ANTI-MONEY LAUNDERING

DATA SCIENCE CASE STUDY

BY: KHALID KAHLOOT

PROBLEM STATEMENT

- Anti Money Laundering Refers To A Set Of Laws, Regulations, And Procedures Intended To Prevent Criminals From Disguising Illegally Obtained Funds As Legitimate Income.

- Though Anti-money-laundering Laws Cover A Relatively Limited Range Of Transactions And Criminal Behaviors, Their Implications Are Far-reaching.

PROBLEM STATEMENT

- Anti Money Laundering Refers To A Set
Procedures Intended To Prevent Cri
Obtained Funds As Legitimate Inc

- Though Anti-money-laundering Of
Transactions And Criminal Beh reaching.

The transactions are very very very similar to each other

ANOMALY

- Anomaly is a data point of interest in in this case it's something that stands out it's an unusual an unusual data point when a data generating process

- behaves and usually it results in an anomaly

FEATURES DESCRIPTION

feature name	description
user_id	Unique ID for the customer
request_id	Unique ID for the transfer request
target_recipient_id	Unique ID for recipient
date_user_created	Date when user was created
addr_country_code	Sender Address country code
addr_city	Sender Address city
recipient_country_code	Recipient country code
flag_personal_business	Business payment vs personal
payment_type	Payment method used to upload money
date_request_submitted	Date at which the customer set up a transfer
date_request_received	Date at which we received the customer?s money
date_request_transferred	Date at which we payed out to the recipient

FEATURES DESCRIPTION

payment_status	Payment status
ccy_send	Currency where the customer sends from
ccy_target	Currency where the customer sends to
transfer_to_self	Recipient type
sending_bank_name	Sending bank name
sending_bank_country	Sending bank country
payment_reference_classification	Reason of the transfer if the customer has entered it
device	Platform of the customer
transfer_sequence	How many transfers has the customer made so far
days_since_previous_req	Days since previous request
first_attempt_date	Date of the first transfer attempt
first_success_date	Date of the first successful transfer

STATISTICS

Dataset info

Number of variables	28
Number of observations	100000
Total Missing (%)	9.3%
Total size in memory	21.4 MiB
Average record size in memory	224.0 B

Variables types

Numeric	4
Categorical	16
Boolean	1
Date	6
Text (Unique)	1
Rejected	0
Unsupported	0

STATISTICS

- addr_city has a high cardinality: 16674 distinct values Warning
- addr_country_code has a high cardinality: 148 distinct values Warning
- date_request_cancelled has 77741 / 77.7% missing values Missing
- date_request_cancelled has a high cardinality: 20250 distinct values Warning
- date_request_received has 21386 / 21.4% missing values Missing
- date_request_transferred has 22624 / 22.6% missing values Missing
- days_since_previous_req has 15223 / 15.2% missing values Missing
- days_since_previous_req has 20883 / 20.9% zeros Zeros
- first_success_date has 4539 / 4.5% missing values Missing
- invoice_value is highly skewed ($\gamma 1 = 43.324$) Skewed
- invoice_value has 22263 / 22.3% missing values Missing
- invoice_value_cancel is highly skewed ($\gamma 1 = 28.846$) Skewed
- invoice_value_cancel has 77741 / 77.7% missing values Missing
- payment_type has 18777 / 18.8% missing values Missing
- recipient_country_code has a high cardinality: 68 distinct values Warning
- sending_bank_name has a high cardinality: 824 distinct values Warning
- target_recipient_id has a high cardinality: 94774 distinct values Warning
- transfer_sequence is highly skewed ($\gamma 1 = 20.528$) Skewed
- user_id has a high cardinality: 89436 distinct values Warning

STATISTICS - ADDR_CITY

Value	Count	Frequency (%)
LONDON	19219	19.2%
DUBLIN	980	1.0%
MADRID	962	1.0%
BRISTOL	845	0.8%
MANCHESTER	843	0.8%
EDINBURGH	793	0.8%
BERLIN	734	0.7%
PARIS	683	0.7%
TALLINN	636	0.6%
NEW YORK	635	0.6%
Other values (16663)	73668	73.7%

STATISTICS - ADDR_COUNTRY_CODE

Value	Count	Frequency (%)
GBR	52512	52.5%
USA	8306	8.3%
ESP	5631	5.6%
DEU	5158	5.2%
FRA	3609	3.6%
AUS	2967	3.0%
IRL	2216	2.2%
NLD	1618	1.6%
ITA	1487	1.5%
CHE	1351	1.4%
Other values (138)	15145	15.1%

STATISTICS

ccy_send
Categorical

Distinct count	20
Unique (%)	0.0%
Missing (%)	0.0%
Missing (n)	0

GBP	54430
EUR	24726
USD	8769
Other values (17)	12075

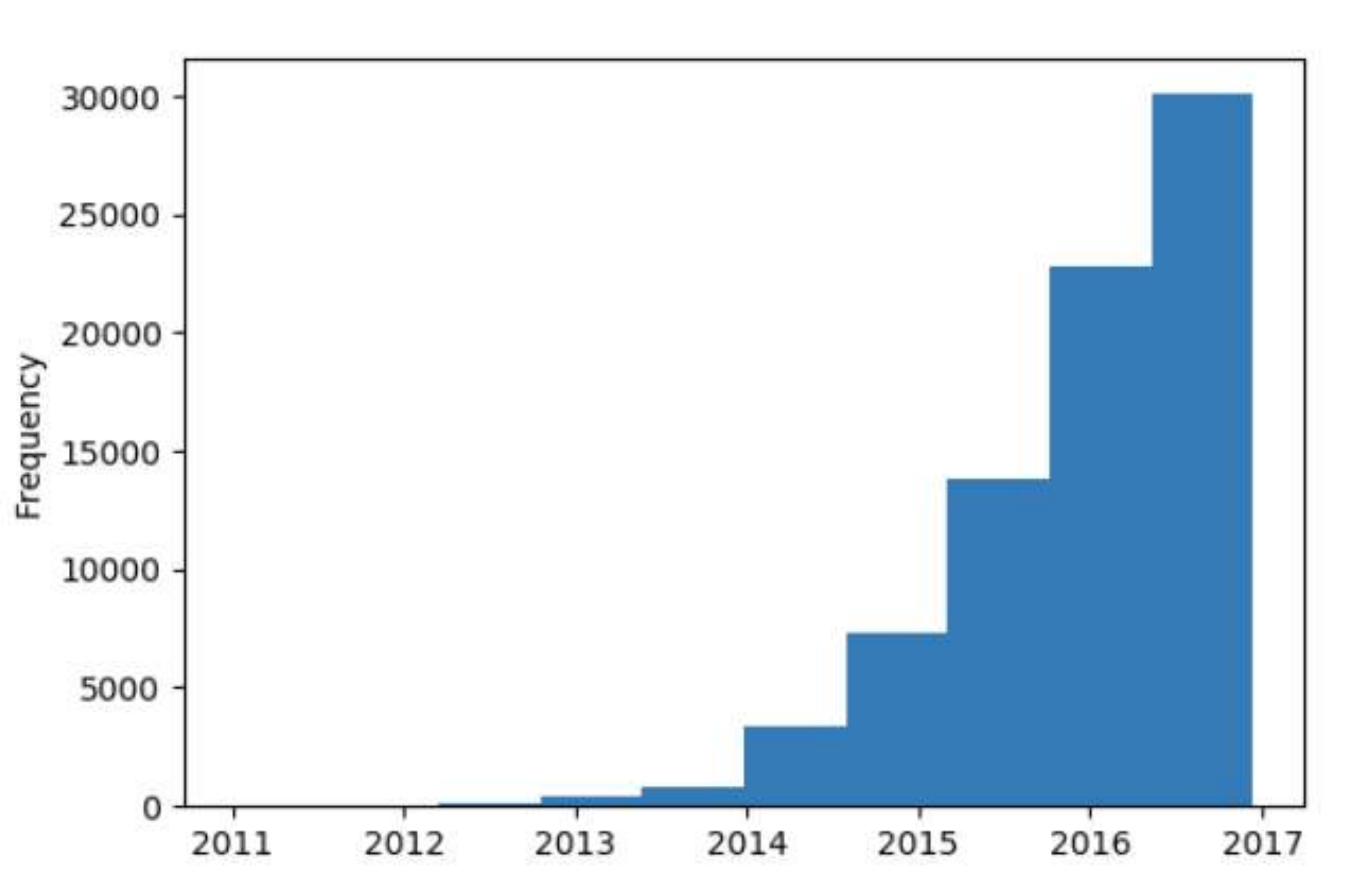
Toggle details

ccy_target
Categorical

Distinct count	41
Unique (%)	0.0%
Missing (%)	0.0%
Missing (n)	0

EUR	33579
GBP	17060
INR	12814
Other values (38)	36547

STATISTICS - DATE_REQUEST_RECEIVED



STATISTICS - DEVICE

Value	Count	Frequency (%)
Desktop Web	70187	70.2%
iOS App	14325	14.3%
Android App	8164	8.2%
Mobile Web	7324	7.3%

STATISTICS - PAYMENT_REFERENCE_CLASSIFICATION

Value	Count	Frequency (%)
blank	49755	49.8%
Other/unknown	28079	28.1%
invoice	5378	5.4%
monthly	3676	3.7%
family	2068	2.1%
rent	1542	1.5%
generic	1407	1.4%
self_transfer	1042	1.0%
gift	841	0.8%
house	807	0.8%
Other values (15)	5405	5.4%

STATISTICS - TRANSFER_TO_SELF

Value	Count	Frequency (%)
Self-recipient: Email match	26385	26.4%
Self-recipient: Exact name match	16435	16.4%
N.A. Recipient Email Unknown	15433	15.4%
Other Recipient	14278	14.3%
N.A. Sender or Recipient is business	13246	13.2%
Family (Last Matches, 1st name different)	10003	10.0%
Self-recipient: Name match	4220	4.2%

ANOMALY DETECTION

- Isolation Forest Algorithm

- I will use scikit-learn implementation, For large dataset,
- we can use spark implementation
- ***https://github.com/titicaca/spark-iforest***

