CV

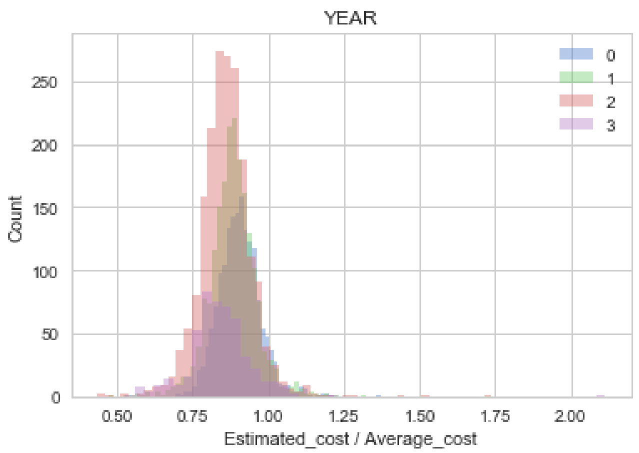
Task 1, Task 2 analysis

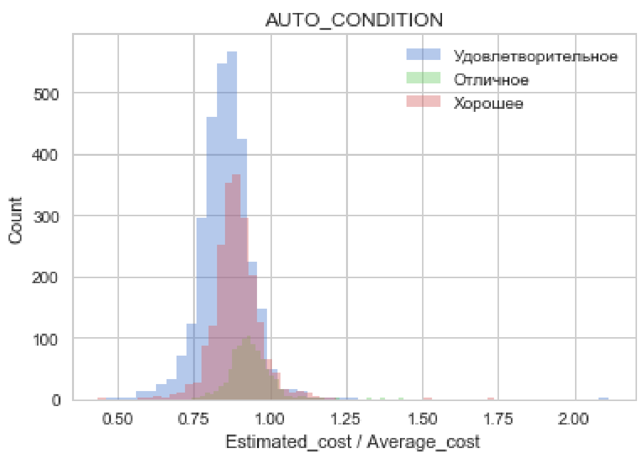
Name:	Zhumazhenis Dairabay
City:	Astana
University:	Nazarbayev University, School of Engineering, 2nd year student
School:	KTL, Kyzylorda
Skills:	1. Machine learning https://www.udemy.com/machinelearning/ course is completed, use Python. In 1st year, I had Programming for Engineers course and I did group project related to Machine Learning. We evaluated true errors of different Discriminant Analysis classifiers and SVM classifiers, and compared them: LDA, QDA, DLDA, RLDA, G13, KSVM, LSVM. https://github.com/zhumazhenis/Classifier/blob/master/Report_of_project_Grade_114_out_of_115.pdf 2. Programming languages: C++, Java, Python 3. Few experience in Excel, SQL, Matlab, Wolfram Mathematica. I can improve my skills in Excel, SQL during working process. 4. Android App Development. Completed nFactorial Incubator 2017, published Makhal-matel app. http://bit.ly/makhal-matel
Achievements:	Math Olympiads winner: 1. International Mathematics Competition 2017 (IMC 2017), Bulgaria – participant 2. International Zhautykov Math Olympiad 2014 – bronze. 3. Asian-Pacific Math Olympiad 2015 – bronze 4. Silk-Road Math Olympiad 2014 – bronze 5. Republican Math Olympiad 2015 – bronze, and other reginal 6. International Tuymaada Math Olympiad 2012, Yakutia – participant 7. NU Math Battle 2016 – 3rd prize Algorithm skills: ACM ICPC Quarter Final – participant

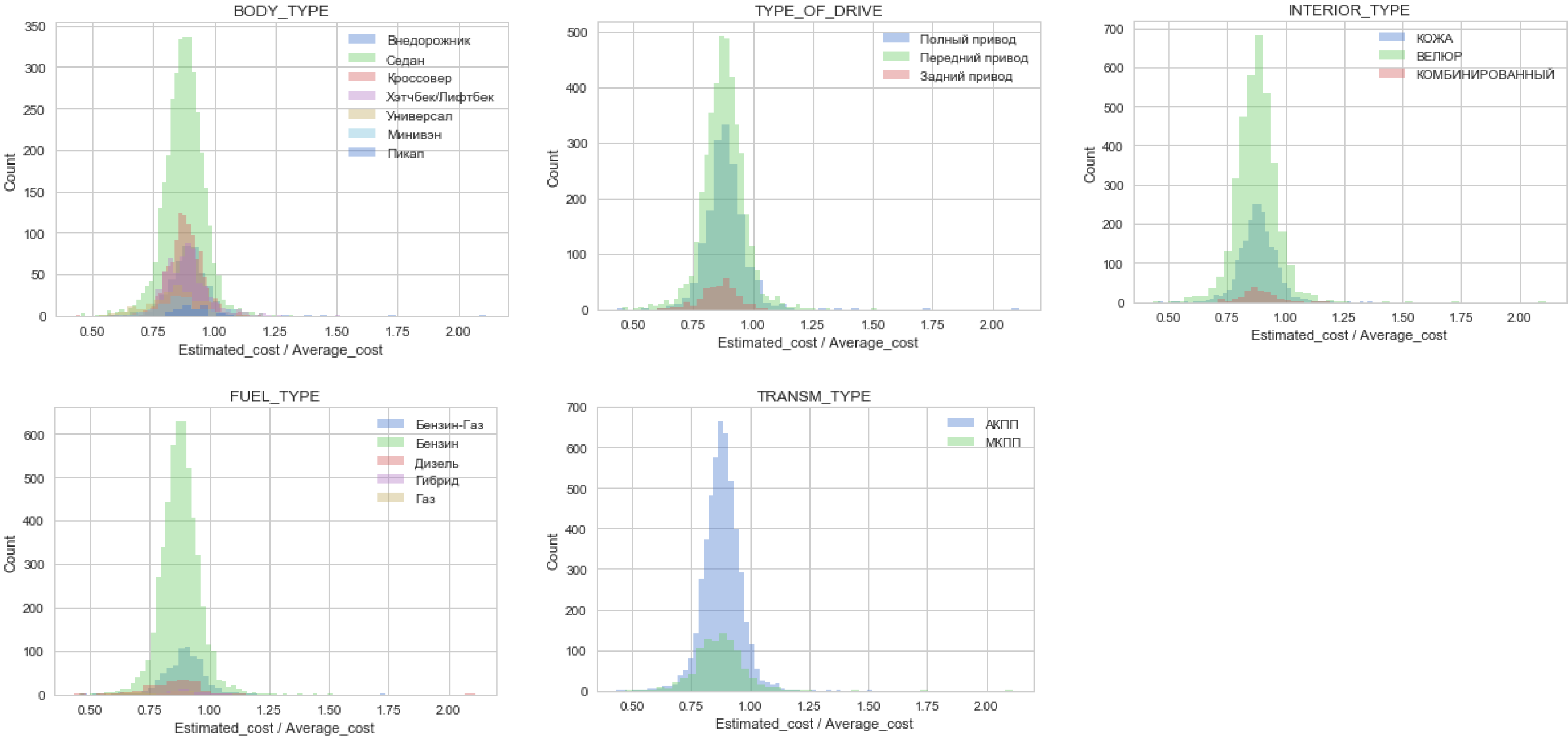
Done by Zhumazhenis Dairabay

Task 1

Aim:	To estimate the cost of cars in TEST data, according to TRAIN data.
Process:	1. Split Мошенники_training data into training and test data (proportion 0.8 and 0.2). 2. Remove redundant features like ID, VIN_# 3. It is clearly seen that, the ESTIM_COST is almost linearly dependent to AVG_COST. But, also we have to check how other features affect. $coefficient = \frac{ESTIM\_COST}{AVG\_COST}$ Distribution graphics of each feature are shown below. The graphic is $coefficient$ vs number of cars. From the graphs, it is clearly seen that almost all features are not discriminative and biases ESTIM_COST. Despite that, AUTO_CONDITION and YEAR features are selected for prediction. 20 different variables in YEAR feature are grouped and labeled as below. $YEAR = \begin{cases} 0, & 0 \leq YEAR < 5 \\ 1, & 5 \leq YEAR < 10 \\ 2, & 10 \leq YEAR < 15 \\ 3, & 15 \leq YEAR < 20 \end{cases}$ 4. Apply Linear Regression (sklearn library).
Results:	**Training data split with proportion 0.8 and 0.2:** X_train (4800 samples): Number of predicted samples with -10%<error<10%: 3885 Accuracy: 3885 / 4800 * 100% = 80.94% X_test (1200 samples): Number of predicted samples with -10%<error<10%: 975 Accuracy: 975 / 1200 * 100% = 81.25%
Kernel:	





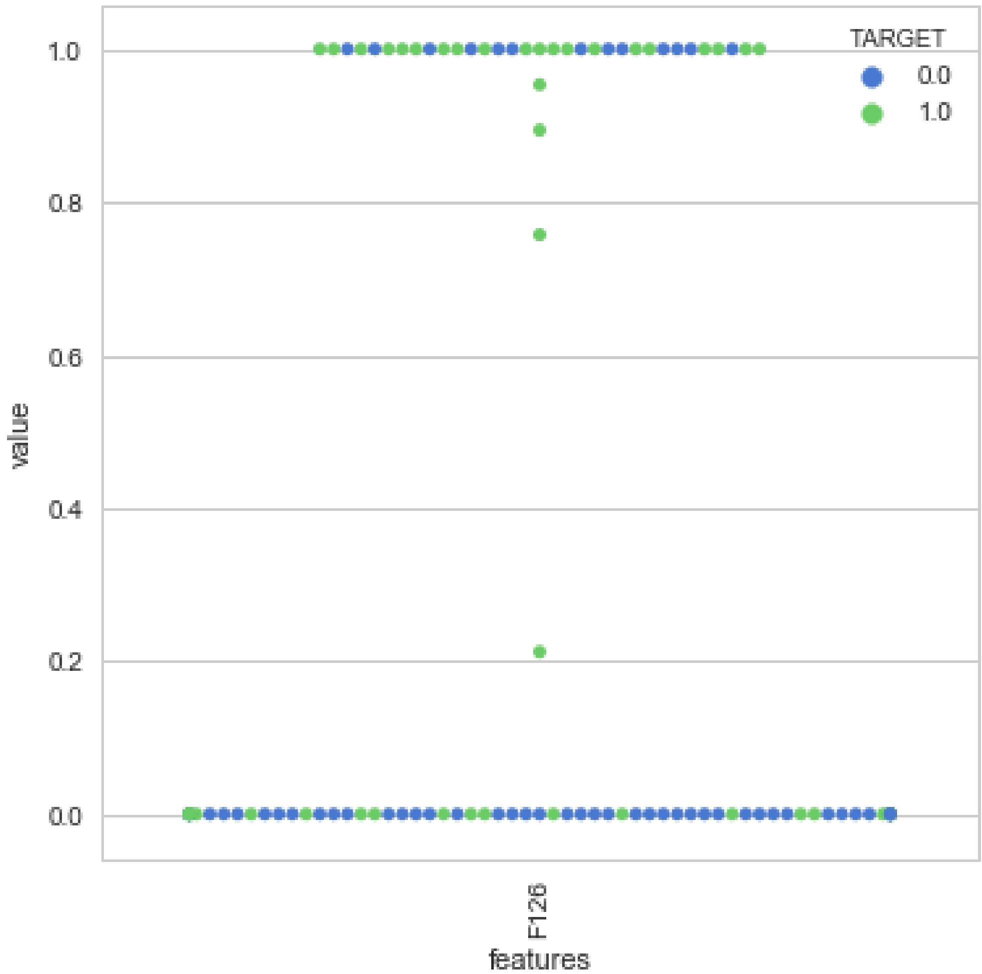

BODY_TYPE
Внедорожник
Седан
Кроссовер
Хэтчбек/Лифтбек
Универсал
Минивэн
Пикап
Count
Estimated_cost / Average_cost
TYPE_OF_DRIVE
Полный привод
Передний привод
Задний привод
Count
Estimated_cost / Average_cost
INTERIOR_TYPE
КОЖА
ВЕЛЮР
КОМБИНИРОВАННЫЙ
Count
Estimated_cost / Average_cost
FUEL_TYPE
Бензин-Газ
Бензин
Дизель
Гибрид
Газ
Count
Estimated_cost / Average_cost
TRANSM_TYPE
АКПП
МКПП
Count
Estimated_cost / Average_cost

