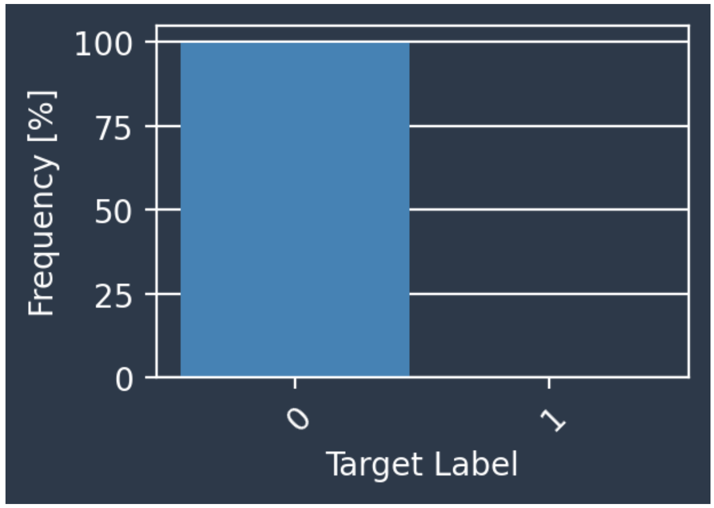
Target Analysis

The column **Class** is used as the target column. See the distribution of values (labels) in the target column below:

Number of Classes	Invalid Percentage	Missing Percentage

Number of Classes	Invalid Percentage	Missing Percentage
2	0.00%	0.00%

Target Label	Frequency Percentage	Label Count
0	99.83%	227452
1	0.17%	394



Histogram of the target column labels.

Data Sample

The following table contains a random sample of **10** rows from the dataset. The top two rows provide the type and prediction power of each column. 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).

	Class	V11	V10	V14	
Prediction Power	-	0.999332	0.998998	0.998832	0.9
Column Types	-	numeric	numeric	numeric	n
35642	0	-0.186932507856365	-0.264405052696032	0.277410194920294	-0.45021475070
177333	0	0.45086981730199	10.4257234389354	-5.17902075992527	-1.4106656213
35888	0	-1.35177566019963	0.0115004687886872	0.449300945031224	-0.0229029350
139319	0	0.732920443807004	0.42290485404323297	1.08186580512093	0.6218514328
52395	0	-0.8734506040280859	-0.0729804710441808	-0.403224912588031	0.2149000329
133683	0	-0.402279381897754	-0.6742142284192599	0.116034731132129	1.26067847
99418	0	-0.5411441360375371	0.8089280692602231	-0.487030878596944	-0.769883812
58506	0	-1.13185727623873	-0.21240650410944897	-0.7158303988167	-0.0222911858

Duplicate Rows

⚠ **Low severity insight: “Duplicate rows”**
0.53% of the rows were found to be duplicates when testing a random sample of 10000 rows from the dataset. Some data sources could include valid duplicates, but in some cases these duplicates could point to problems in data collection. Unintended duplicate rows could disrupt the automatic hyperparameter tuning of Amazon SageMaker Autopilot and result in sub-par model. Thus should be removed for more accurate results. This preprocessing can be done with Amazon SageMaker Data Wrangler using the “Drop duplicates” transform under “Manage rows”.

