Select Rows — Tue Jun 07 22, 12:30:04

Data

Instances: 1000000
Condition: fraud is 1.0

Output

Matching data: 87403 instances
Non-matching data: 912597 instances

Select Columns — Tue Jun 07 22, 12:30:17

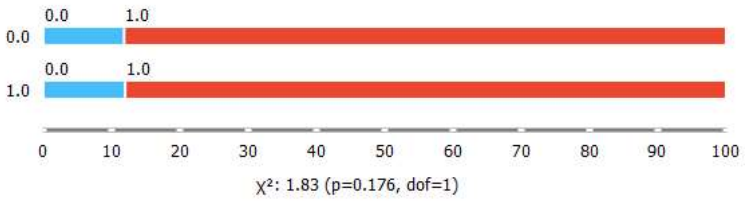
Input data

Features: distance_from_home, distance_from_last_transaction, ratio_to_median_purchase_price, repeat_retailer, used_chip, used_pin_number, online_order, fraud

Output data

Features: distance_from_home, distance_from_last_transaction, ratio_to_median_purchase_price, repeat_retailer, used_chip, used_pin_number, online_order
Target: fraud

Box Plot — Tue Jun 07 22, 12:30:34

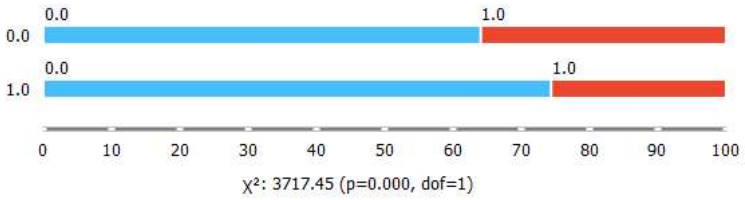


Box plot for attribute 'repeat_retailer' grouped by 'fraud'

How to read this boxplot:
The top bar represents Repeat Retailer = 0 (not a repeat retailer).
The bottom bar represents Repeat Retailer = 1 (repeat retailer).
The left/blue part of each bar represents Fraud = 0 (not fraud).
The right/red part of each bar represents Fraud = 1 (fraudulent transaction).

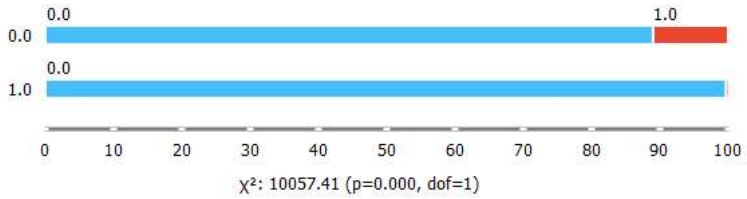
Same applies to the next box plots.

Box Plot — Tue Jun 07 22, 12:30:39



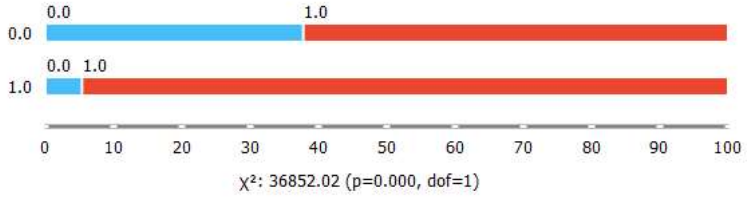
Box plot for attribute 'used_chip' grouped by 'fraud'

Box Plot — Tue Jun 07 22, 12:30:41



Box plot for attribute 'used_pin_number' grouped by 'fraud'

Box Plot — Tue Jun 07 22, 12:30:45



Box plot for attribute 'online_order' grouped by 'fraud'

Rank — Tue Jun 07 22, 12:32:37

Input

Features: distance_from_home, distance_from_last_transaction, ratio_to_median_purchase_price, repeat_retailer, used_chip, used_pin_number, online_order
Target: fraud

Ranks

	#	Info. gain
ratio_to_median_purchase_price		0.0893072903068266
online_order	2.0	0.03433445747401781
used_pin_number	2.0	0.012185683344356102
distance_from_home		0.008266999392740282
used_chip	2.0	0.0028073756166503427
distance_from_last_transaction		0.0008270180657372617
repeat_retailer	2.0	1.3250818776122664e-06

Output

Features: ratio_to_median_purchase_price, online_order, used_pin_number, distance_from_home, used_chip
Target: fraud

Data Sampler — Tue Jun 07 22, 12:32:42

Sampling type: Random sample with 60 % of data, stratified (if possible), deterministic
Input: 1000000 instances
Sample: 600000 instances
Remaining: 400000 instances

Test and Score — Tue Jun 07 22, 12:33:29

Settings

Sampling type: No sampling, test on testing data
Target class: Average over classes

Scores

Model	AUC	CA	F1	Precision	Recall
kNN dist-weight 6	0.9970074437768868	0.9880275	0.9880858550557525	0.9881651849408919	0.9880275
Logistic Regression	0.9666141913003695	0.9590675	0.9554642905690897	0.9570528422721966	0.9590675
Naive Bayes	0.9243354343604238	0.9144025	0.8913619361720656	0.8898722086188269	0.9144025
Tree	0.871611184122134	0.9770275	0.9754323847183966	0.9775915793914405	0.9770275

Confusion Matrix — Tue Jun 07 22, 12:33:55

Confusion matrix for kNN dist-weight 6 (showing number of instances)

		Predicted 0.0	1.0	∑
Actual	**0.0**	362263	2776	**365039**
	1.0	2013	32948	**34961**
	∑	**364276**	**35724**	**400000**

Confusion Matrix — Tue Jun 07 22, 12:34:02

Confusion matrix for Logistic Regression (showing number of instances)

		Predicted 0.0	1.0	∑
Actual	**0.0**	362464	2575	**365039**
	1.0	13798	21163	**34961**
	∑	**376262**	**23738**	**400000**

Confusion Matrix — Tue Jun 07 22, 12:34:09

Confusion matrix for Naive Bayes (showing number of instances)

		Predicted 0.0	1.0	∑
Actual	**0.0**	360641	4398	**365039**
	1.0	29841	5120	**34961**
	∑	**390482**	**9518**	**400000**

Confusion Matrix — Tue Jun 07 22, 12:34:18

Confusion matrix for Tree (showing number of instances)

		Predicted 0.0	1.0	∑
Actual	**0.0**	365039	0	**365039**
	1.0	9189	25772	**34961**
	∑	**374228**	**25772**	**400000**