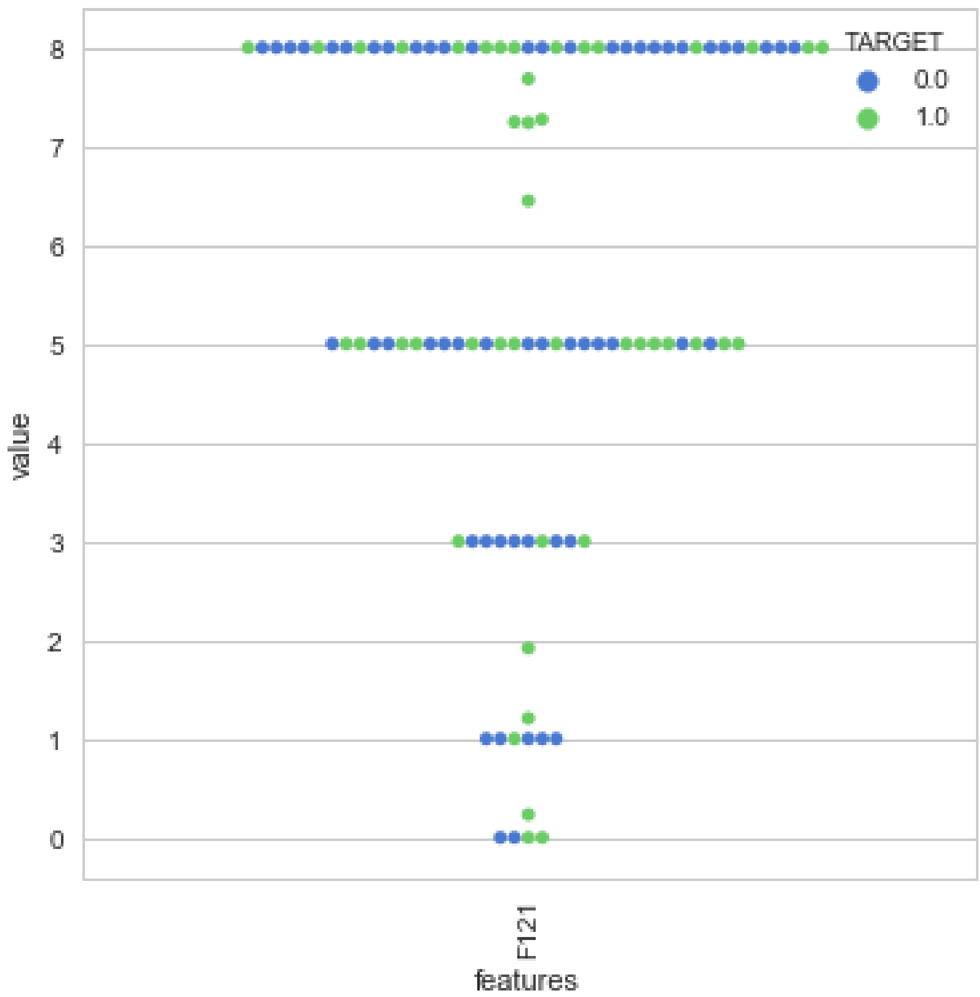
Task 2

Aim:	To classify Мошенники_test data, whether “мошенник” or not, according to Мошенники_training data.
Process:	1. Split Мошенники_training data into training and test data (proportion 0.8 and 0.2). 2. Remove redundant features like ID. 3. Identify highly discriminative features by visualizing them (27 features are selected). According, to graphics below, almost all values are not continuous (e.g. [0 or 1], [1,2,3]), the data best fits with tree algorithms. 4. Apply Random Forest Classifier (sklearn library).
Results:	**Training data split with proportion 0.8 and 0.2:** X_train (24000 samples): 23690 / 24000 * 100% = 98.7% X_test (6000 samples): 5592 / 6000 * 100% = 93.2%
Kernel:	

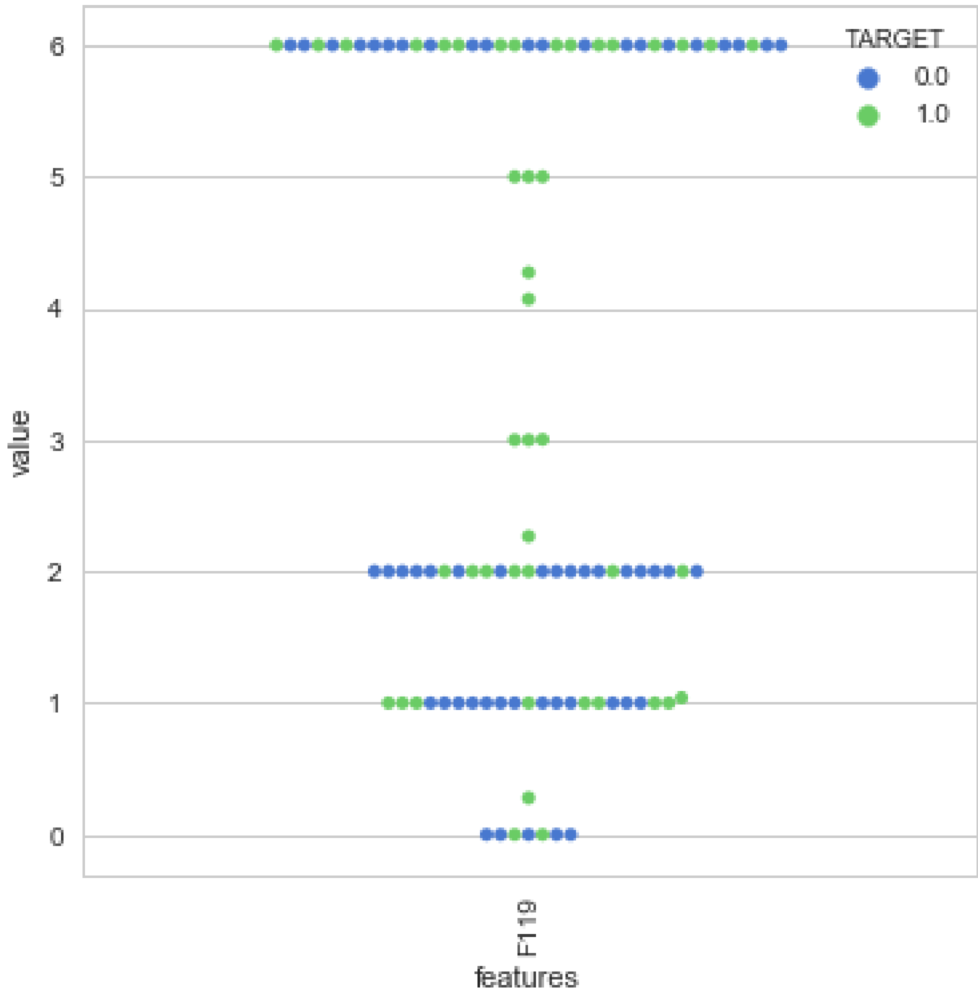
Highly discriminative features:

F2, F7, F8, F10, F11, F19, F32, F33, F34, F35, F38, F39, F40, F46, F63, F86, F87, F88, F107, F111, F77, F78, F79, F96, F108, F109, F119, F121, F126

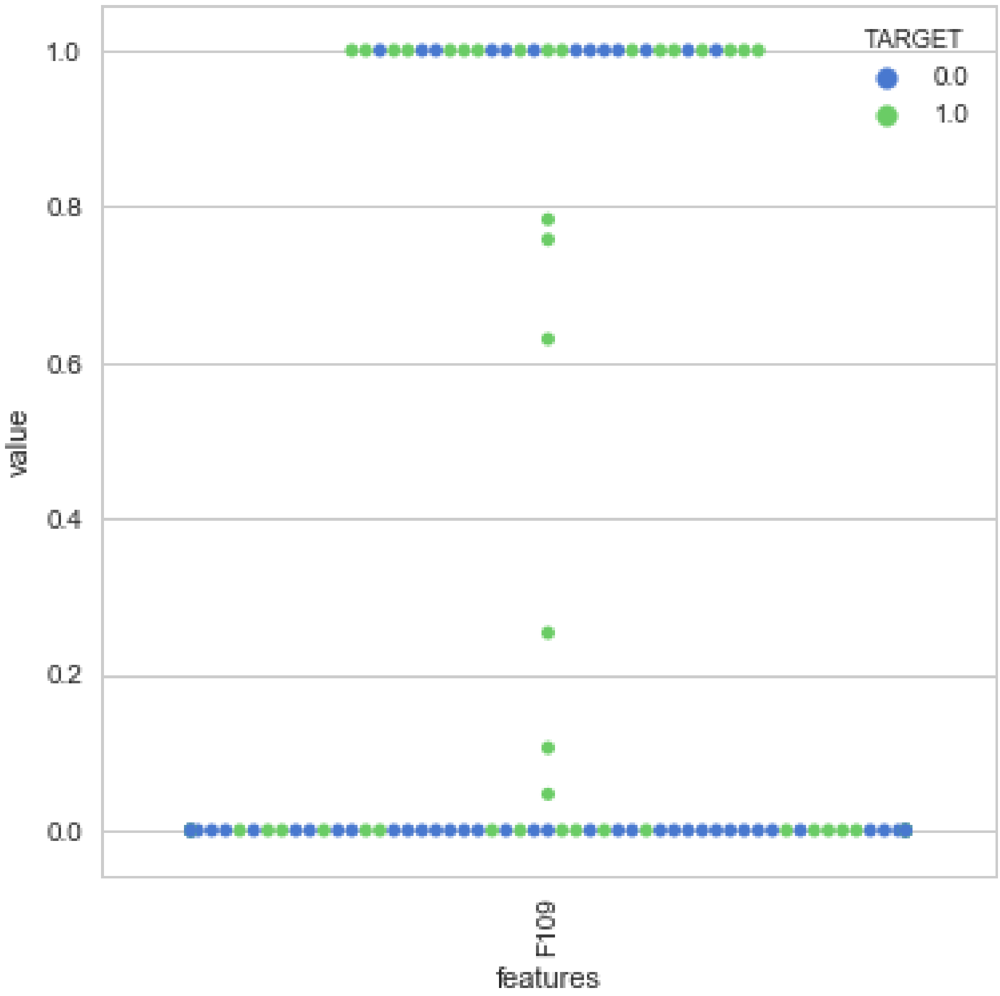




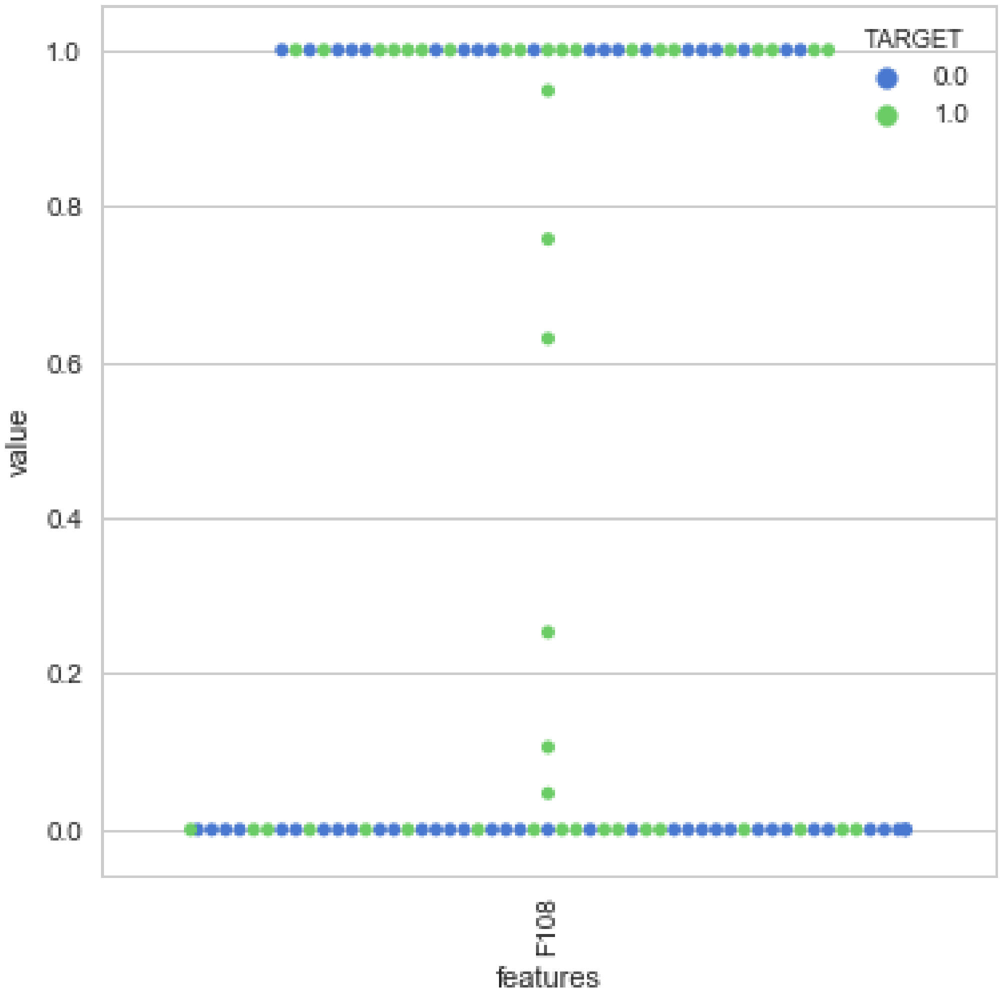
TARGET
0.0
1.0
8
7
6
5
4
3
2
1
0
value
F121
features