- *Train* ***Unsupervised*** Isolation Forest, ***No labels are used***

ANOMALY DETECTION

Anomalous Transfers count

```
ds[ds['anomalous']==1]['anomalous'].count()
```

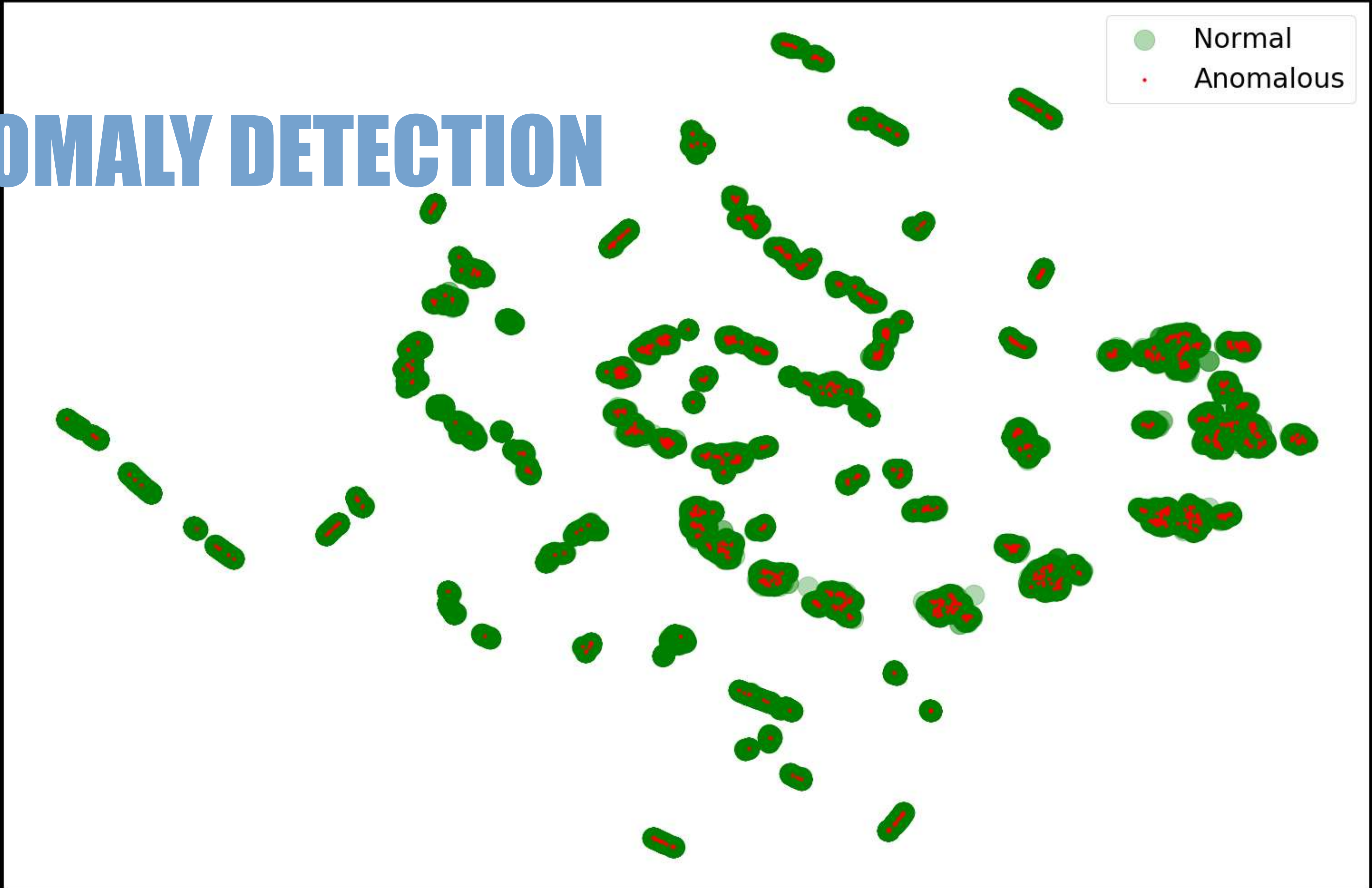
```
1500
```

Non-Anomalous Transfer count

```
ds[ds['anomalous']==0]['anomalous'].count()
```

```
98500
```

ANOMALY DETECTION



QUESTIONS TO ANSWER

1. Looking at the data, which customers would you deem risky in terms of Money Laundering based on their behavior?

Sort the Transfers by there Anomalous Behavior

```
ds.sort_values(['anomalous_score'])[['user_id', 'addr_country_code', 'addr_city
```

	user_id	addr_country_code	addr_city	anomalous
57362	69fd02c4fbd5bfa6533f7a5eac3bd81c	FIN	HELSINKI	1
48695	69fd02c4fbd5bfa6533f7a5eac3bd81c	FIN	HELSINKI	1
5779	69fd02c4fbd5bfa6533f7a5eac3bd81c	FIN	HELSINKI	1
35815	69fd02c4fbd5bfa6533f7a5eac3bd81c	FIN	HELSINKI	1
47108	69fd02c4fbd5bfa6533f7a5eac3bd81c	FIN	HELSINKI	1
14917	69fd02c4fbd5bfa6533f7a5eac3bd81c	FIN	HELSINKI	1
15194	69fd02c4fbd5bfa6533f7a5eac3bd81c	FIN	HELSINKI	1
10023	69fd02c4fbd5bfa6533f7a5eac3bd81c	FIN	HELSINKI	1
34178	69fd02c4fbd5bfa6533f7a5eac3bd81c	FIN	HELSINKI	1
44262	69fd02c4fbd5bfa6533f7a5eac3bd81c	FIN	HELSINKI	1

QUESTIONS TO ANSWER

2. What kind of info would you like to acquire from/about these customers in order to trust our service to them or deny it? How would you go about getting this info?

Reporting the Missing features for those Customers

```
import pandas as pd
```

```
print('index', 'Number of Missing')
ds.sort_values(['anomalous_score']).isnull().sum(axis=1).head(10)
```

```
index Number of Missing

57362    2
48695    2
5779     2
35815    2
47108    2
14917    2
15194    2
10023    2
34178    2
44262    2
dtype: int64
```

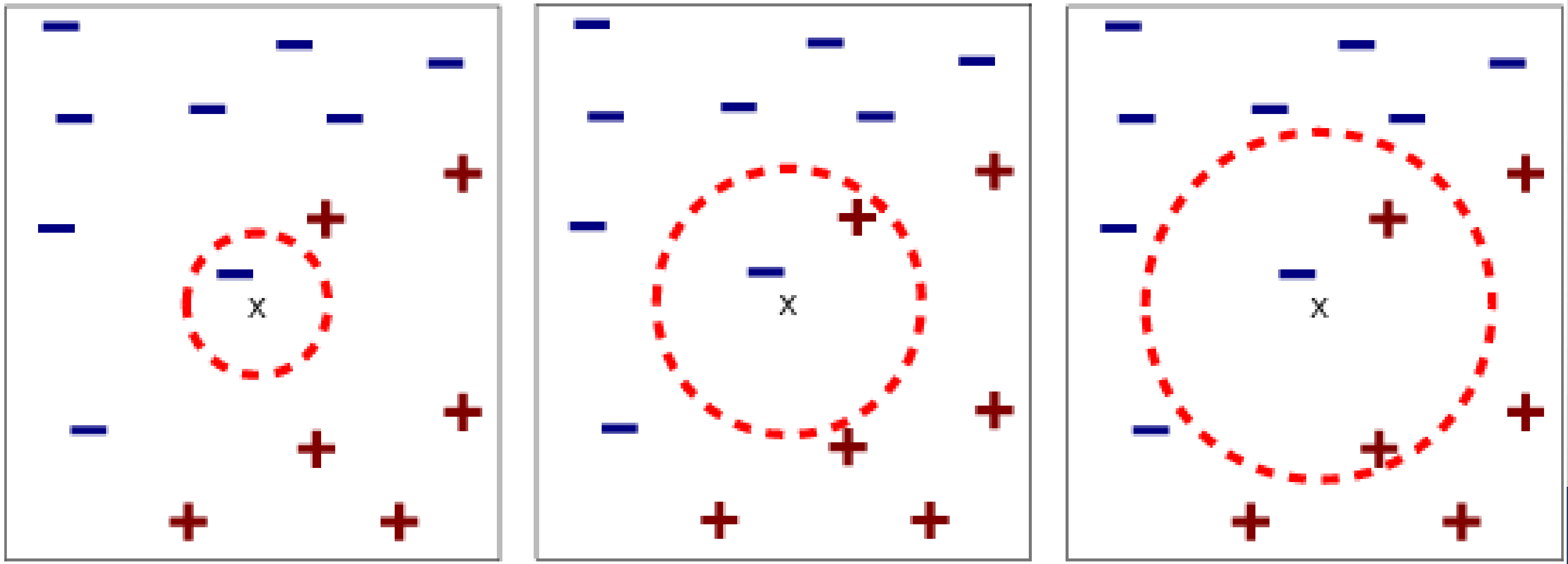
QUESTIONS TO ANSWER

```
user_id             69fd02c4fbd5bfa6533f7a5eac3bd81c
addr_country_code                                FIN
addr_city                                   HELSINKI
Name: 57362, dtype: object

missing values:
['date_request_cancelled', 'invoice_value_cancel']
-------------------------------------------------------

user_id             69fd02c4fbd5bfa6533f7a5eac3bd81c
addr_country_code                                FIN
addr_city                                   HELSINKI
Name: 48695, dtype: object

missing values:
['date_request_cancelled', 'invoice_value_cancel']
-------------------------------------------------------

user_id             69fd02c4fbd5bfa6533f7a5eac3bd81c
addr_country_code                                FIN
addr_city                                   HELSINKI
Name: 5779, dtype: object

missing values:
['date_request_cancelled', 'invoice_value_cancel']
-------------------------------------------------------

user_id             69fd02c4fbd5bfa6533f7a5eac3bd81c
addr_country_code                                FIN
addr_city                                   HELSINKI
Name: 35815, dtype: object

missing values:
['date_request_cancelled', 'invoice_value_cancel']
-------------------------------------------------------
```

SIMILAR TRANSFERS

Nearest neighbor Algorithm

example of K nearest neighbor.