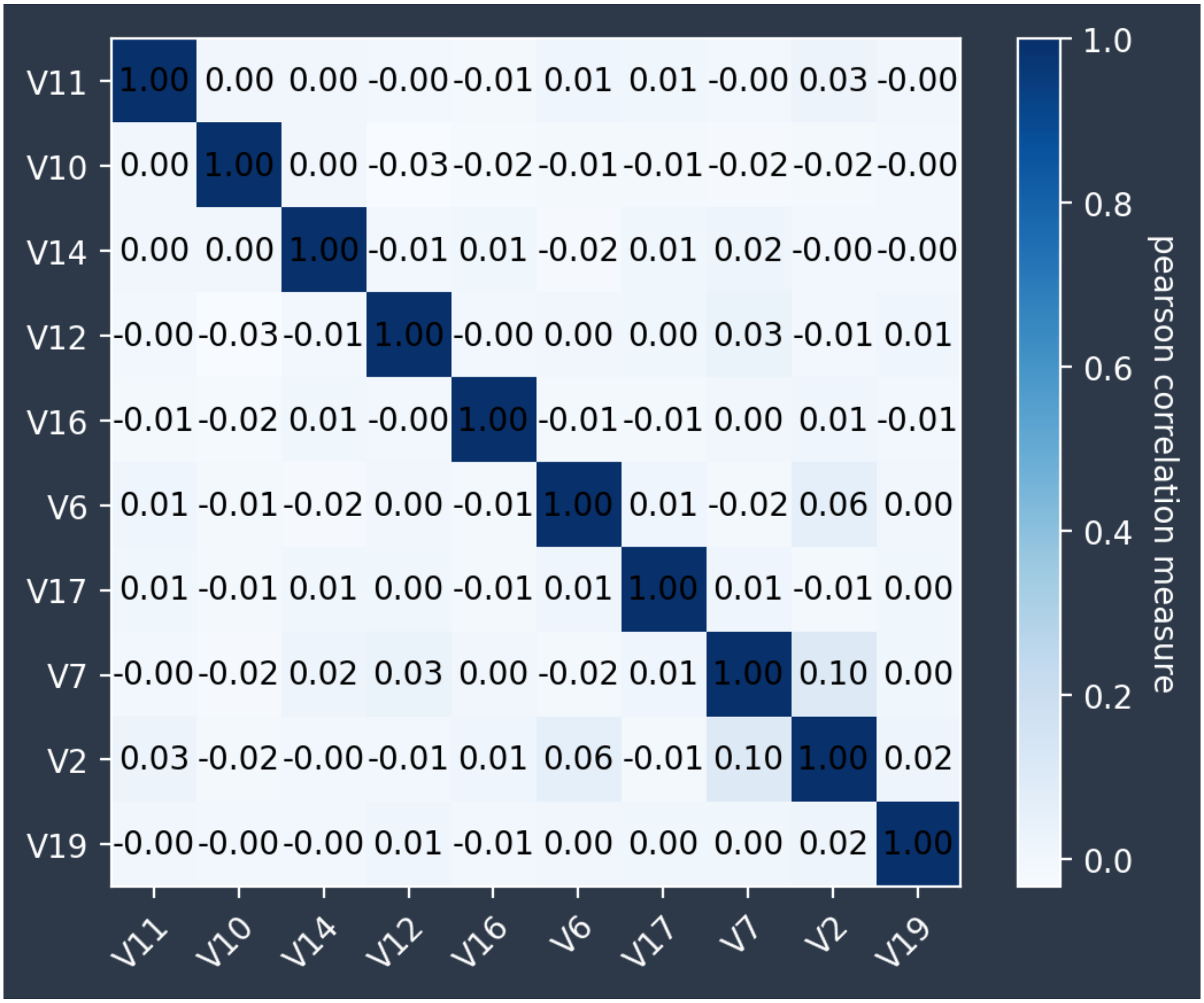
A sample of duplicate rows is presented below. The number of occurrences of a row is given in left most **Duplicate count** column.

Duplicate count	V1	V2	V3	V4
5	1.24567381944824	0.166975019545401	0.488305742562781	0.6353219207244001
4	2.0533112135278504	0.08973464781763099	-1.68183566862495	0.45421196023303295
3	2.04862857999224	-0.367489471288573	-2.5440649129570003	-0.7284718306852042
3	1.16095677944219	1.26562134412589	-1.57647298025721	1.4729879637544598
3	2.03499679478385	0.273057863689555	-3.8485521533257097	-0.36653356556066496
3	1.79012266278276	-0.676921329244653	-1.4073821822197898	-0.10049693190563999
2	-1.9260094377309498	1.4110933564510602	1.23024310264001	1.1801495588163
2	1.38552585500187	-0.990191903333903	-0.624478619475939	-1.78961355254733
2	2.06408608374138	-0.0738024376904158	-1.4916281809924399	0.142715842735288
2	2.01070503619608	0.14791145611498002	-1.59646654329458	0.33024521316543604

Cross Column Statistics

Amazon SageMaker Autopilot calculates Pearson’s correlation between columns in your dataset. Removing highly correlated columns can reduce overfitting and training time. Pearson’s correlation is in the range [-1, 1] where 0 implies no correlation, 1 implies perfect correlation, and -1 implies perfect inverse correlation.

The full correlation matrix between the 10 most predictive numeric features is presented below.