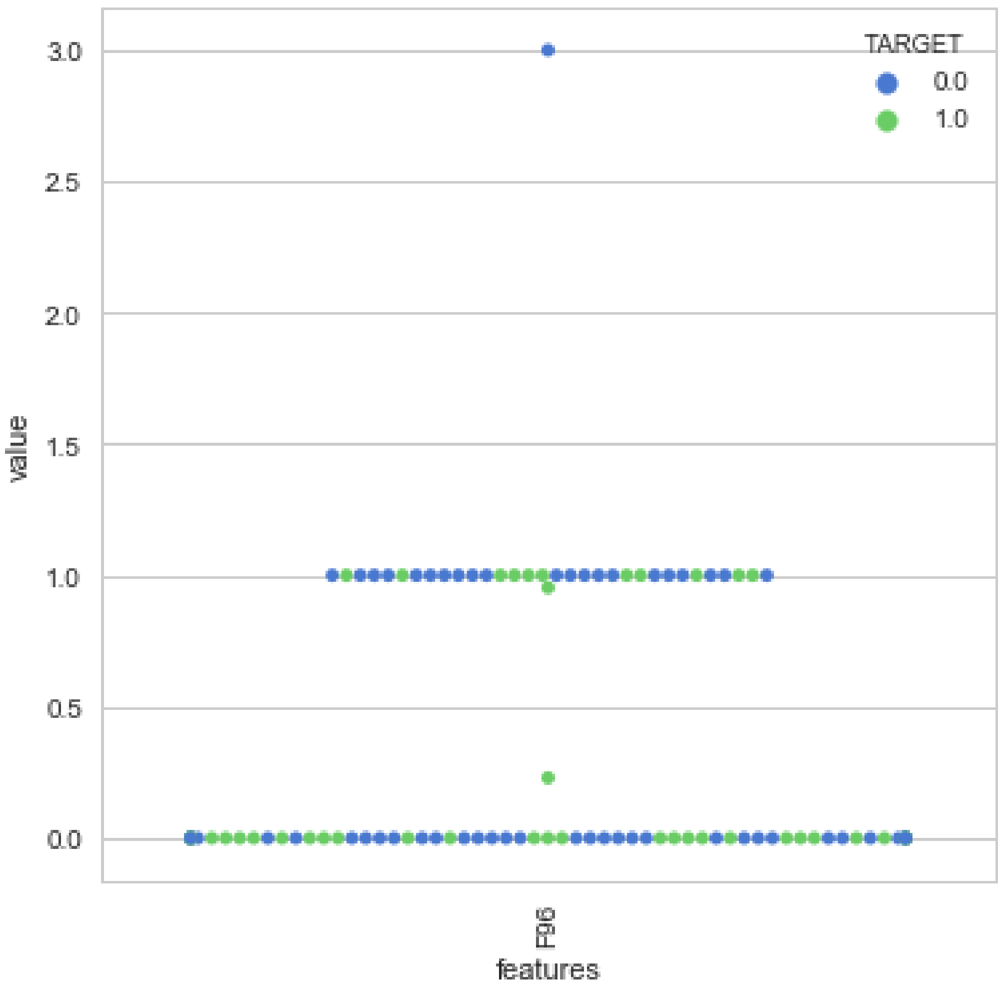


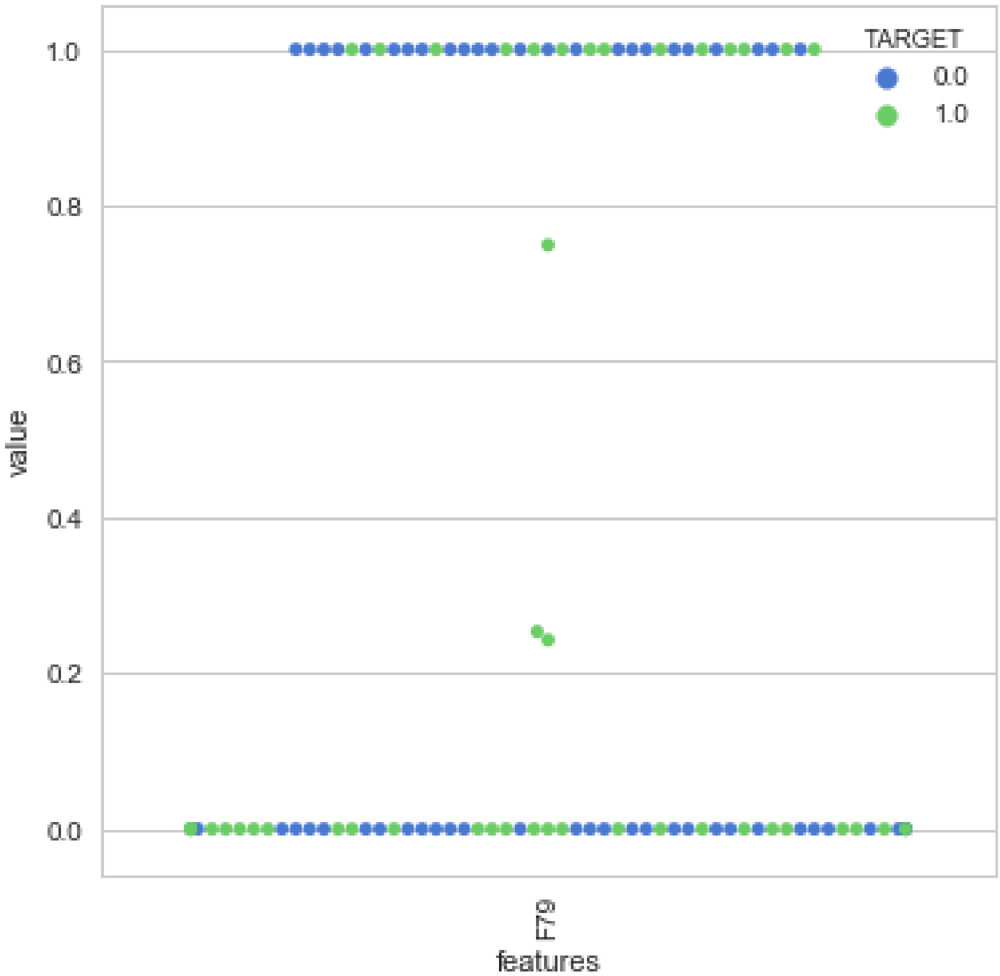
TARGET
0.0
1.0
value
0
1
2
3
4
5
6
F119
features



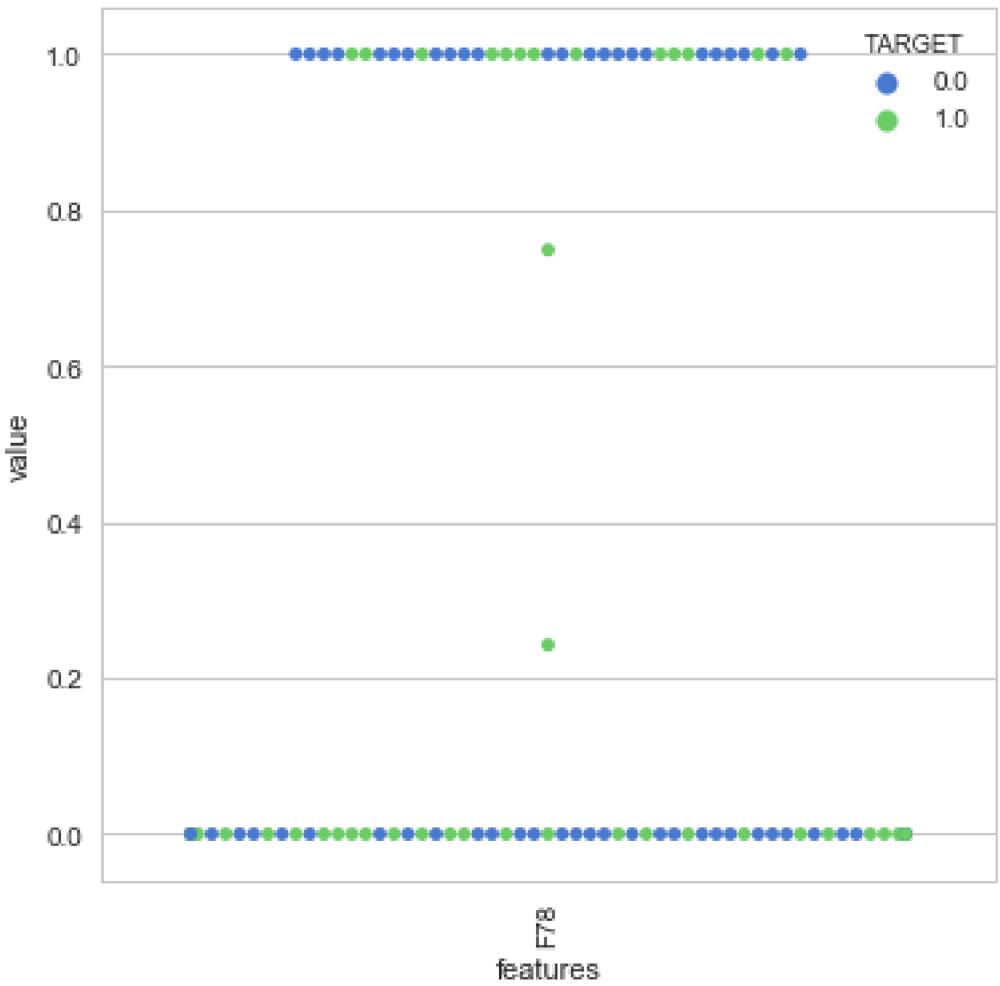
TARGET
0.0
1.0
1.0
0.8
0.6
0.4
0.2
0.0
value
F109
features