(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

As an example of similarity, take the top anomalous transfer

Top 5 similar transfers to that instant

only the first one is anomalous

the top five similar transfers are comming from the same user_id

```
ds.iloc[similars[1][0]].head(5)
```

	user_id	request_id	target_recipient_id	anomalous_score	anomalous
57362	69fd02c4fbd5bfa6533f7a5eac3bd81c	8a0e22659ce2072497165a4ddefbe631	946bd1f5e726291c99dadd06a5189718	-0.014018	1
96952	69fd02c4fbd5bfa6533f7a5eac3bd81c	0ed6d004fbbdef3368ef24bddc199c02	86cafaba5ec5232913bc70eb24be2494	-0.007760	1
29164	69fd02c4fbd5bfa6533f7a5eac3bd81c	924c04a44f88568f9e43522b9c5dd299	d915efb5e361801178694fee0ecdf621	-0.012309	1
39007	69fd02c4fbd5bfa6533f7a5eac3bd81c	2a04354be989d9562b0c2c7e24208332	ca5c23f854fd03bd8cdb26522da7bd22	-0.007517	1
10440	69fd02c4fbd5bfa6533f7a5eac3bd81c	36d468dc36c6490edfcb16db5109639c	35d783e4ffc5042ce4d49e924e6a8448	-0.007837	1

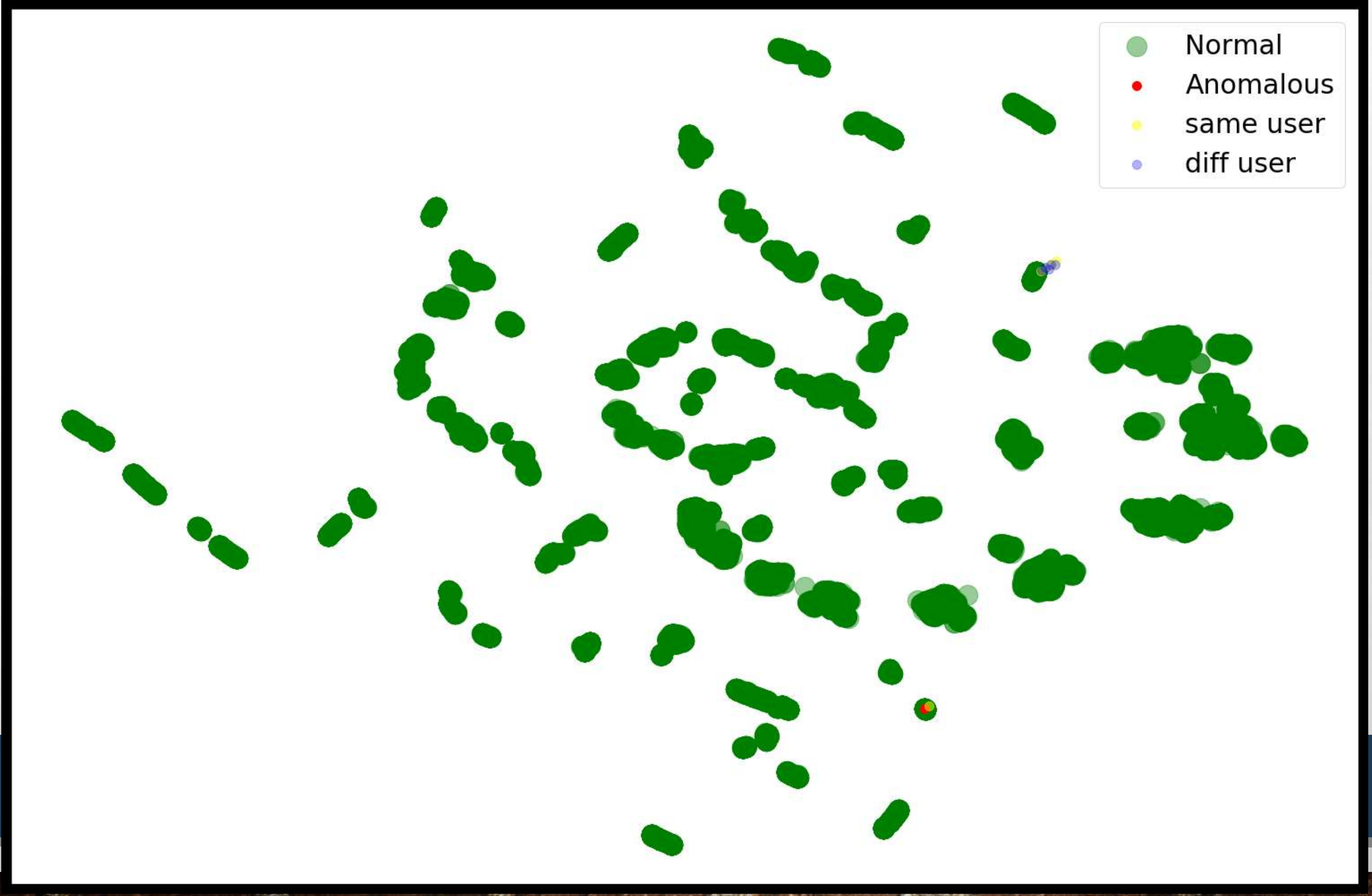
the top five similar transfers are comming from the different user_id

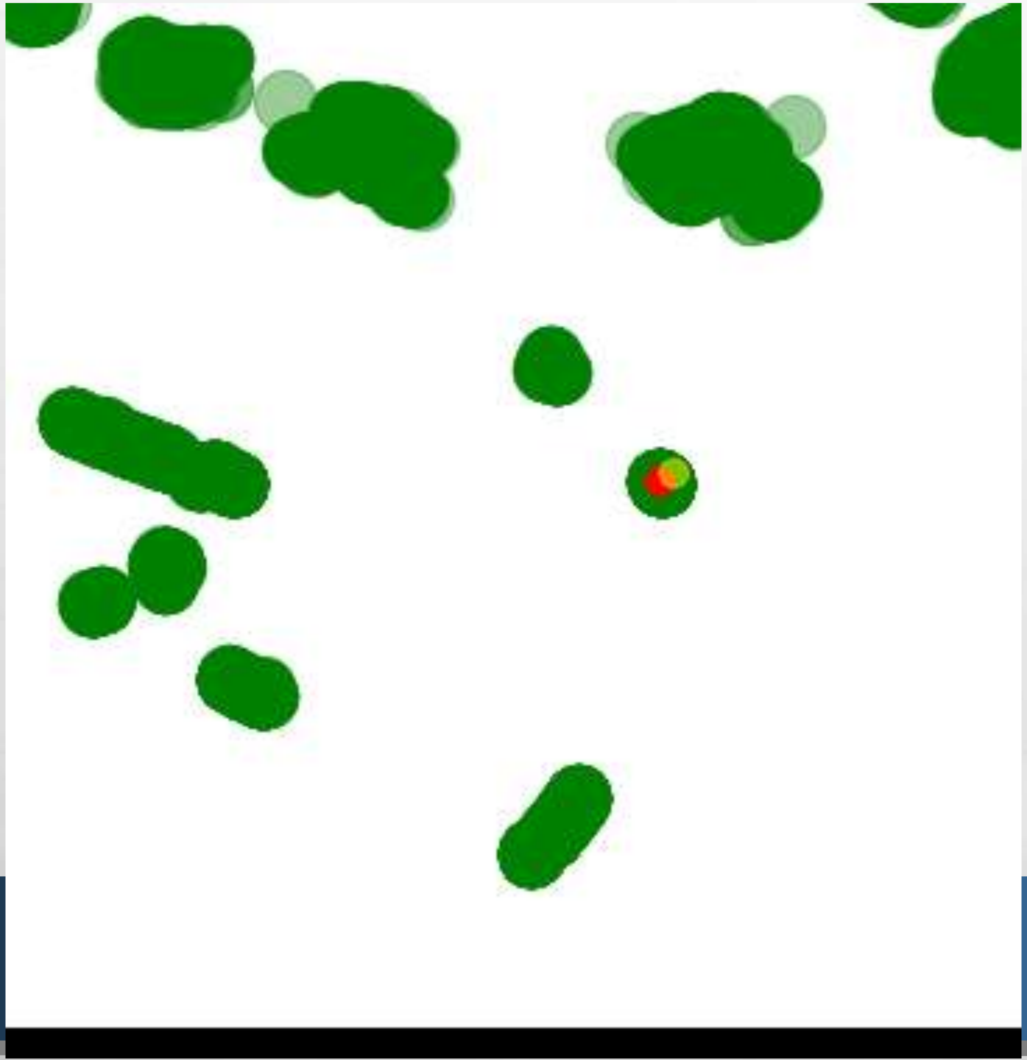
```
ds.iloc[similars[1][0]][ds.user_id!='69fd02c4fbd5bfa6533f7a5eac3bd81c'].head(5)
```