Cross column correlation for numeric features

Anomalous Rows

Anomalous rows are detected using the Isolation forest algorithm on a sample of **10000** randomly chosen rows after basic preprocessing. The isolation forest algorithm associates an anomaly score to each row of the dataset it is trained on. Rows with negative anomaly scores are usually considered anomalous and rows with positive anomaly scores are considered non-anomalous. When investigating an anomalous row, look for any unusual values - in particular any that might have resulted from errors in the gathering and processing of data. Deciphering whether a row is indeed anomalous, contains errors, or is in fact valid requires domain knowledge and application of business logic.

Inspect the rows below, to see if any of those are anomalous. A subset of rows is presented below. Anomaly score is presented as the left most column; Smaller values indicate a higher chance that the row is anomalous.

	Anomaly Scores	V1	V2	V3	
8288	-0.257247	-13.1926709562391	12.785970638297998	-9.90665002092758	3.320336882889
8303	-0.250446	-36.5105831707971	-40.938048442402206	-5.37798649493194	11.4745895673491
2850	-0.222405	-21.775338877596397	-17.135619792132303	-5.34501684767869	3.75051432777
4857	-0.218579	-29.200328590574397	16.1557014298057	-30.013712485724803	6.47673117996E

	Anomaly Scores	V1	V2	V3	
3916	-0.217312	-7.23466336596796	4.01391071480448	-3.8838941652989303	-2.68681369127
3661	-0.217122	-21.2091195927913	12.6521968313004	-23.5539329441267	6.1740779100435
7280	-0.205716	-28.344757250015803	-40.978852228328705	0.422089971454916	9.264320616983
7328	-0.199004	-19.438377351953303	-17.1641400626533	-8.61024038249712	4.07381273249
7909	-0.195948	-3.7656801220835	5.8907352377921	-10.2022676310229	10.259035976621
1781	-0.195665	-11.918762692066501	8.62611106242464	-15.8957547641341	6.03880956264

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 29** of the columns contained missing values.

Cardinality

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 **25 of the 29** columns ranked by the number of unique entries.

The table below shows **25 of the 26** columns ranked by the number of unique entries.

💡 Suggested Action Items

- Verify the number of unique values of a feature is as expected. One explanation for unexpected number of unique values 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 multiple unique values can be grouped into a smaller set of possible values.

	Number of Unique Entries	Number of Unique Words (if Text)
Class	2	n/a
V11	221384	n/a
V27	221450	n/a
V14	221475	n/a
V16	221726	n/a
V3	221844	n/a
V15	221895	n/a
V12	221970	n/a
V10	222037	n/a
V19	222039	n/a
V7	222100	n/a
V2	222283	n/a
...	...	...
V24	222838	n/a
V22	222930	n/a
V18	222946	n/a
V1	223028	n/a
V26	223105	n/a
V4	223132	n/a
V20	223213	n/a
V17	223246	n/a
V23	223314	n/a
V21	223443	n/a
V9	223459	n/a
V13	223616	n/a

Descriptive Stats

For each of the input features that has at least one numeric value, 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 **29 of the 29** columns contained at least one numerical value. The table below shows the **25** columns which have the largest percentage of numerical values. Percentage of outliers is calculated only for columns which Autopilot detected to be of numeric type. Percentage of outliers is not calculated for the target column.

💡 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	% of Outlier Values
Class	100.0%	0.001729	0.0	0.0	1.0	nan
V15	100.0%	-0.001139	0.0477206	-4.49894	5.78451	0.0
V27	100.0%	-6.9e-05	0.00114182	-9.89524	12.1524	2.7
V26	100.0%	0.000713	-0.0462435	-2.60455	3.46325	0.1
V25	100.0%	0.000449	0.00434888	-7.49574	7.51959	0.2
V24	100.0%	-0.000336	0.0384062	-2.83663	4.02287	0.1
V23	100.0%	0.001336	-0.0076995	-36.666	22.5284	2.5
V22	100.0%	-0.00026	0.00168712	-10.9331	10.5031	0.1
V21	100.0%	0.000439	-0.0293923	-34.8304	27.2028	2.6
V20	100.0%	-0.000845	-0.0626164	-28.0096	39.4209	2.8
V19	100.0%	0.000186	0.0102449	-7.21353	5.59197	0.1
V18	100.0%	0.000477	-0.011147	-9.49875	5.04107	0.1
V17	100.0%	0.000243	-0.0968522	-24.0191	9.25353	0.5
V16	100.0%	0.000104	0.0672927	-14.1299	8.28989	0.2
V14	100.0%	0.001217	0.0495381	-19.2143	10.5268	0.6
V1	100.0%	0.002949	0.0240586	-56.4075	2.45189	0.7
V13	100.0%	-6.1e-05	-0.0149279	-5.79188	4.56901	0.0
V12	100.0%	-0.001772	0.153916	-18.6837	7.84839	0.2
V11	100.0%	-0.000115	-0.0284942	-4.68293	12.0189	0.1
V10	100.0%	-0.000586	-0.0953705	-24.5883	15.3317	1.0
V9	100.0%	0.000581	-0.0608232	-13.4341	10.3929	0.3
V8	100.0%	0.002861	0.0292485	-73.2167	20.0072	3.0