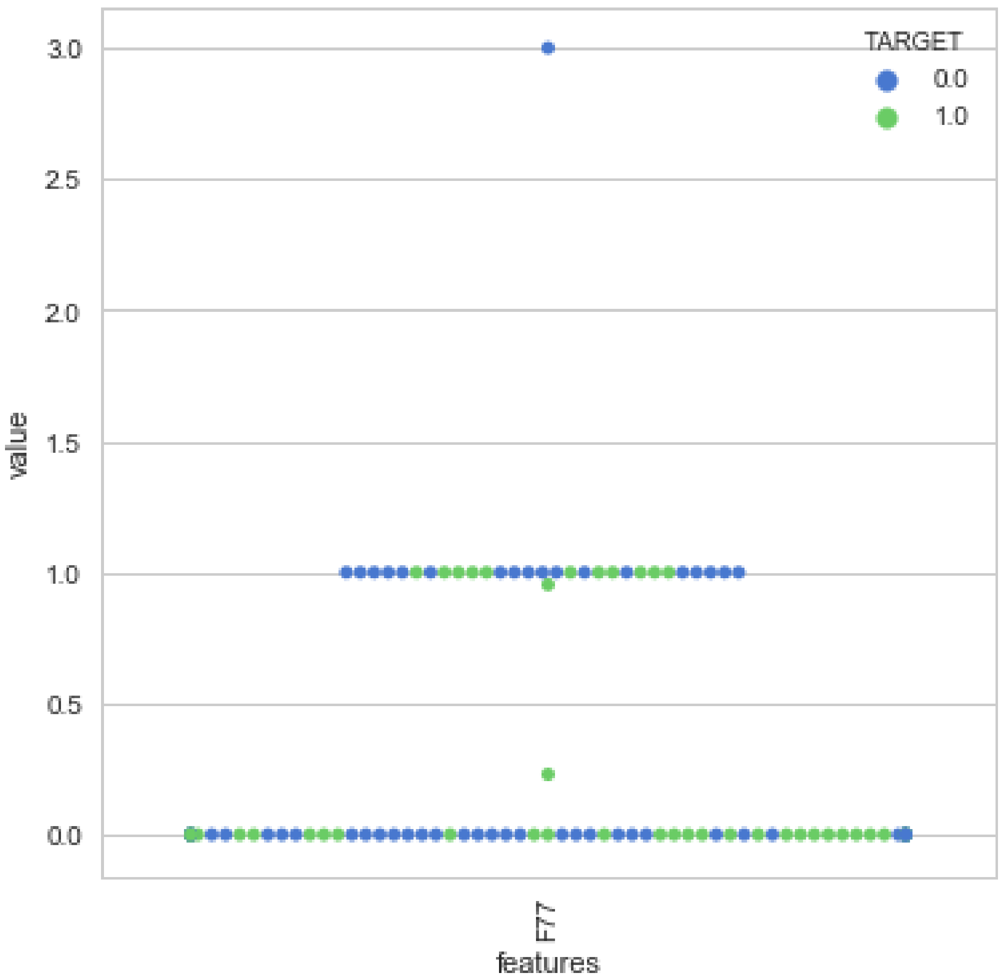

TARGET
0.0
1.0
1.0
0.8
0.6
0.4
0.2
0.0
value
F108
features

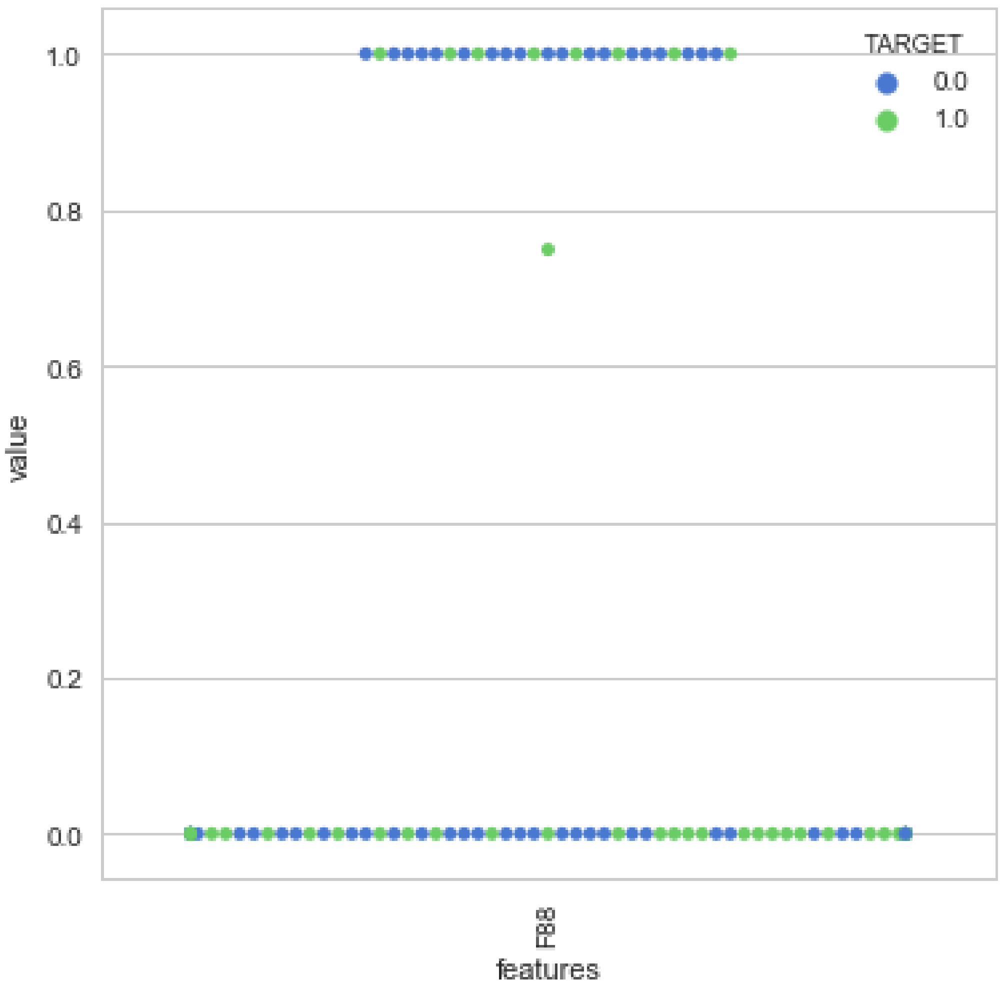

TARGET
0.0
1.0
3.0
2.5
2.0
1.5
1.0
0.5
0.0
value
F96
features



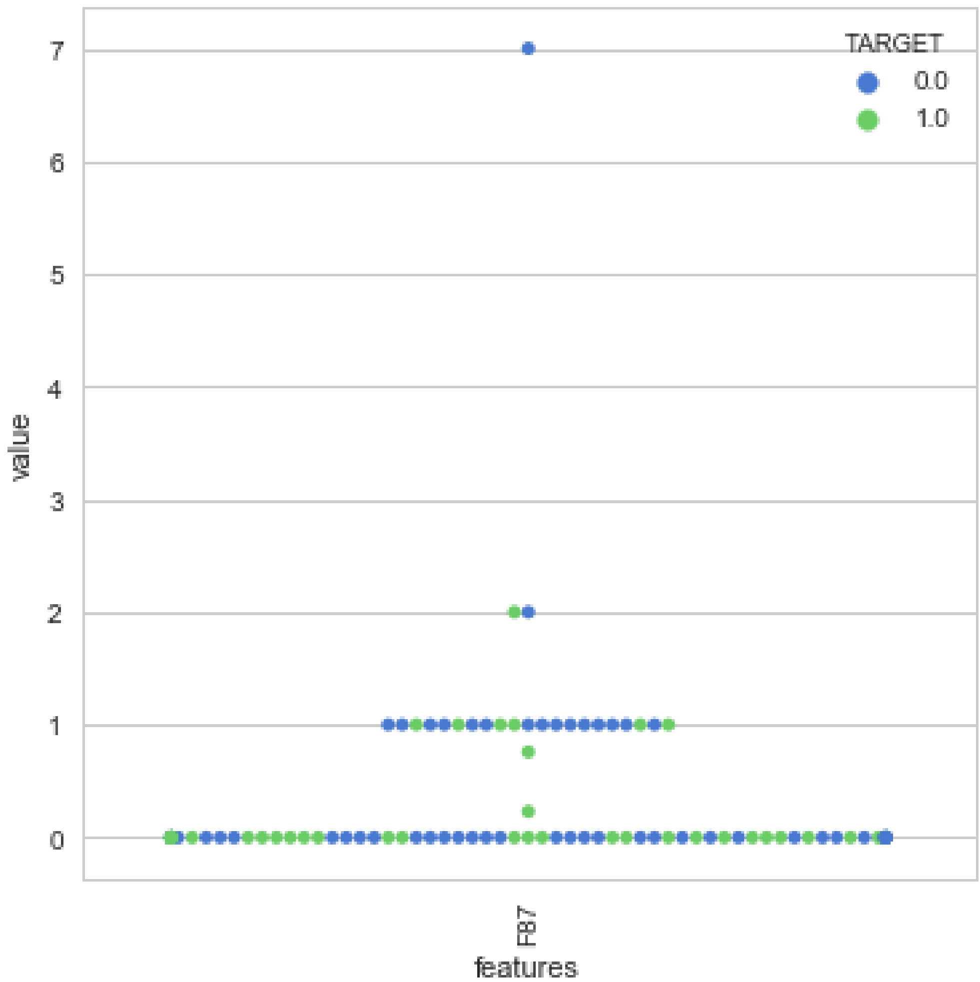
TARGET
0.0
1.0
1.0
0.8
0.6
0.4
0.2
0.0
value
F79
features


