kdb+/q AutoML Procedure Report

This report outlines the results for a classification problem achieved through running kdb+/q AutoML.

This run started on 2024.09.06 at 18:37:06.224.

Description of Input Data

The following is a breakdown of information for each of the relevant columns in the dataset:

col	count	unique	mean	std	min	max	type
comment	900	900					text

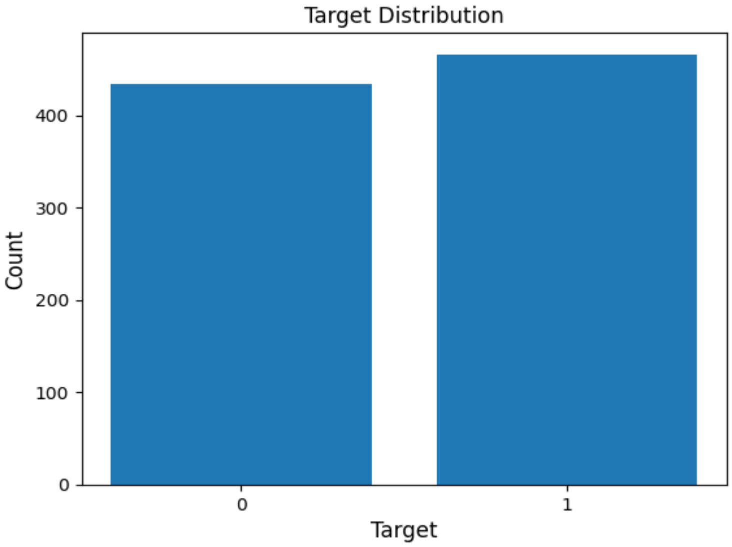

Figure 1: Distribution of input target data

Breakdown of Pre-Processing

Nlp feature extraction and selection was performed with a total of 52 features produced.

Feature extraction took 00:01:05.060 time in total.

Initial Scores

5-

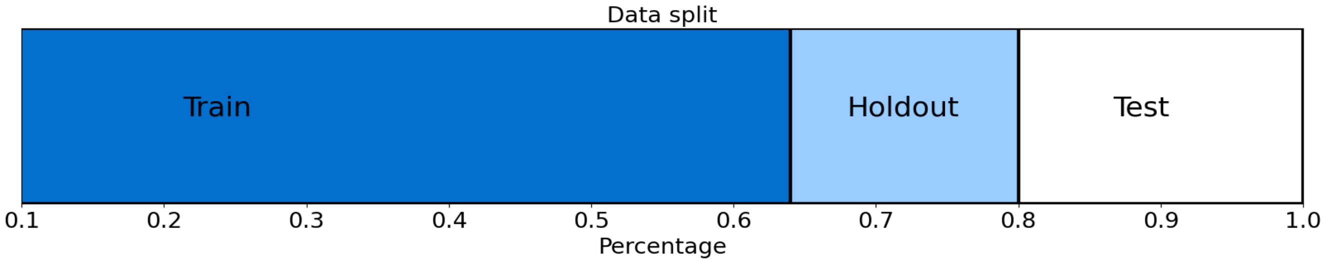

Figure 2: The data split used within this run of AutoML, with data split into training, holdout and testing sets

The total time taken to carry out cross validation for each model on the training set was 00:00:07.410 where models were scored and optimized using .ml.accuracy.

Model scores:

RandomForestClassifier = 0.7516342
LinearSVC = 0.7465517
MLPClassifier = 0.7447976
AdaBoostClassifier = 0.7412294
KNeighborsClassifier = 0.7342429
GradientBoostingClassifier = 0.7309745
SVC = 0.7256822
LogisticRegression = 0.7223388
BinaryKeras = 0.6857871
GaussianNB = 0.657931

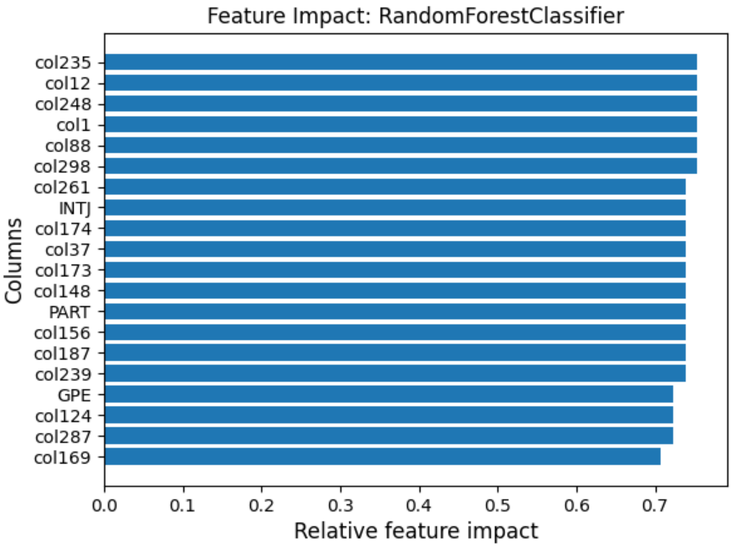

Figure 3: Feature impact of each significant feature as determined by the training set

Model selection summary

Best scoring model = RandomForestClassifier

The score on the holdout set for this model was = 0.7986111.

The total time taken to complete the running of this model on the holdout set was: 00:00:00.333.

Best Model

A 5-fold grid search was performed on the training set to find the best model using, .automl.gs.kfShuff.

The following are the hyperparameters which have been deemed optimal for the model:

criterion = gini
min_samples_split = 2
min_samples_leaf = 1

The score for the best model fit on the entire training set and scored on the testing set was = 0.7333333

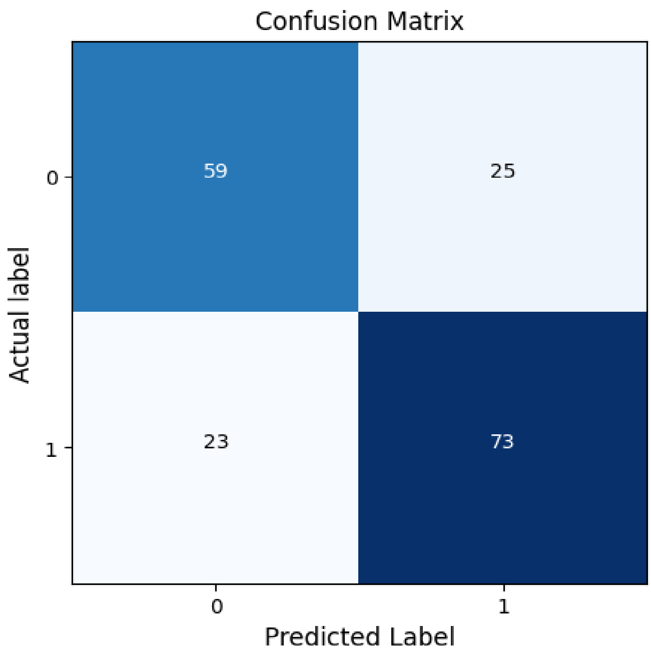

Figure 4: This is the confusion matrix produced for predictions made on the testing set