	% of Numerical Values	Mean	Median	Min	Max	% of Outlier Values
V7	100.0%	2.1e-05	0.0407405	-43.5572	44.0545	1.1
V6	100.0%	-0.000578	-0.281031	-26.1605	23.9178	0.3
V5	100.0%	0.001302	-0.0654268	-42.1479	34.8017	0.6

Definitions

Feature types

Numeric: Numeric values, either floats or integers. For example: age, income. When training a machine learning model, it is assumed that numeric values are ordered and a distance is defined between them. For example, 3 is closer to 4 than to 10 and 3 < 4 < 10.

Categorical: The column entries belong to a set of unique values that is usually much smaller than number of rows in the dataset. For example, a column from datasets with 100 rows with the unique values "Dog", "Cat" and "Mouse". The values could be numeric, textual, or combination of both. For example, "Horse", "House", 8, "Love" and 3.1 are all valid values and can be found in the same categorical column. When manipulating column of categorical values, a machine learning model does not assume that they are ordered or that distance function is defined on them, even if all of the values are numbers.

Binary: A special case of categorical column for which the cardinality of the set of unique values is 2.

Text: A text column that contains many non-numeric unique values, often a human readable text. In extreme cases, all the elements of the column are unique, so no two entries are the same.

Datetime: This column contains date and/or time information.

Feature statistics

Prediction power: Prediction power of a column (feature) is a measure of how useful it is for predicting the target variable. It is measured using a stratified split 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]. A higher prediction power score near 1 indicate that a column is more useful for predicting the target on its own. A lower score near 0 indicate that a column contains little useful information for predicting the target on their own. Although it is possible that a column that is uninformative on its own can be 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 present in dataset that is hard or impossible to obtain at prediction time, such as a duplicate of the target.

Outliers: Outliers are detected using two statistics that are robust to outliers: median and robust standard deviation (RSTD). RSTD is derived by clipping the feature values to the range [5 percentile, 95 percentile] and calculating the standard deviation of the clipped vector. All values larger than median + 5 * RSTD or smaller than median - 5 * RSTD are considered to be outliers.

Skew: Skew measures the symmetry of the distribution and is defined as the third moment of the distribution divided by the third power of the standard deviation. The skewness of the normal distribution or any other symmetric distribution is zero. Positive values imply that the right tail of the distribution is longer than the left tail. Negative values imply that the left tail of the distribution is longer than the right tail. As a thumb rule, a distribution is considered skewed when the absolute value of the skew is larger than 3.

Kurtosis: Pearson's kurtosis measures the heaviness of the tail of the distribution and is defined as the fourth moment of the distribution divided by the fourth power of the standard deviation. The kurtosis of the normal distribution is 3. Thus, kurtosis values lower than 3 imply that the distribution is more concentrated around the mean and the tails are lighter than the tails of the normal distribution. Kurtosis values higher than 3 imply heavier tails than the normal distribution or that the data contains outliers.

Missing Values: Empty strings and strings composed of only white spaces are considered missing.

Valid values:

- **Numeric features / regression target:** All values that could be casted to finite floats are valid. Missing values are not valid.
- **Categorical / binary / text features / classification target:** All values that are not missing are valid.
- **Datetime features:** All values that could be casted to datetime object are valid. Missing values are not valid.

Invalid values: values that are either missing or that could not be casted to the desired type. See the definition of valid values for more information