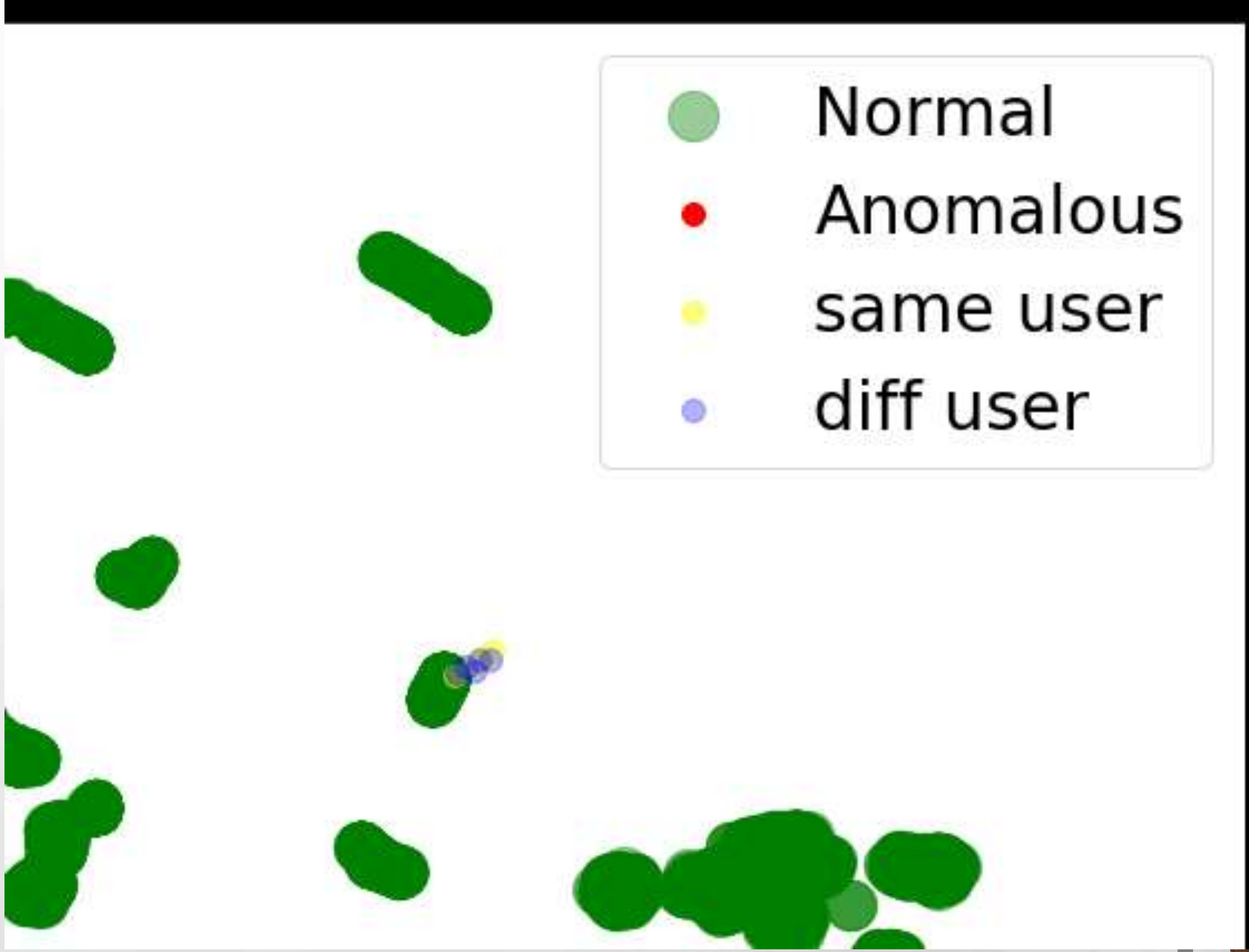
```
C:\ProgramData\Anaconda34\envs\gpu_env\lib\site-packages\ipykernel_launcher.py:1: UserWarning
exed to match DataFrame index.
  """Entry point for launching an IPython kernel.
```

	user_id	request_id	target_recipient_id	da
94646	b78c8ce3f52611c3df21da6b9effe911	13336575f77b6bcf8150d4706dfe5f70	616592c92d5822778f6dbd24e8a8a9ce	
93843	b78c8ce3f52611c3df21da6b9effe911	7a1844430c4dbe11e5e45a736f84e1bd	d1da4339464786aae1e6a5f9a41541be	
95568	b78c8ce3f52611c3df21da6b9effe911	e04933e9c3f9810a5f57ad4996db4033	52afe70eb729ee853cbc47061ff70dd2	
22197	b78c8ce3f52611c3df21da6b9effe911	c203eac78b04e58eeb062e51c2612c73	65a97954685381c7ae12f4c2de002a7e	
81866	b78c8ce3f52611c3df21da6b9effe911	67f4f4c889a4e28b2244a7f2ce3bf7e1	1af8e6781d9e37b2f330dddeba4dedf0	

anomalous_score	anomalous
0.00219	0
0.00219	0
0.00219	0
0.00219	0
0.00219	0


Normal
Anomalous
same user
diff user




Normal
Anomalous
same user
diff user

RISK ESTIMATOR

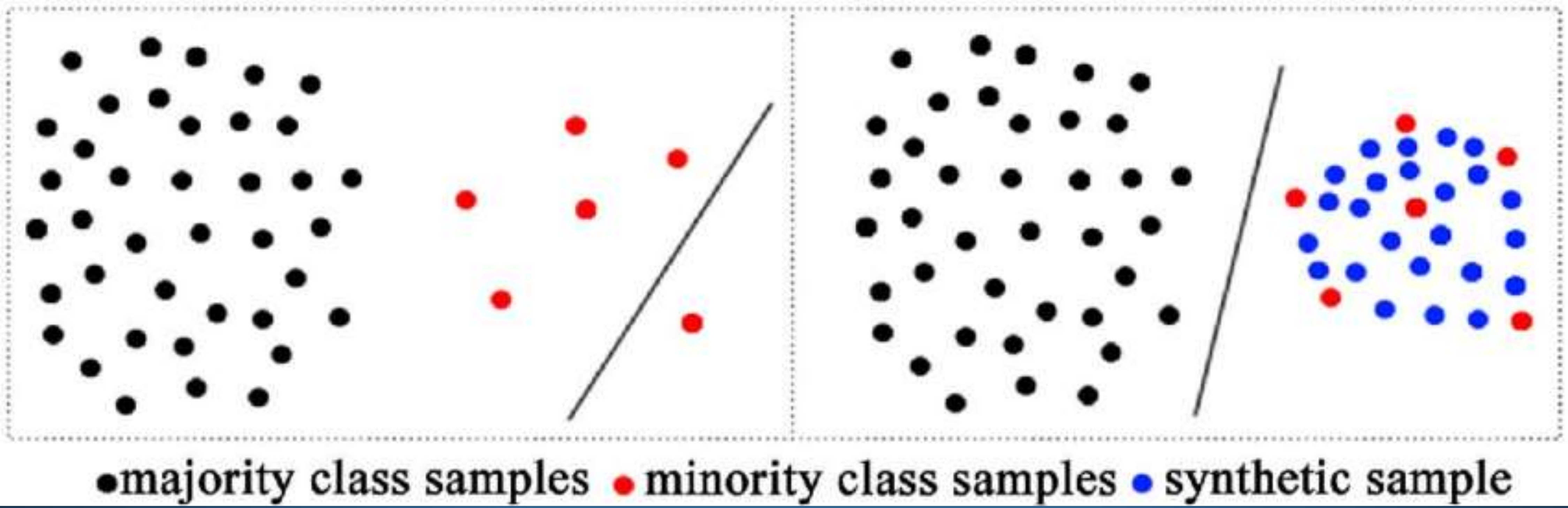
Imbalanced Dataset

```
ds['anomalous'].value_counts()
```

```
0    98500
1     1500
Name: anomalous, dtype: int64
```

RISK ESTIMATOR

We should do Over-sampling using **SMOTE** algorithm



RISK ESTIMATOR

Then fit another **supervised** Isolation Forest

Recall score is very important (score of catching the anomalous transfer)

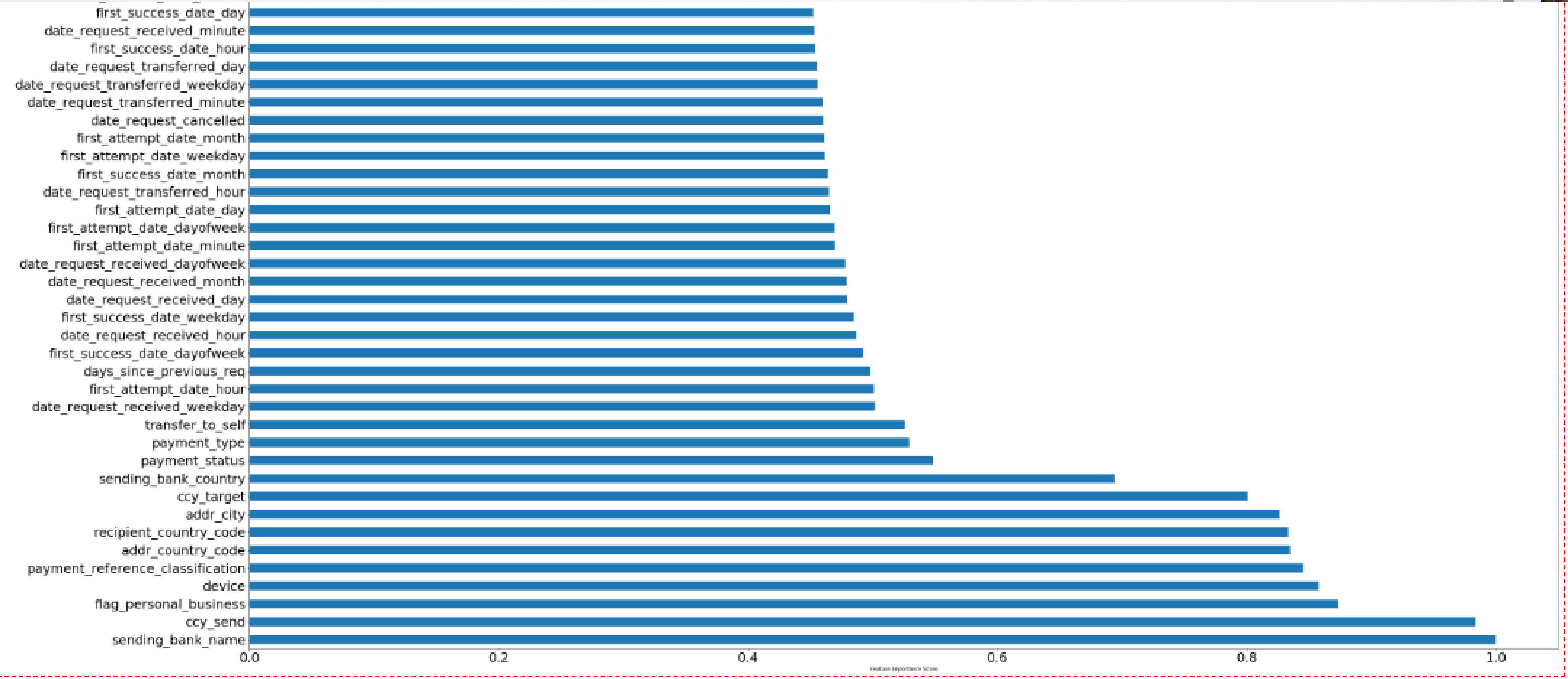
Holdout dataset

	precision	recall	f1-score	support
0	0.00	0.00	0.00	9850
1	0.01	1.00	0.03	150
micro avg	0.01	0.01	0.01	10000
macro avg	0.01	0.50	0.01	10000
weighted avg	0.00	0.01	0.00	10000