TARGET
0.0
1.0
1.0
0.8
0.6
0.4
0.2
0.0
value
F78
features

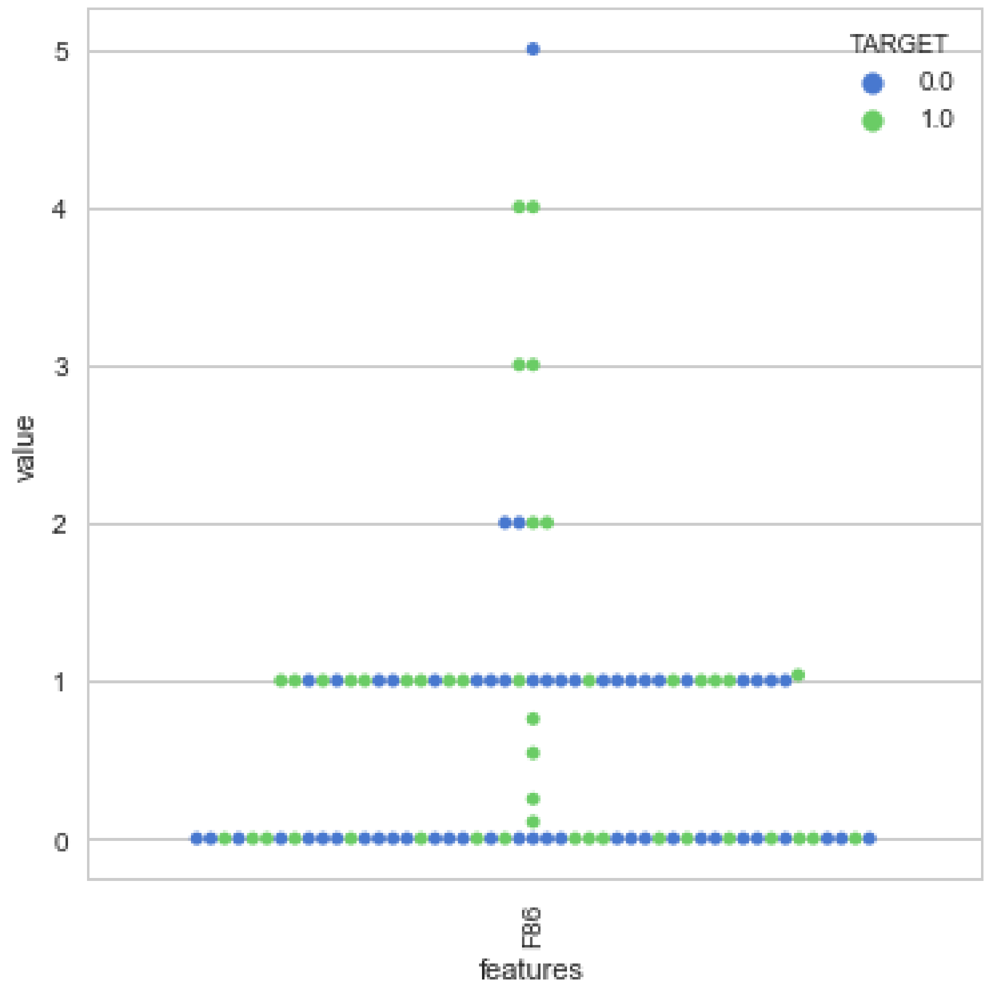

TARGET
0.0
1.0
3.0
2.5
2.0
1.5
1.0
0.5
0.0
value
F77
features

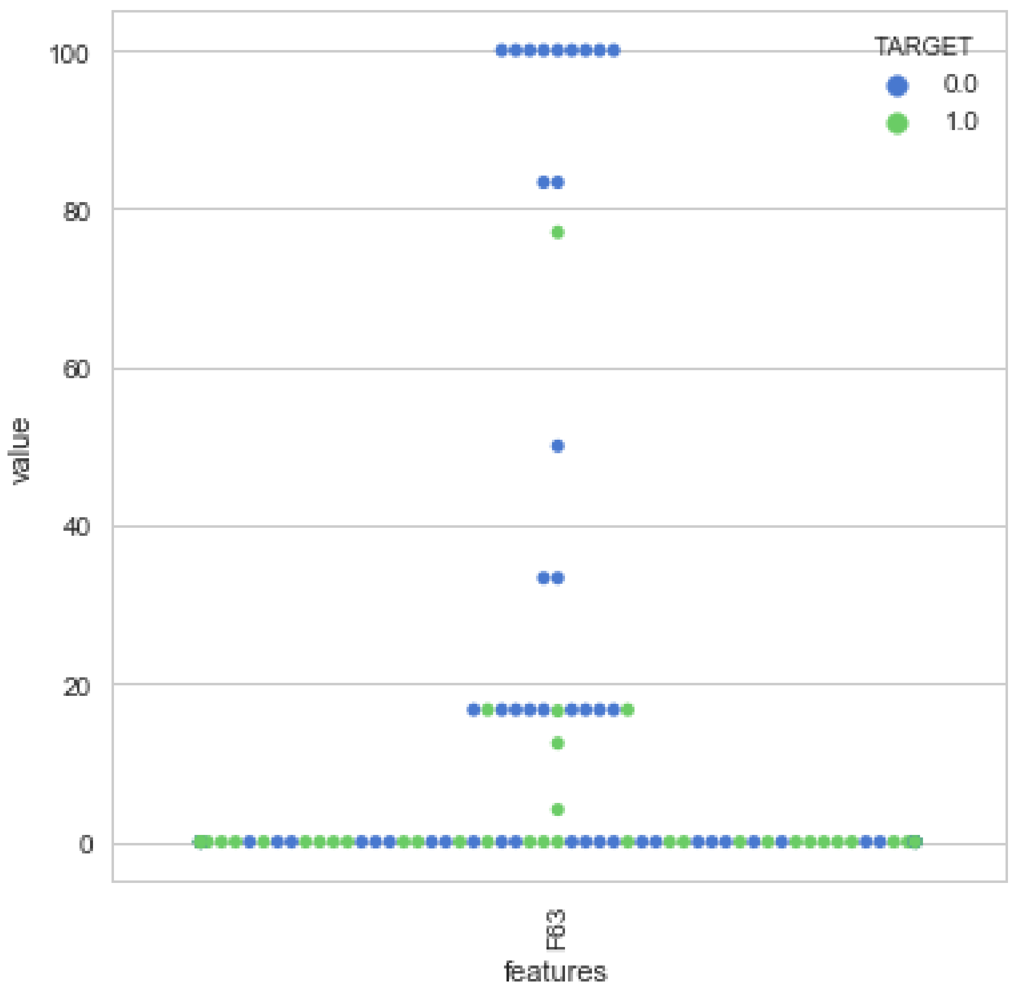


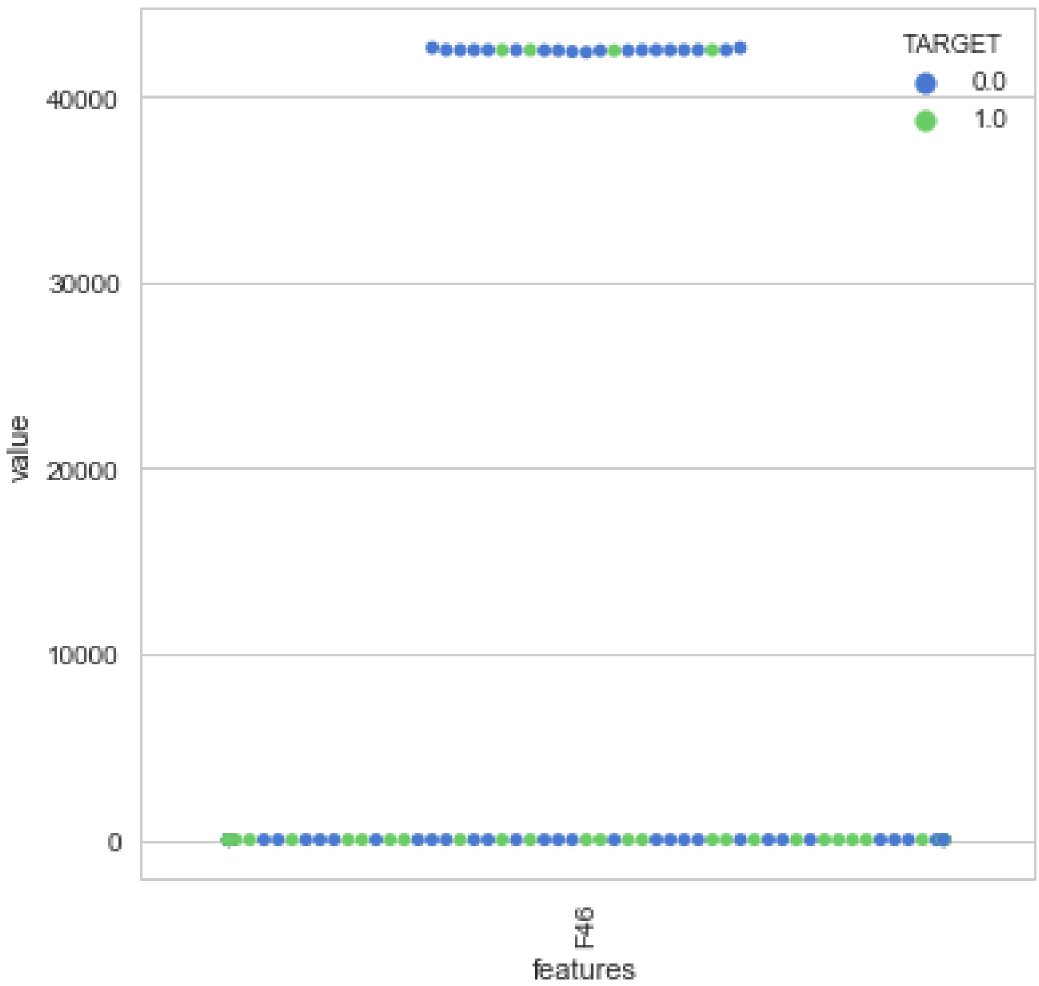
value
1.0
0.8
0.6
0.4
0.2
0.0
TARGET
0.0
1.0
F88
features

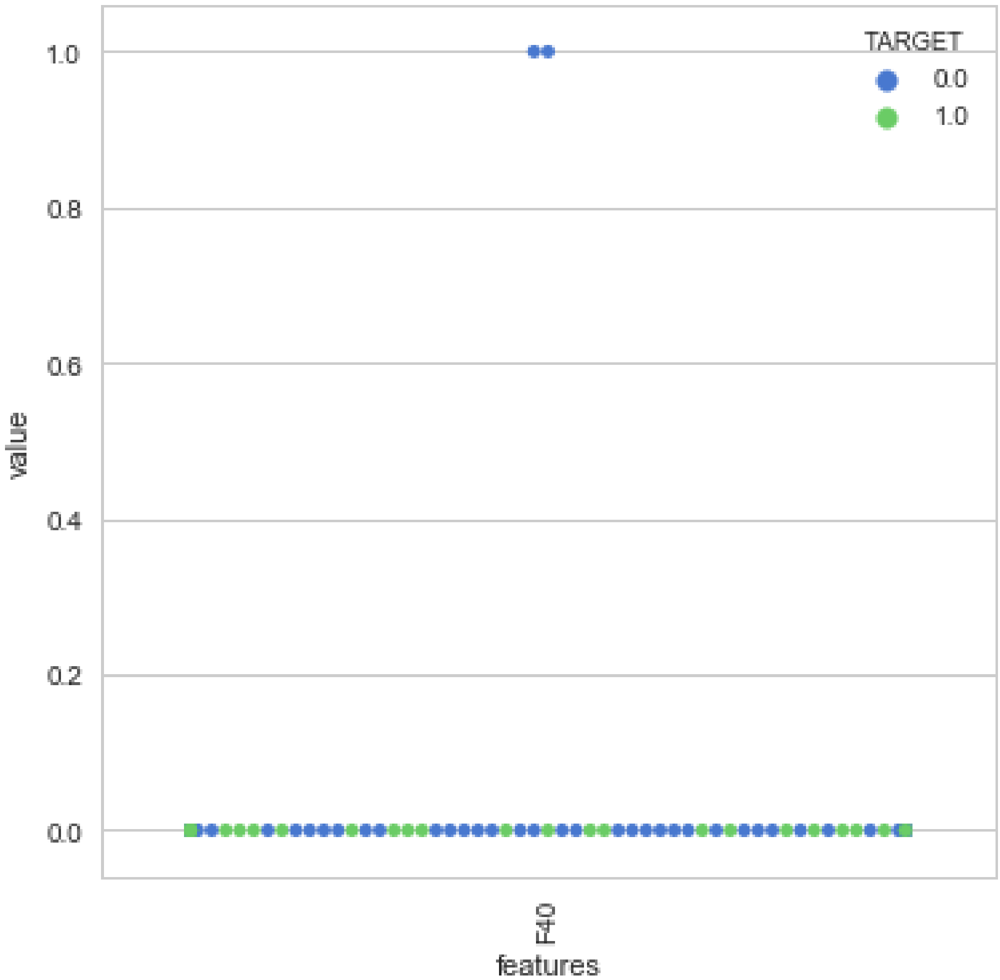


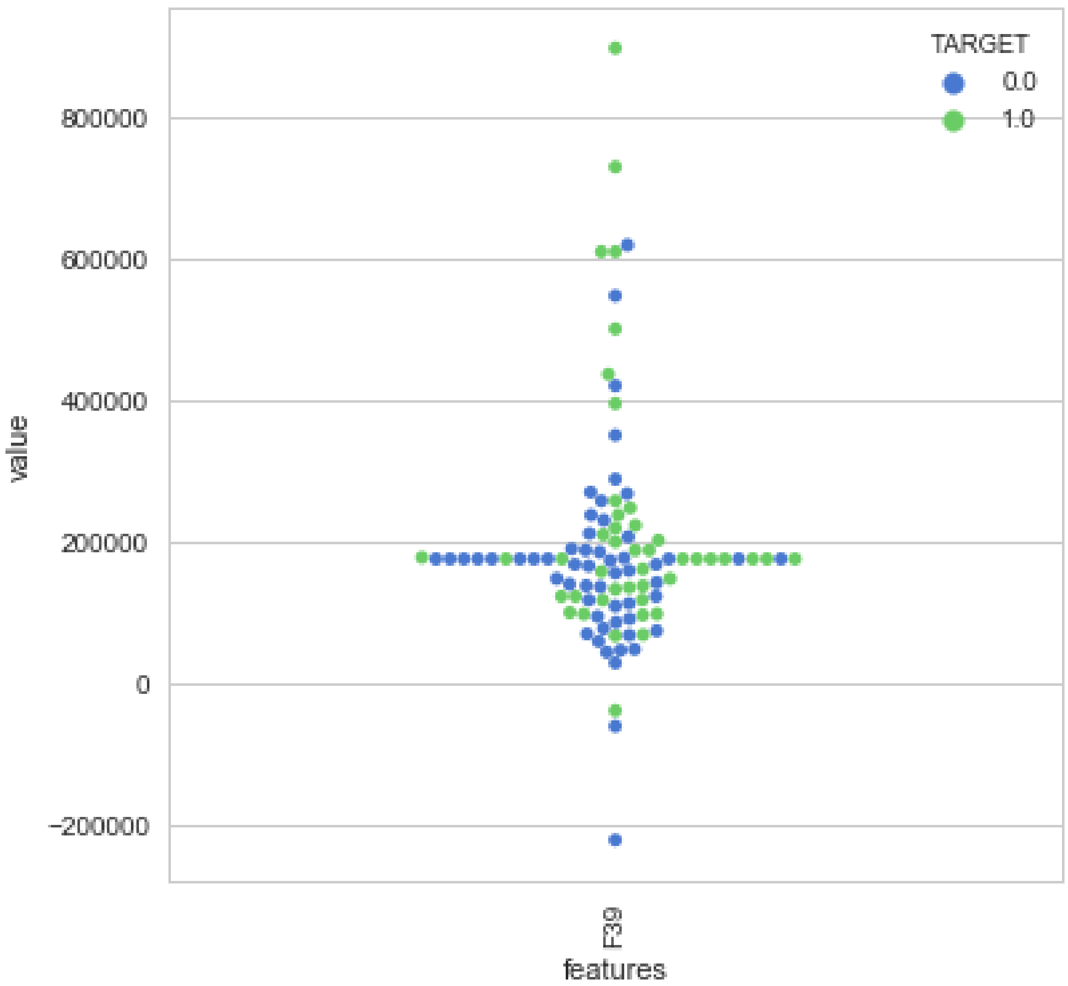
value
7
6
5
4
3
2
1
0
TARGET
0.0
1.0
F87
features


