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 19:11:21.220.

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
tenure	4500	73	32.326	24.559306529918306	0	72	numeric
MonthlyCharges	4500	1251	64.88497777777778	30.497952771442122	18.55	118.75	numeric
TotalCharges	4500	3178	2284.2517034068146	2275.07802803704	18.85	8672.45	numeric
customerID	4500	3310					categorical
gender	4500	2					categorical
Partner	4500	2					categorical
Dependents	4500	2					categorical
PhoneService	4500	2					categorical
MultipleLines	4500	3					categorical
InternetService	4500	3					categorical

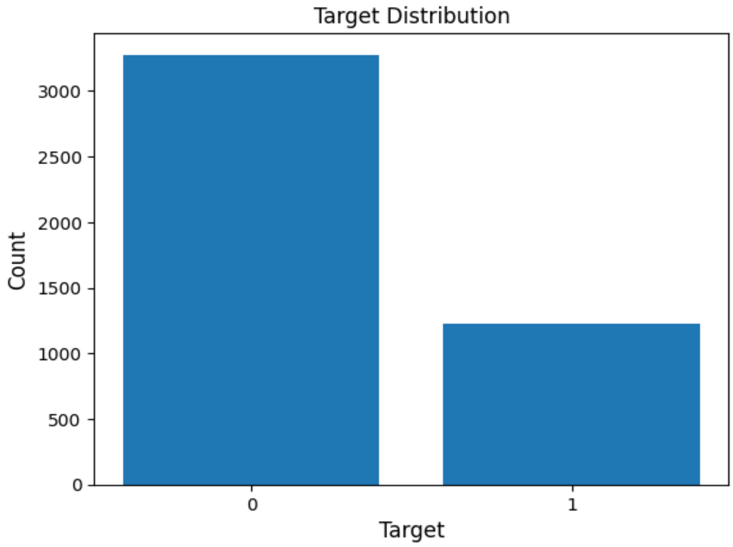

Figure 1: Distribution of input target data

Breakdown of Pre-Processing

Normal feature extraction and selection was performed with a total of 428 features produced.

Feature extraction took 00:00:44.639 time in total.

Initial Scores

5-

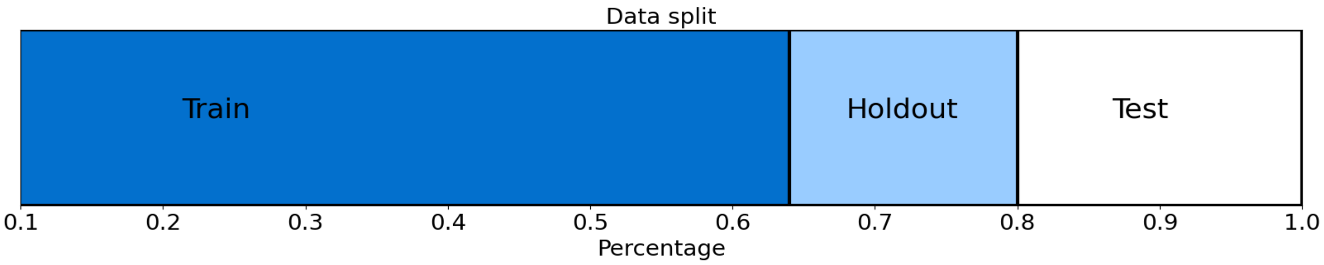

Figure 2: The data split used within this run of AutoML, with data split into training, holdout and testing sets

The total time taken to carry out cross validation for each model on the training set was 00:00:24.509 where models were scored and optimized using .ml.accuracy.

Model scores:

RandomForestClassifier = 0.8322917
GradientBoostingClassifier = 0.8086806
LinearSVC = 0.803125
LogisticRegression = 0.8024306
AdaBoostClassifier = 0.7975694
KNeighborsClassifier = 0.7732639
MLPClassifier = 0.7579861
GaussianNB = 0.7347222
SVC = 0.7298611
BinaryKeras = 0.7104167

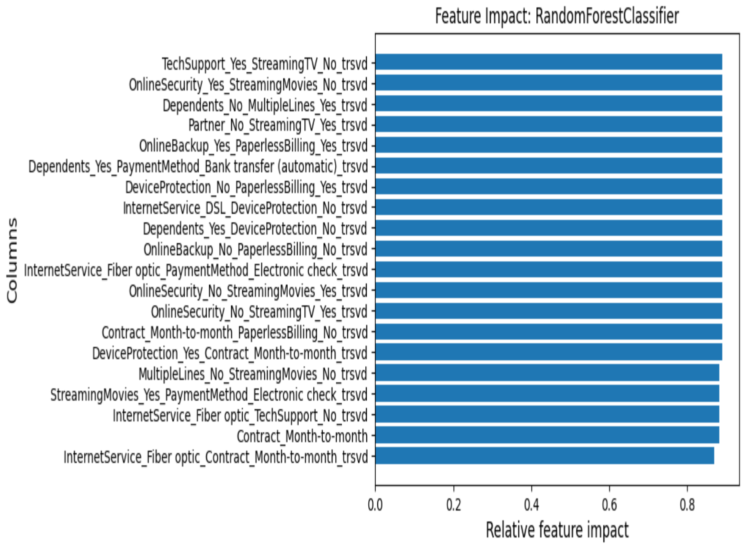

Figure 3: Feature impact of each significant feature as determined by the training set

Model selection summary

Best scoring model = RandomForestClassifier

The score on the holdout set for this model was = 0.8402778.

The total time taken to complete the running of this model on the holdout set was: 00:00:00.933.

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 = 0.8555556

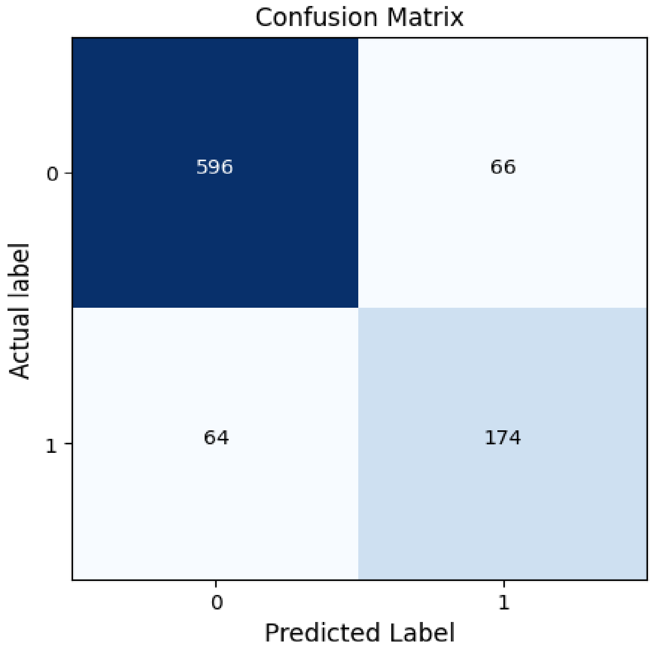

Figure 4: This is the confusion matrix produced for predictions made on the testing set