Train dataset

	precision	recall	f1-score	support
0	0.00	0.00	0.00	88650
1	0.01	1.00	0.03	1350
micro avg	0.01	0.01	0.01	90000
macro avg	0.01	0.50	0.01	90000
weighted avg	0.00	0.01	0.00	90000

FEATURES IMPORTANCE Perturbation Ranking



FEATURES IMPORTANCE

Perturbation Ranking

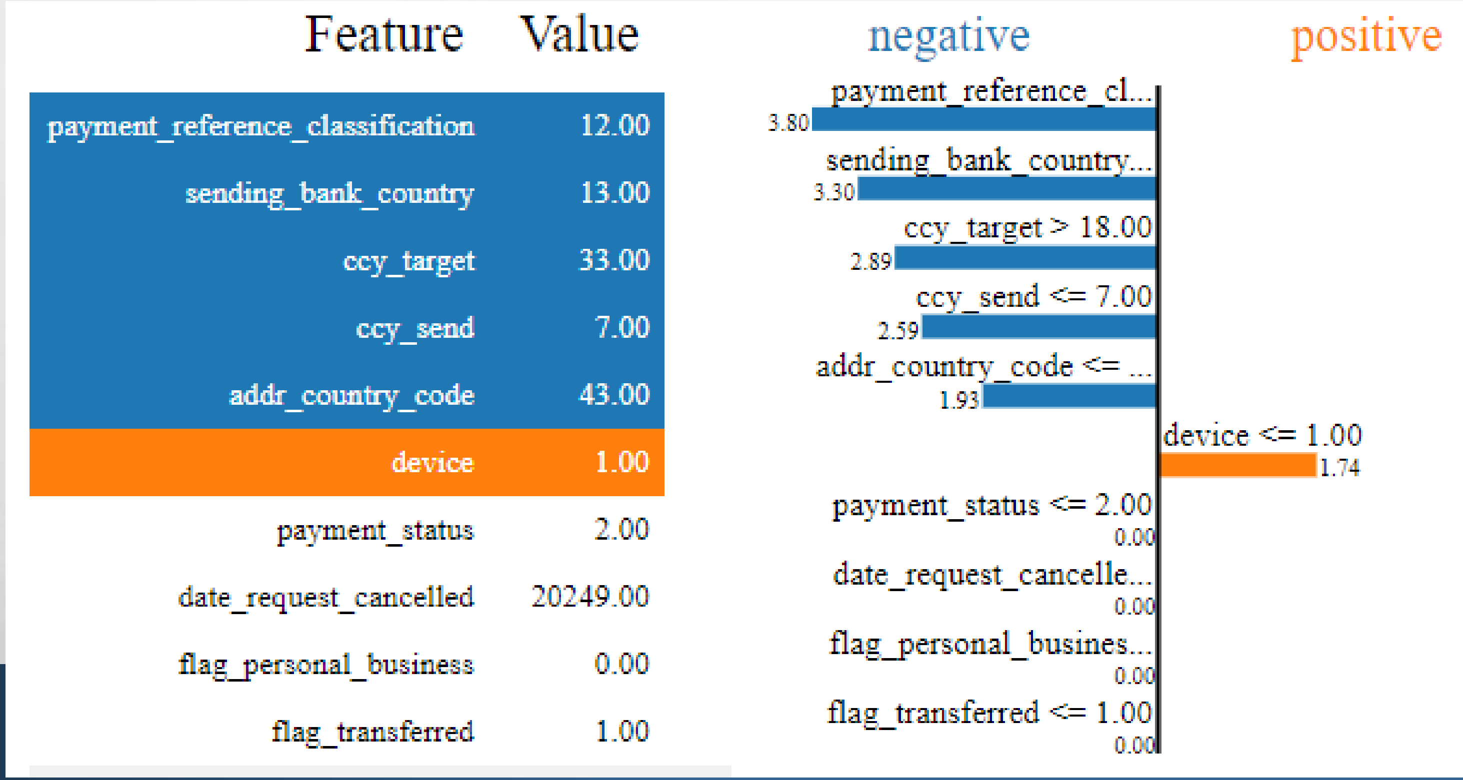
evaluating a trained model's prediction with each of the features.

How much effecting the prediction if we perturb and shuffle a feature.

	importance	error
feature names		
sending_bank_name	1.000000	0.001230
ccy_send	0.983704	0.001210
flag_personal_business	0.873557	0.001075
device	0.857824	0.001055
payment_reference_classification	0.845600	0.001040
addr_country_code	0.834821	0.001027
recipient_country_code	0.833436	0.001025
addr_city	0.826150	0.001016
ccy_target	0.800726	0.000985
sending_bank_country	0.693585	0.000853
payment_status	0.548267	0.000674
payment_type	0.529567	0.000651
transfer_to_self	0.526118	0.000647
date_request_received_weekday	0.501783	0.000617
first_attempt_date_hour	0.500923	0.000616
days_since_previous_req	0.498219	0.000613
first_success_date_dayofweek	0.492112	0.000605
date_request_received_hour	0.486904	0.000599
first_success_date_weekday	0.484964	0.000597

MODEL EXPLAINING

Anomalous Transfer



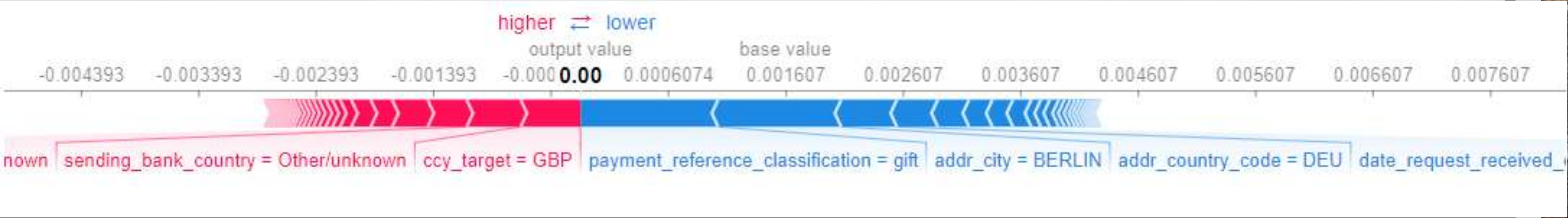
invoice_value	9416.22
invoice_value_cancel	NaN
flag_transferred	1
transfer_sequence	8712
days_since_previous_req	0
date_user_created_dayofweek	2
date_user_created_weekday	2
date_user_created_hour	11
date_user_created_minute	4
date_user_created_day	26
date_user_created_month	11
date_request_submitted_dayofweek	0
date_request_submitted_weekday	0
date_request_submitted_hour	13
date_request_submitted_minute	35
date_request_submitted_day	31
date_request_submitted_month	10
date_request_received_dayofweek	0
date_request_received_weekday	0
date_request_received_hour	7
date_request_received_minute	3
date_request_received_day	11
date_request_received_month	1
date_request_transferred_dayofweek	0
date_request_transferred_weekday	0
date_request_transferred_hour	15
date_request_transferred_minute	16
date_request_transferred_day	11
date_request_transferred_month	1
first_attempt_date_dayofweek	2
first_attempt_date_weekday	2
first_attempt_date_hour	11
first_attempt_date_minute	5
first_attempt_date_day	12
first_attempt_date_month	2
first_success_date_dayofweek	0
first_success_date_weekday	0
first_success_date_hour	8
first_success_date_minute	1
first_success_date_day	12
first_success_date_month	5
addr_country_code	FIN
addr_city	HELSINKI
recipient_country_code	NL
flag_personal_business	Business
payment_type	Bank Transfer
date_request_cancelled	NaN
payment_status	Transferred
ccy_send	EUR
ccy_target	SEK
transfer_to_self	N.A. Sender or Recipient is business
sending_bank_name	POHJOLA PANKKI OYJ (POHJOLA BANK PLC)
sending_bank_country	FI
payment_reference_classification	invoice
device	Desktop Web