TARGET
0.0
1.0
5
4
3
2
1
0
value
F86
features

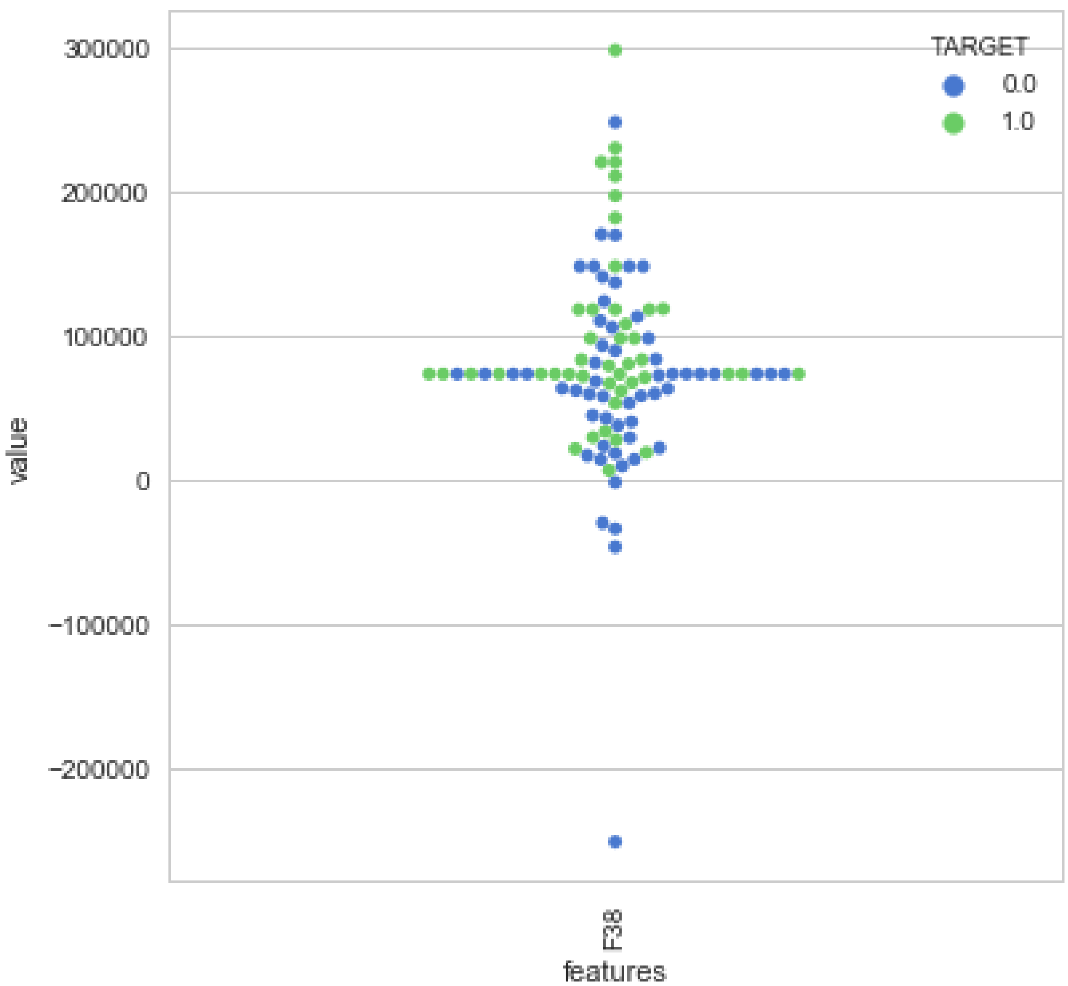

TARGET
0.0
1.0
100
80
60
40
20
0
value
F63
features


TARGET
0.0
1.0
40000
30000
20000
10000
0
value
F46
features

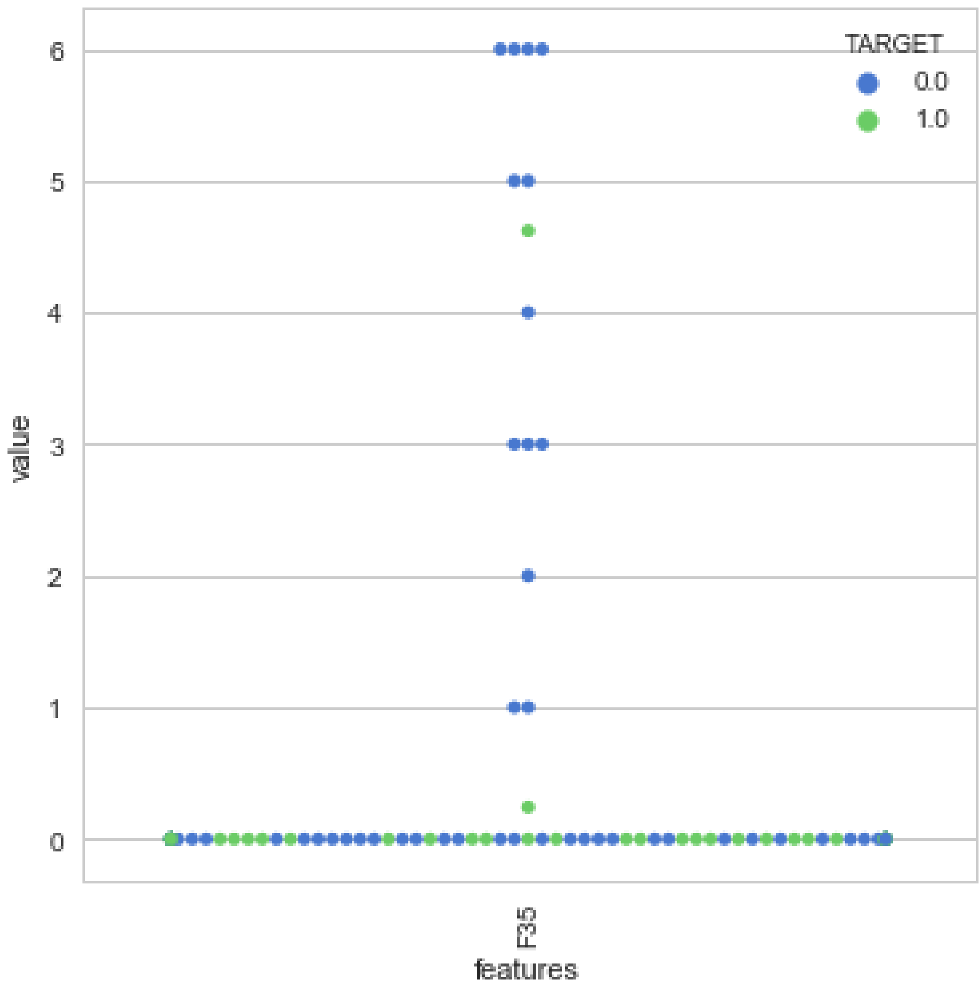

TARGET
0.0
1.0
1.0
0.8
0.6
0.4
0.2
0.0
value
F40
features

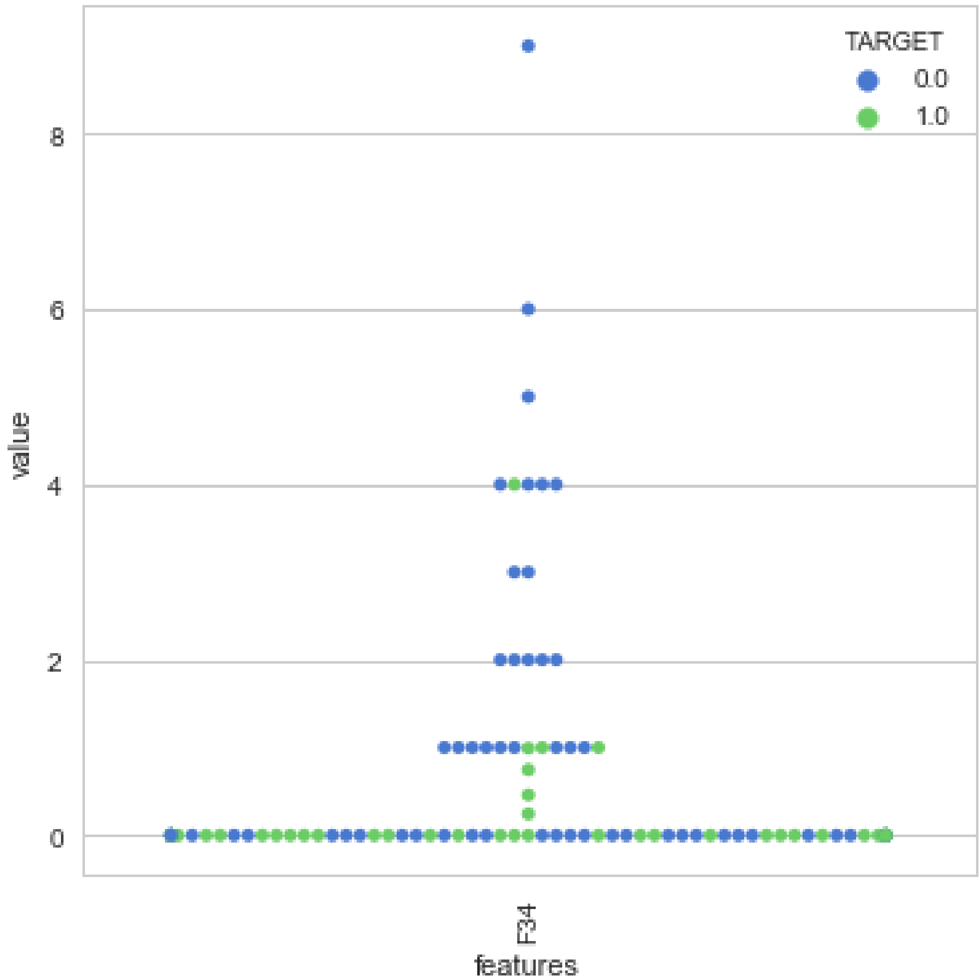


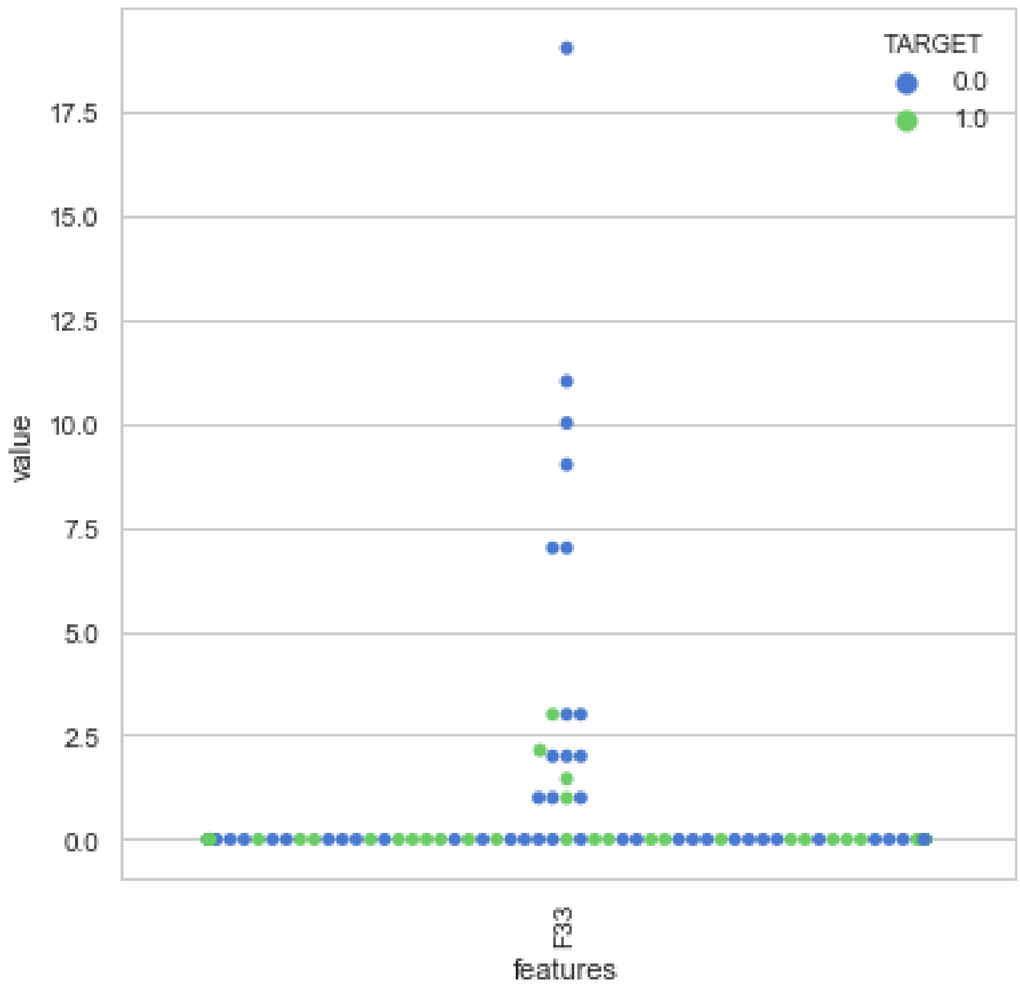
TARGET
0.0
1.0
800000
600000
400000
200000
0
−200000
value
F39
features



