# Data Pre-processing Model

Machine learning DV2578 (Spam Filter)

## Dataset Summary

Number of rows: 4600  
Number of columns: 58  
Problem type: Binary Classification

**Duplicate rows**: 8.50%  
**High Severity Insight:** Duplicate Rows Detected  
**Severity**: High  
**Issue**  
Approximately 8.50% of the rows in the dataset are found to be duplicates. Duplicate rows can significantly impact the accuracy and reliability of data analysis, leading to skewed results and erroneous conclusions.  
Action Item  
Identify and remove duplicate rows from the dataset. This process will improve the quality of the dataset and enhance the effectiveness of any downstream analysis or modeling tasks.

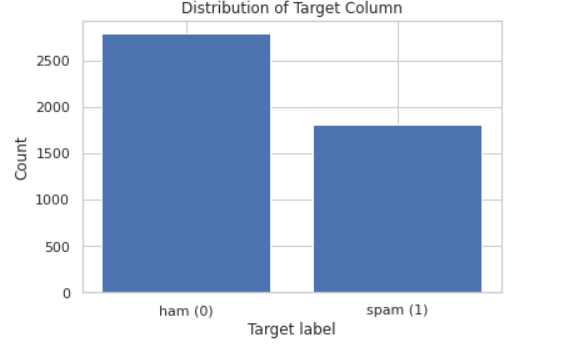
|  |  |
| --- | --- |
| **Missing target values** | **0.00%** |
| **Invalid target values** | **0.04%** |

# Valid values Numeric features

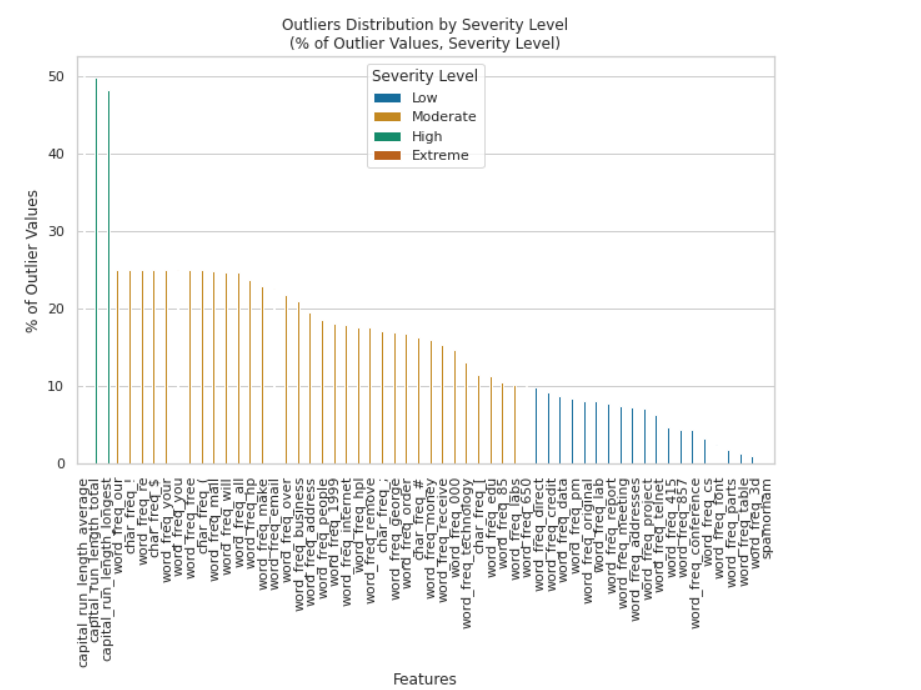
All values that could be cast to finite floats are valid. Missing values are not valid.  
**Column Type Analysis**

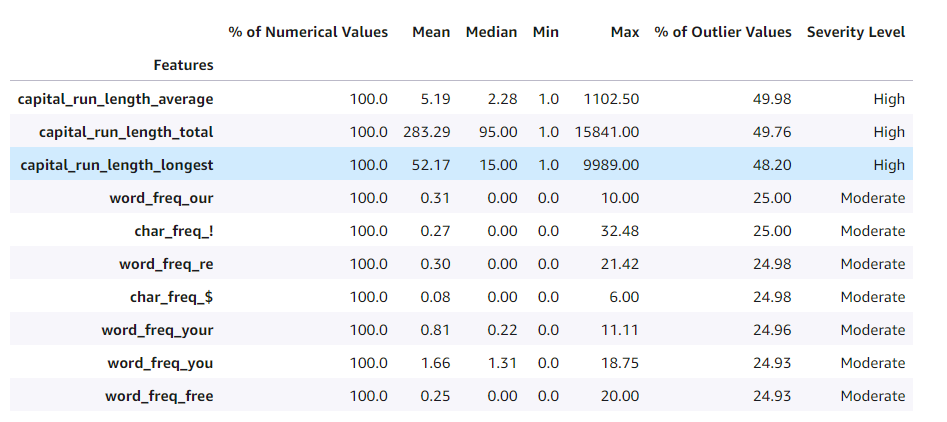
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **target** | **Categories** | **Type** | **Count** | **Percentage** |
| **0** | **Numeric** | **float64** | **55** | **94.827586** |
| **1** | **Numeric** | **int64** | **3** | **5.172414** |

**Target Analysis**  
The column spamorham is used as the target column. See the distribution of target column values (labels) in the target column below:  
Number of Classes: 2  
Positive Label: spam (0)  
Negative Label: ham (1)



**Missing Values**  
Empty strings and strings composed of only white spaces are considered missing.  
Missing target values: 0.00%  
**Invalid values**  
Values that are either missing or that could not be cast to the desired type.  
Invalid target values: 0.00%  
**Descriptive Statistics**  
The Descriptive statistics are computed from the data sample.  
We found 58 of the 58 columns contained at least one numerical value.  
**Outliers**  
The outlined strategy provides a systematic approach for identifying and analyzing outliers in a dataset. The steps involve calculating the percentage of outliers by comparing values to percentiles, selecting the top outliers, creating a comprehensive table with statistics, sorting it by outlier percentages, adding a severity level column, and resetting the index for clarity. This approach quantifies and categorizes outliers, facilitating further analysis and decision-making based on their severity across different features.





**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.   
- Use domain knowledge to check that the range of values for a feature meets expectations.

**Feature Summary**  
Prediction power is measured by stratified splitting the spam data set into 80% and 20% training and validation folds.  
We fit a Random Forest Classifier model for each feature separately on the training fold after applying minimal feature pre-processing and measure prediction performance on the validation data.  
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.  
Note: We are only showing the top 10 predicted results.