Anomaly Threshold
-0.0052

MODEL EXPLAINING

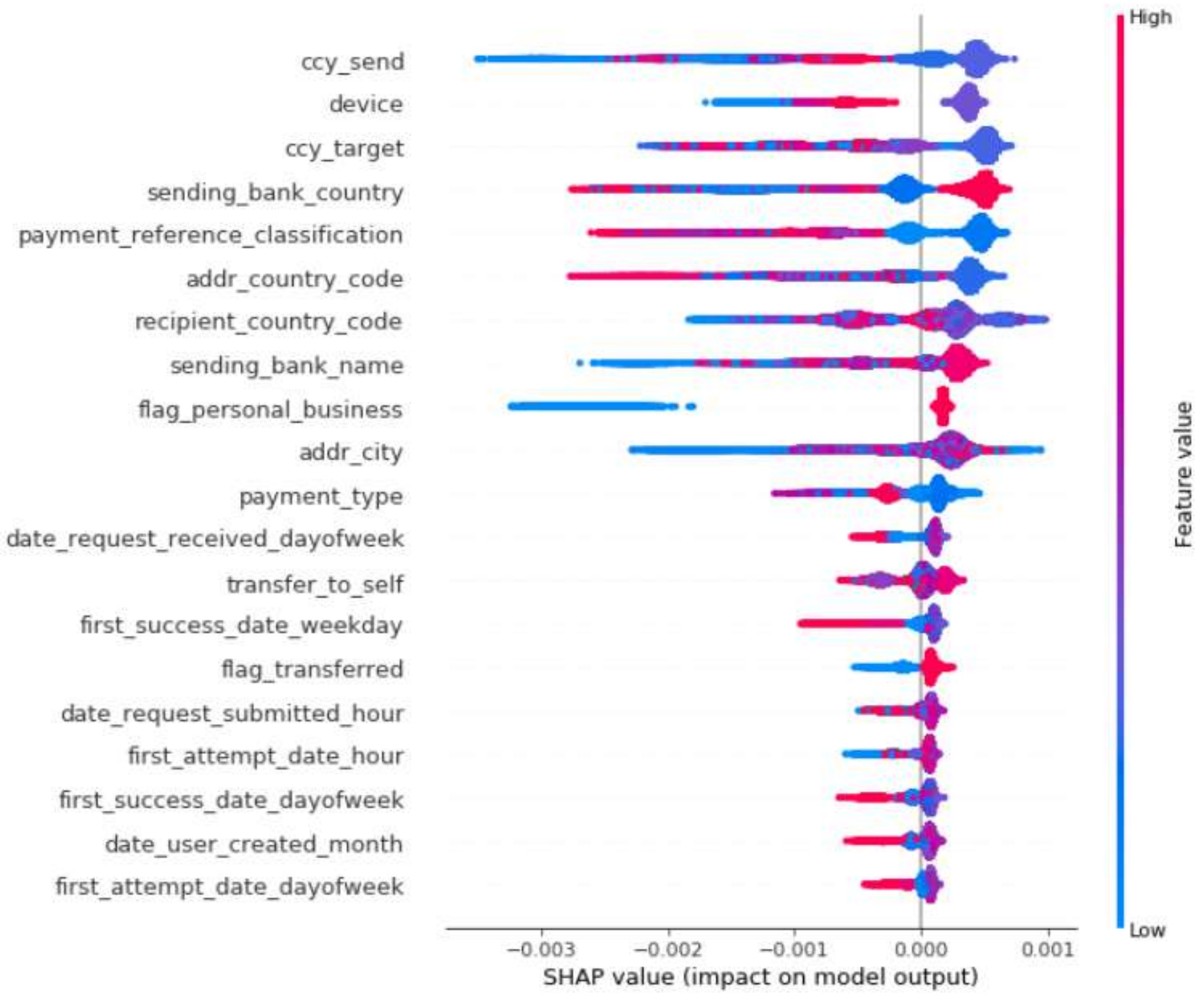
Anomalous Transfer

Not Anomalous Transfer



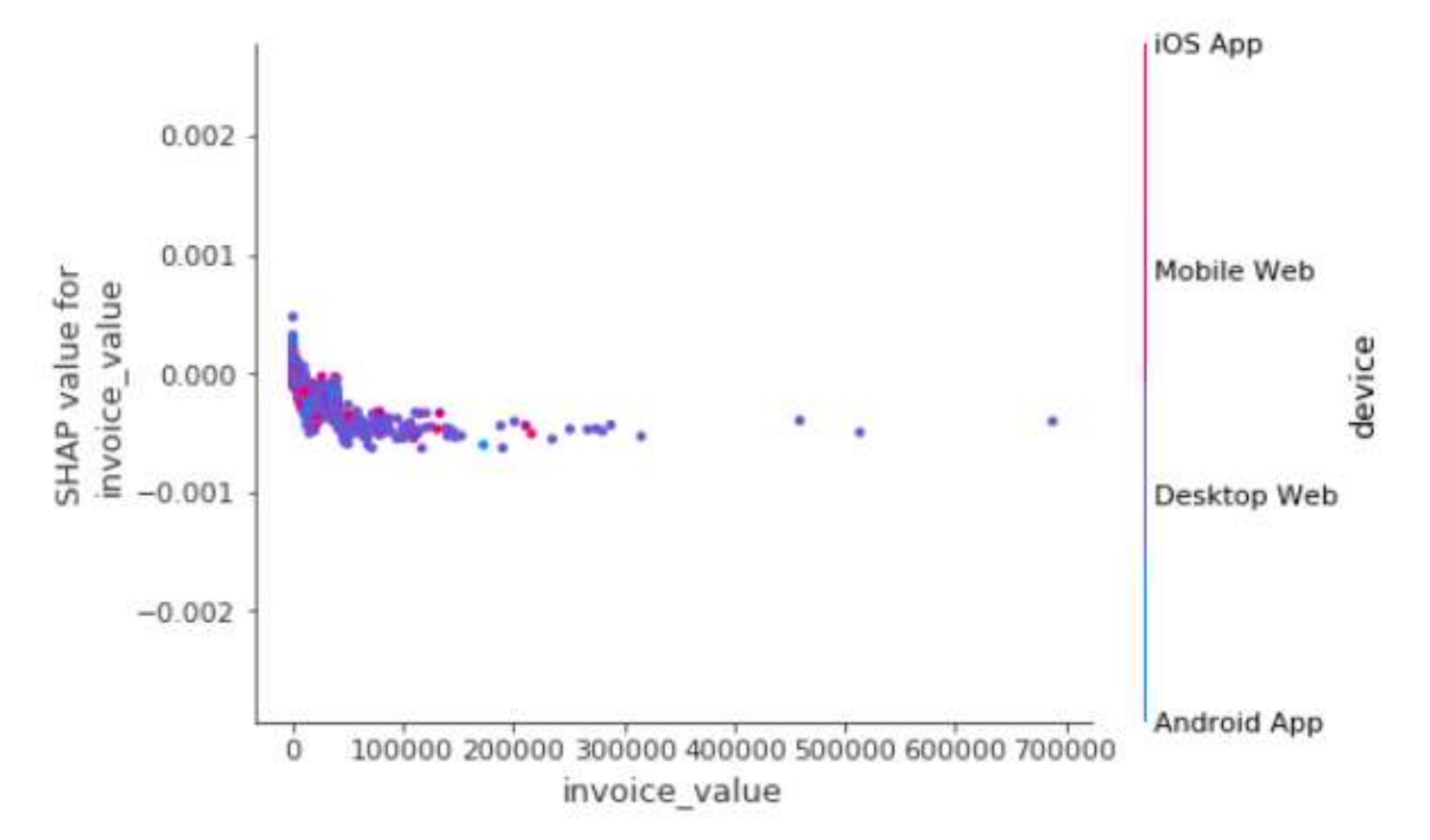
MODEL EXPLAINING

Anomaly Threshold
-0.0052



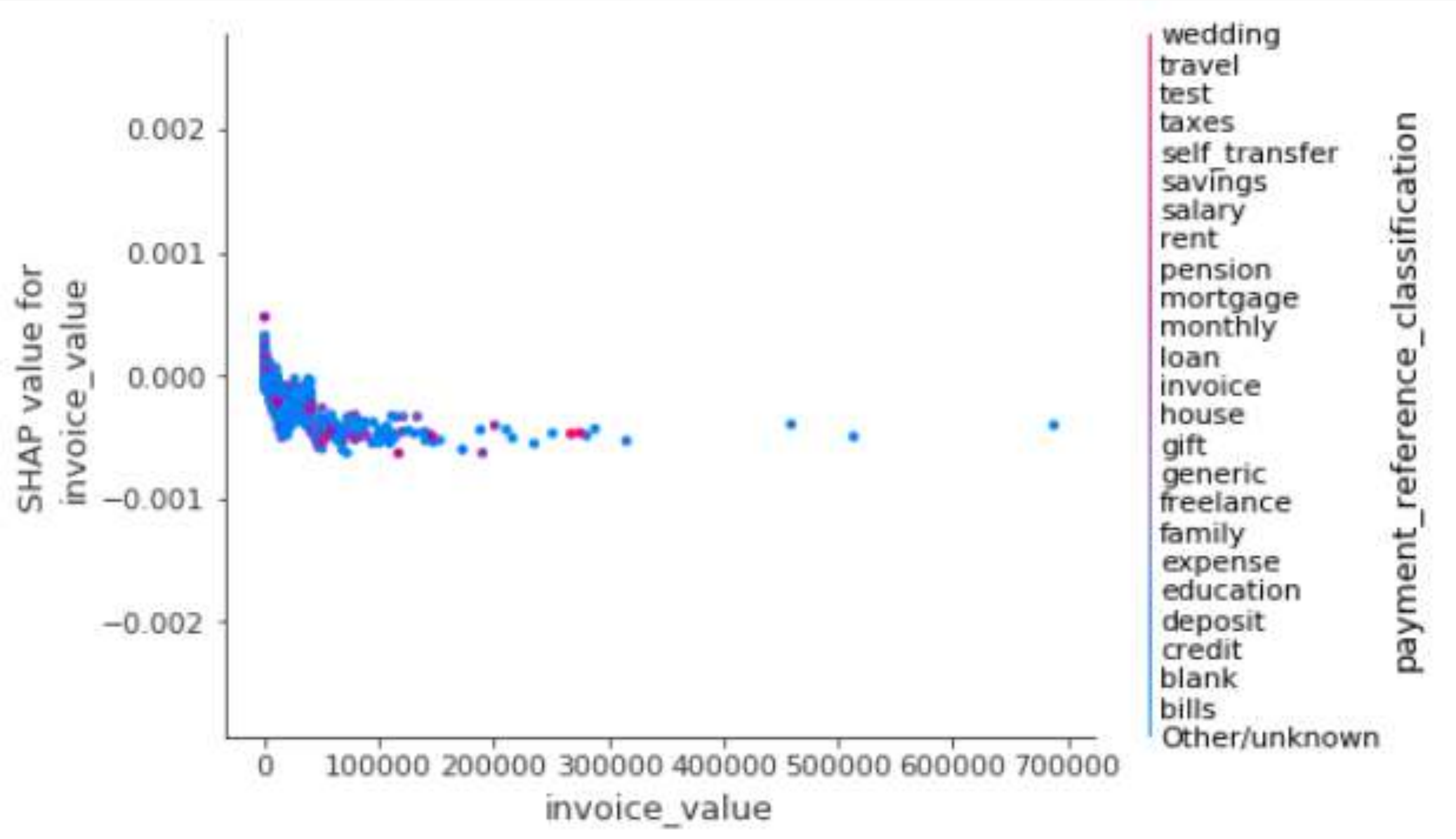