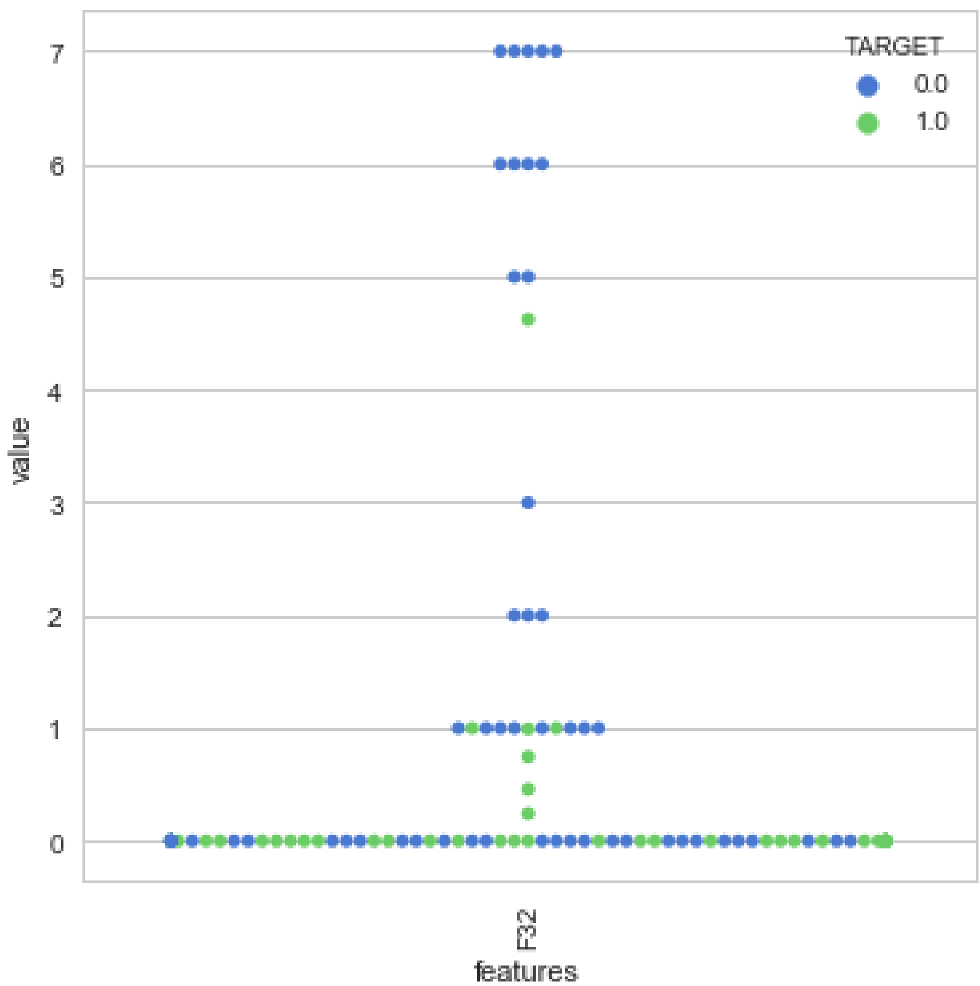
TARGET
0.0
1.0
300000
200000
100000
0
−100000
−200000
value
F38
features


TARGET
0.0
1.0
6
5
4
3
2
1
0
value
F35
features

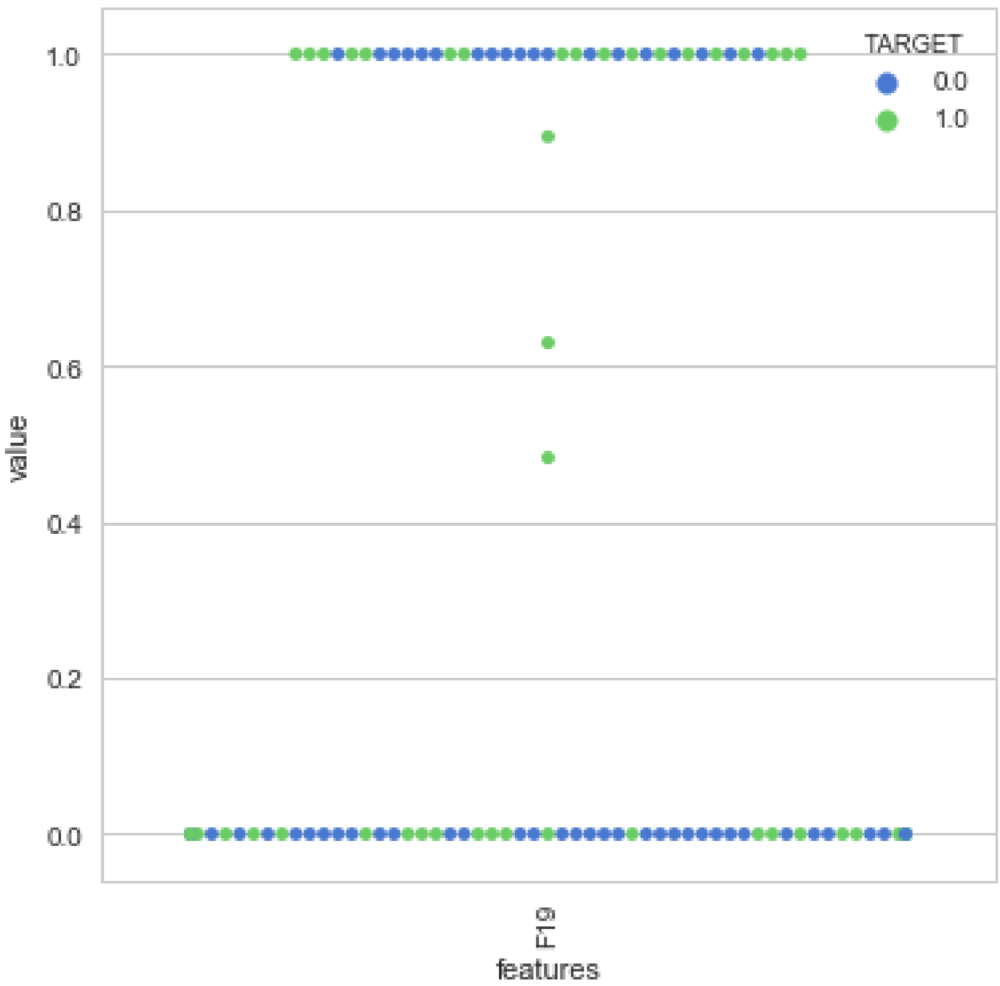

TARGET
0.0
1.0
8
6
4
2
0
value
F34
features



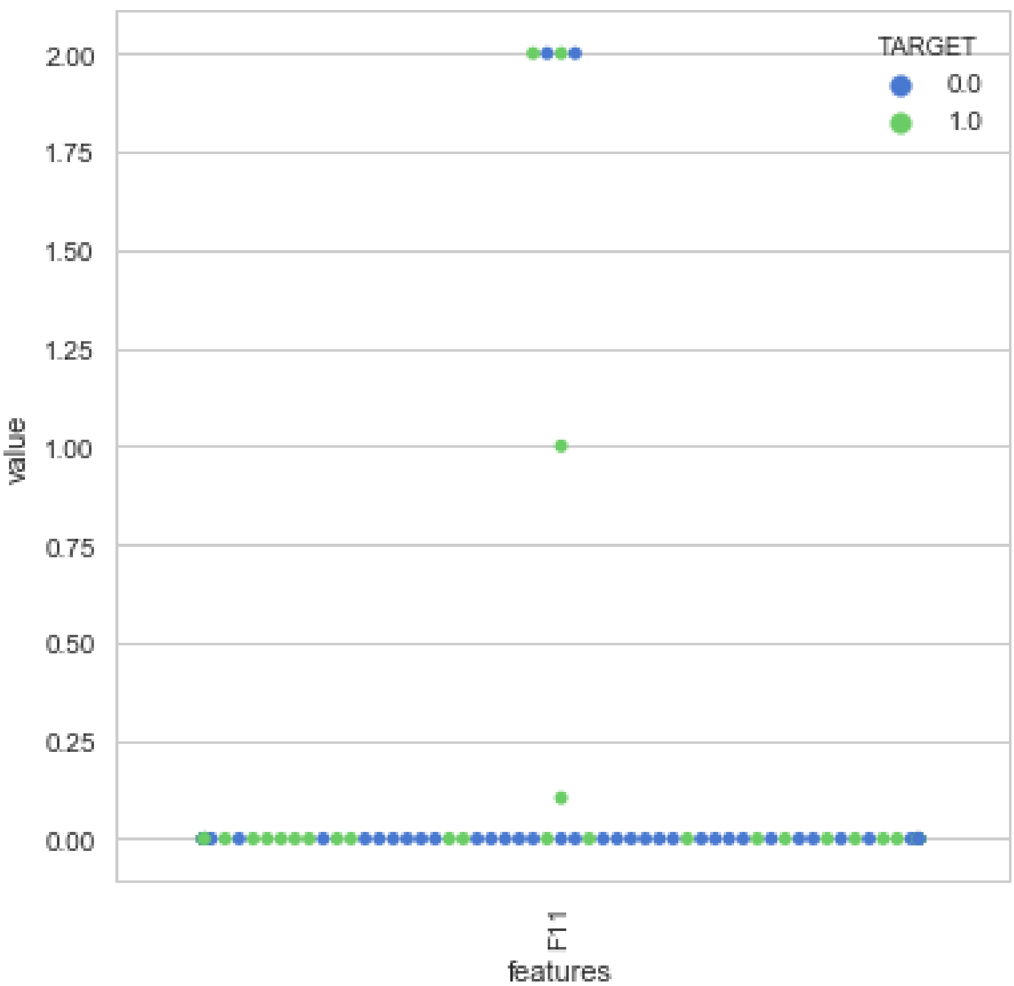
TARGET
0.0
1.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
value
F33
features