Anomaly Threshold
-0.0052

MODEL EXPLAINING



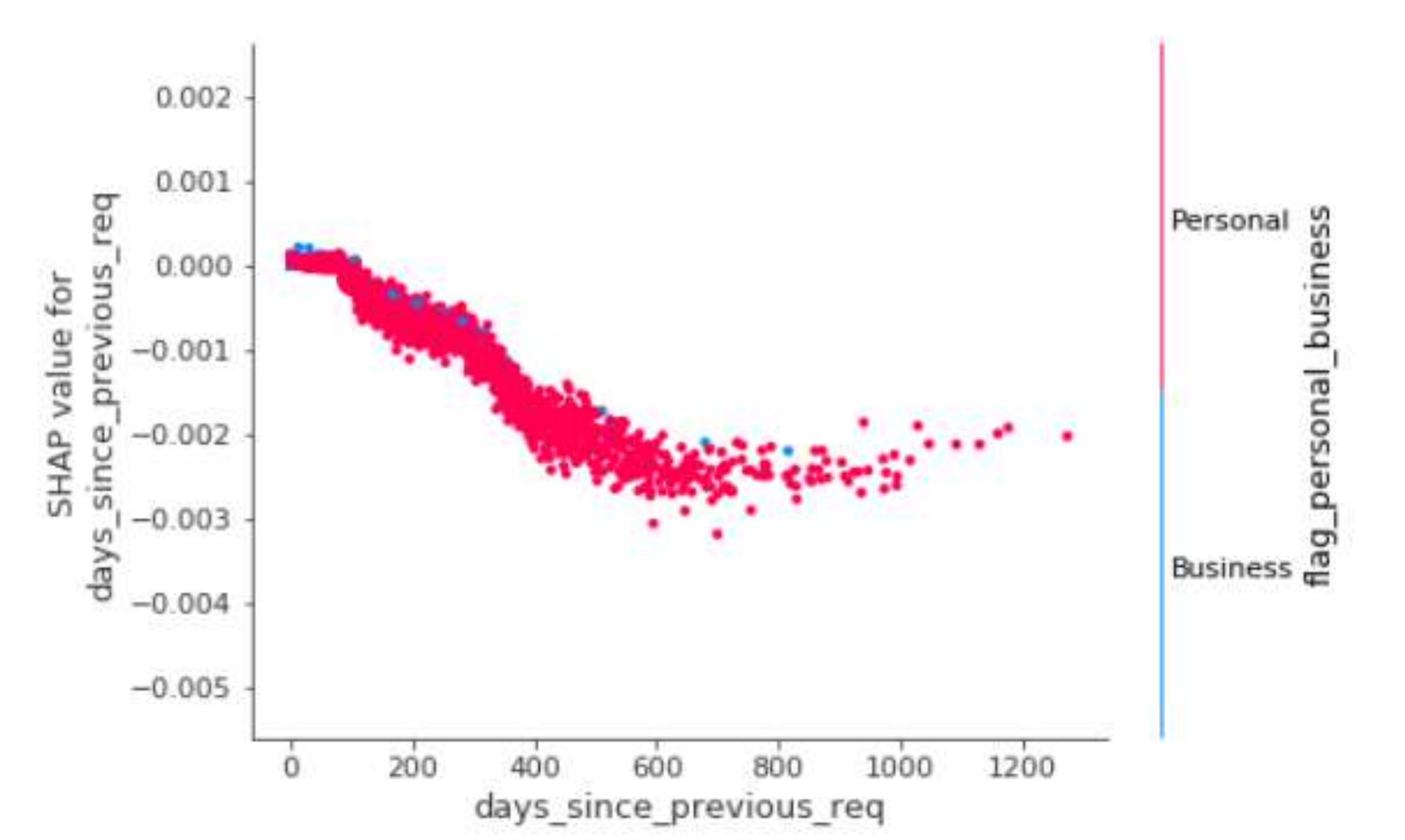
Anomaly Threshold
-0.0052

MODEL EXPLAINING



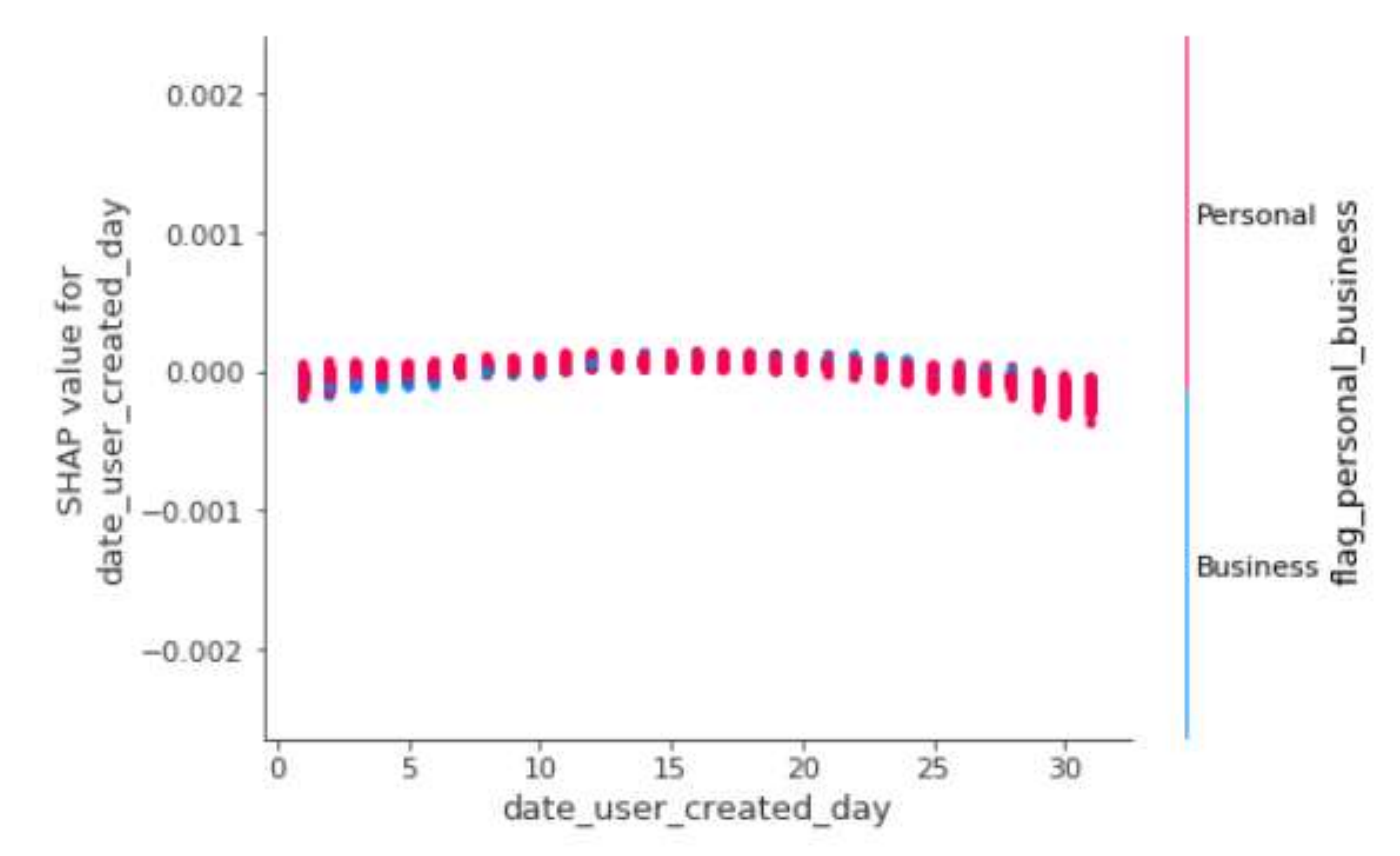
Anomaly Threshold
-0.0052

MODEL EXPLAINING