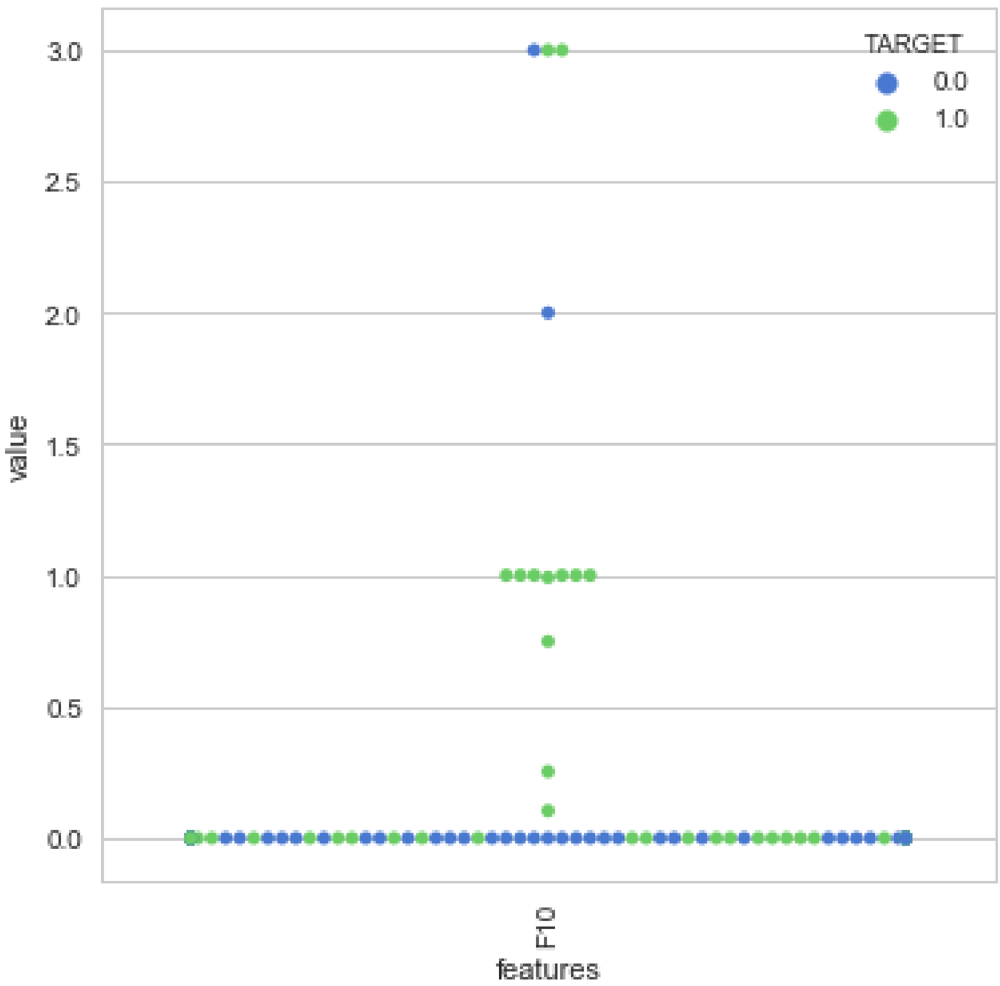
TARGET
0.0
1.0
7
6
5
4
3
2
1
0
value
F32
features



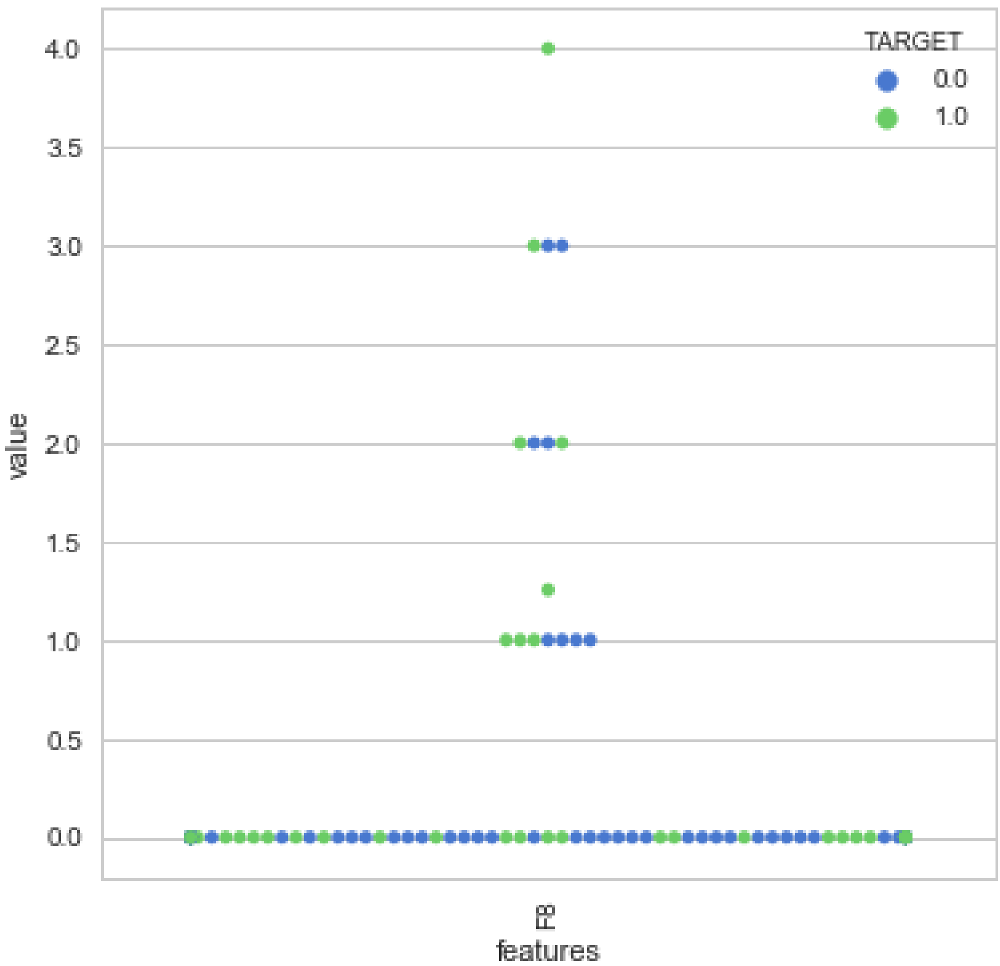
TARGET
0.0
1.0
1.0
0.8
0.6
0.4
0.2
0.0
value
F19
features



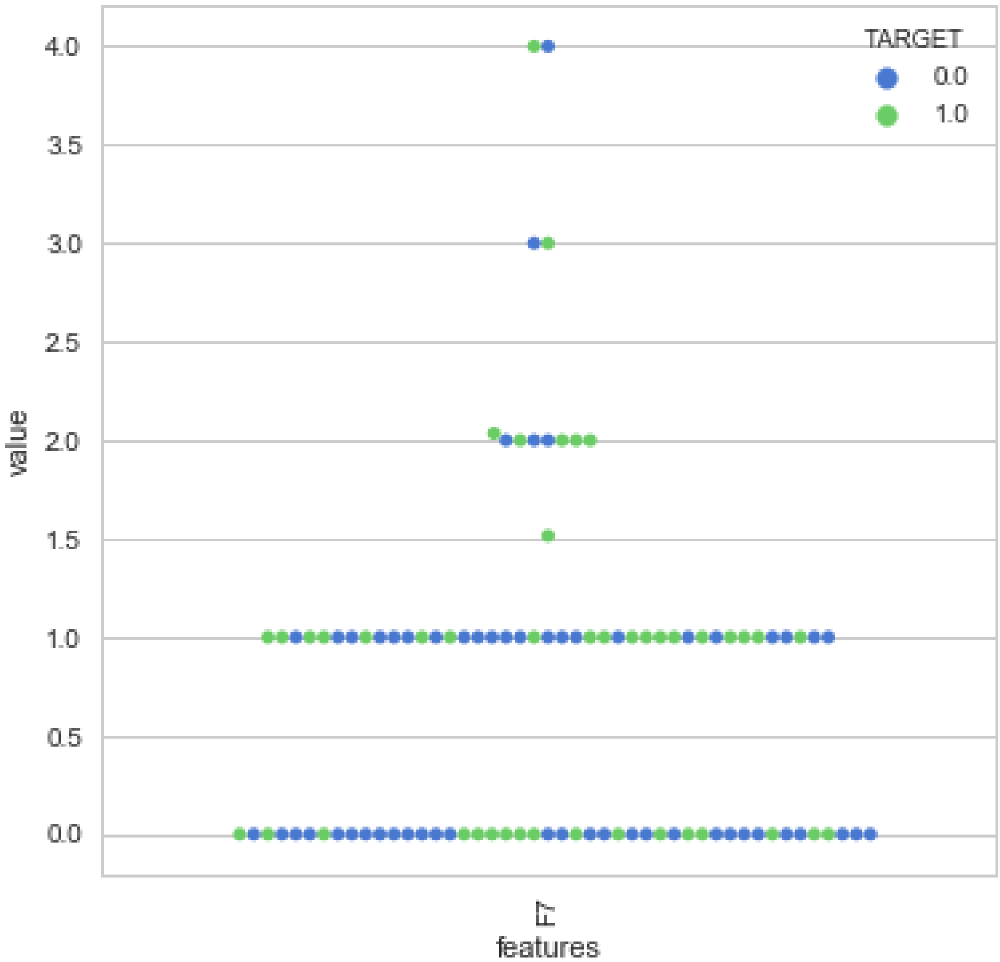
TARGET
0.0
1.0
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
value
F11
features

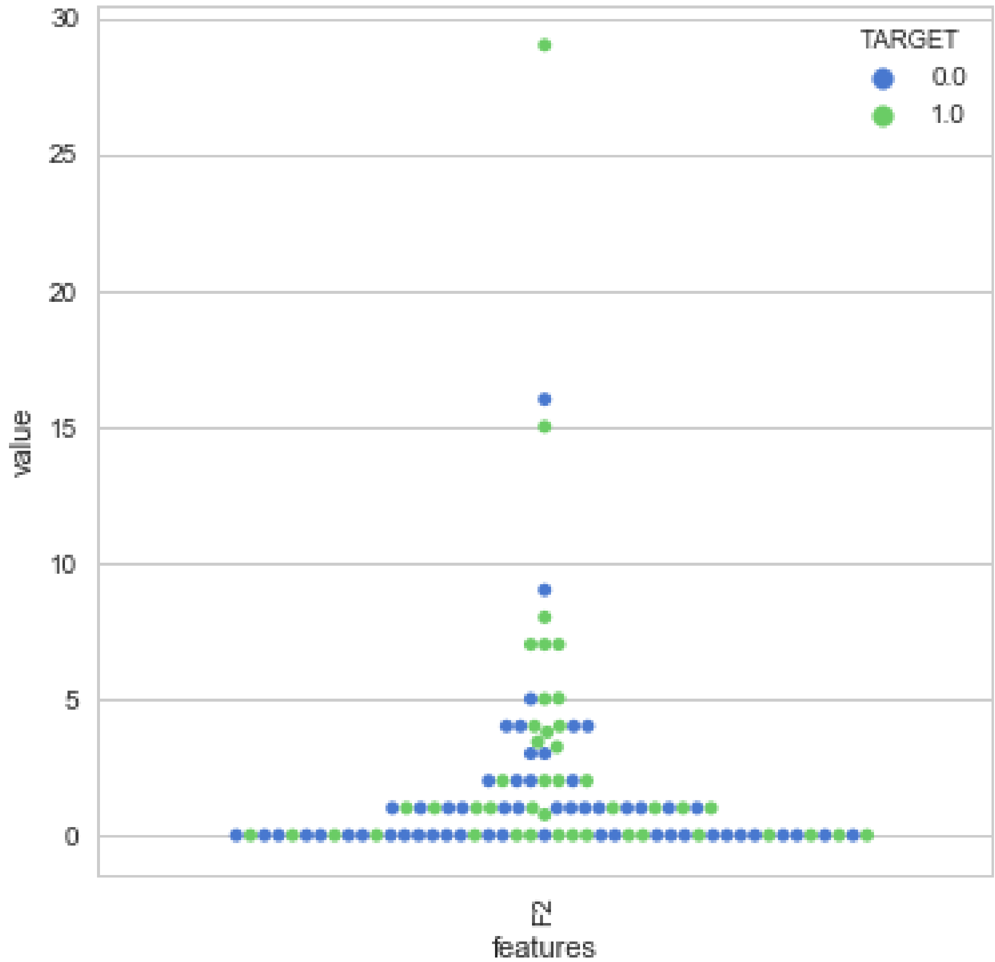


TARGET
0.0
1.0
3.0
2.5
2.0
1.5
1.0
0.5
0.0
value
F10
features



TARGET
0.0
1.0
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
value
F8
features


TARGET
0.0
1.0
value
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
F7
features


TARGET
0.0
1.0
value
30
25
20
15
10
5
0
F2
features