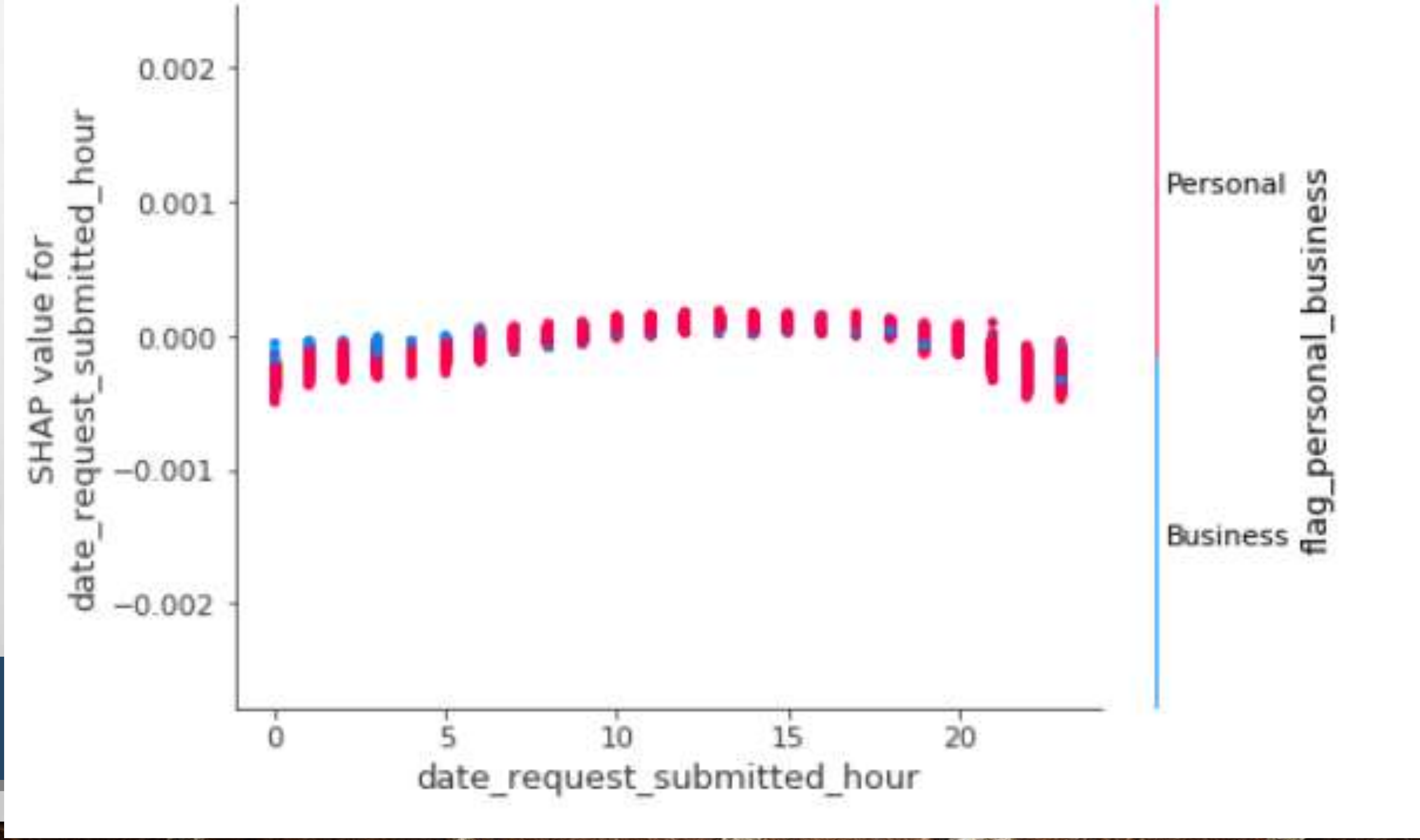
Anomaly Threshold
-0.0052

MODEL EXPLAINING



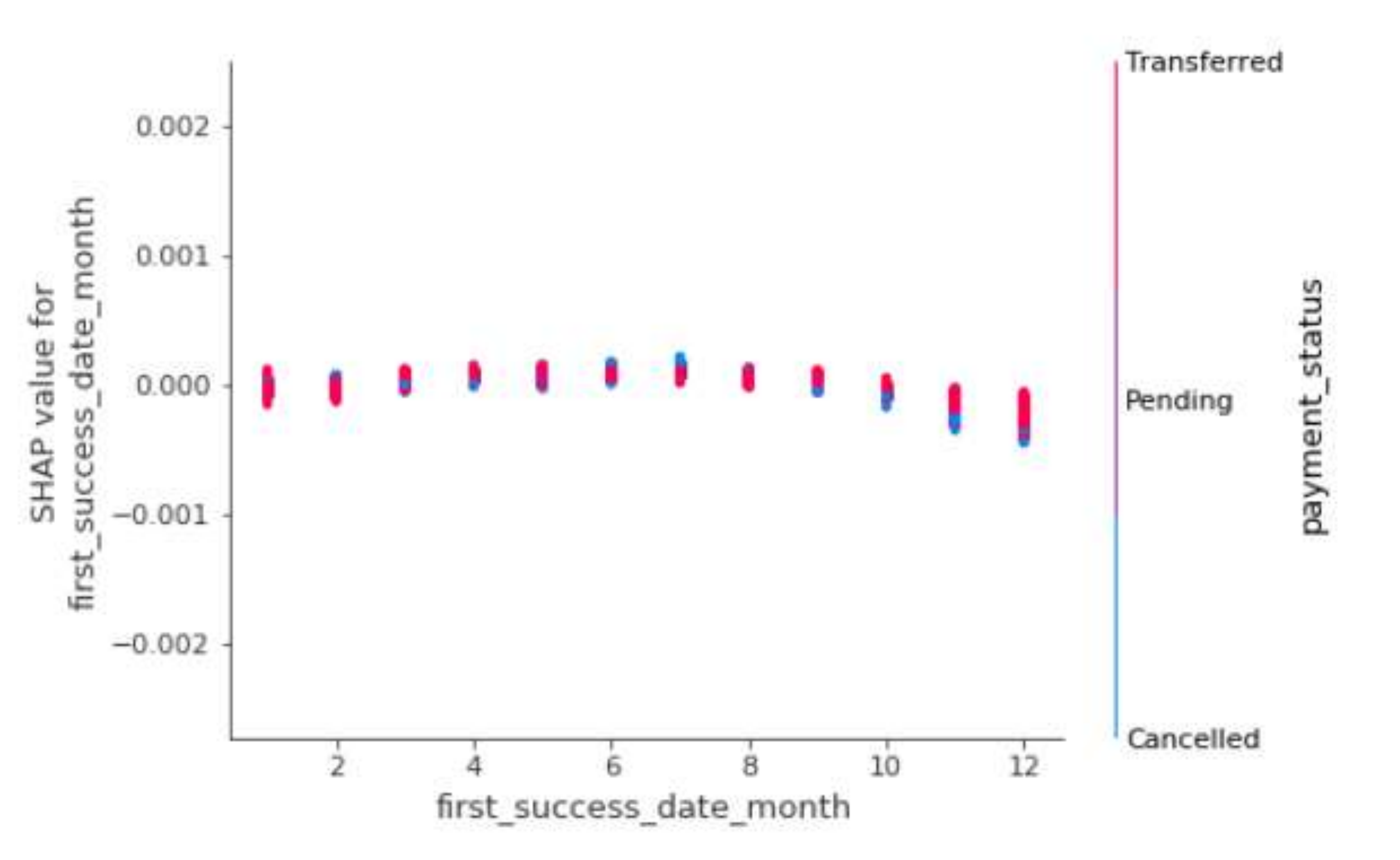
Anomaly Threshold
-0.0052

MODEL EXPLAINING



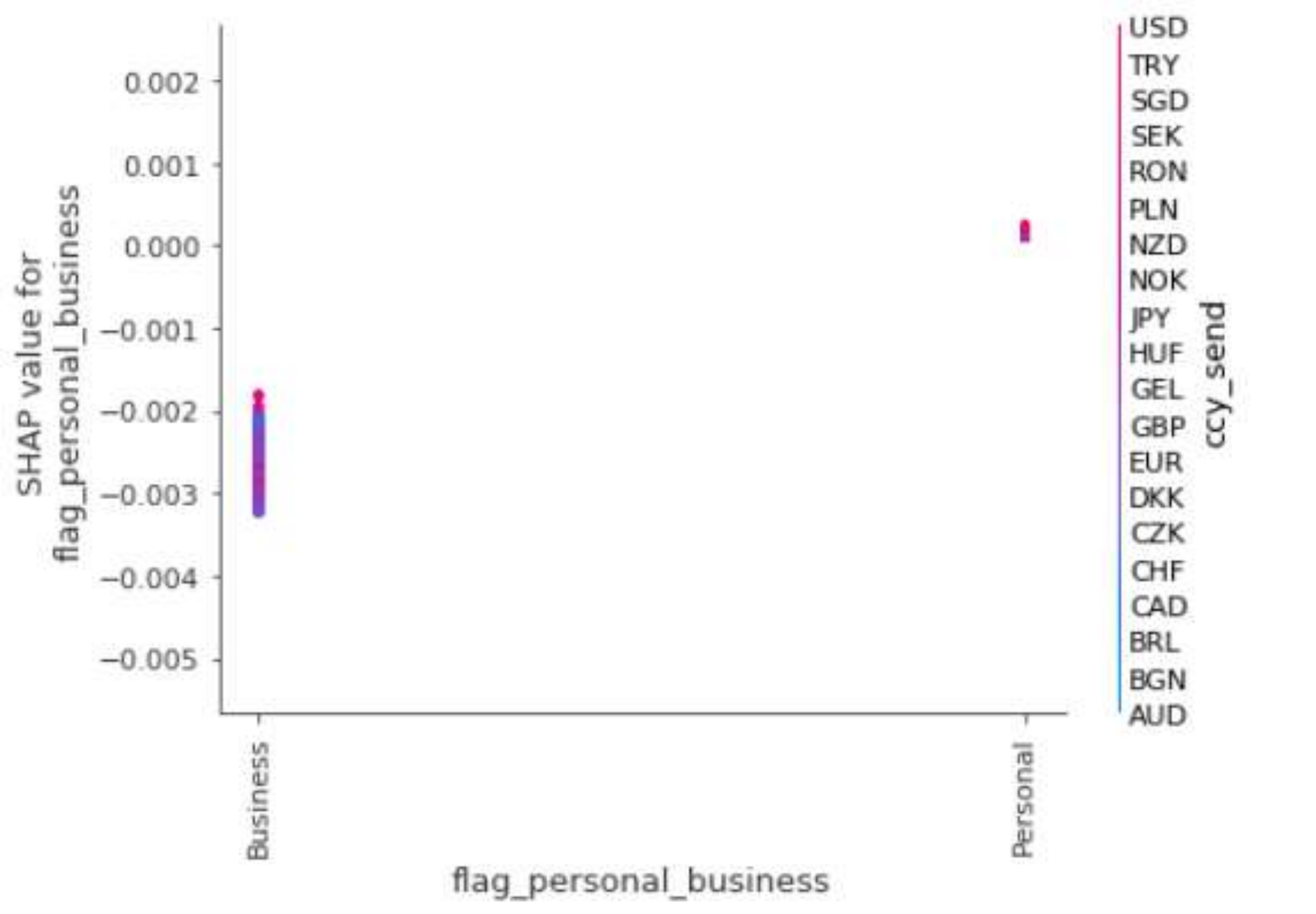
Anomaly Threshold
-0.0052

MODEL EXPLAINING



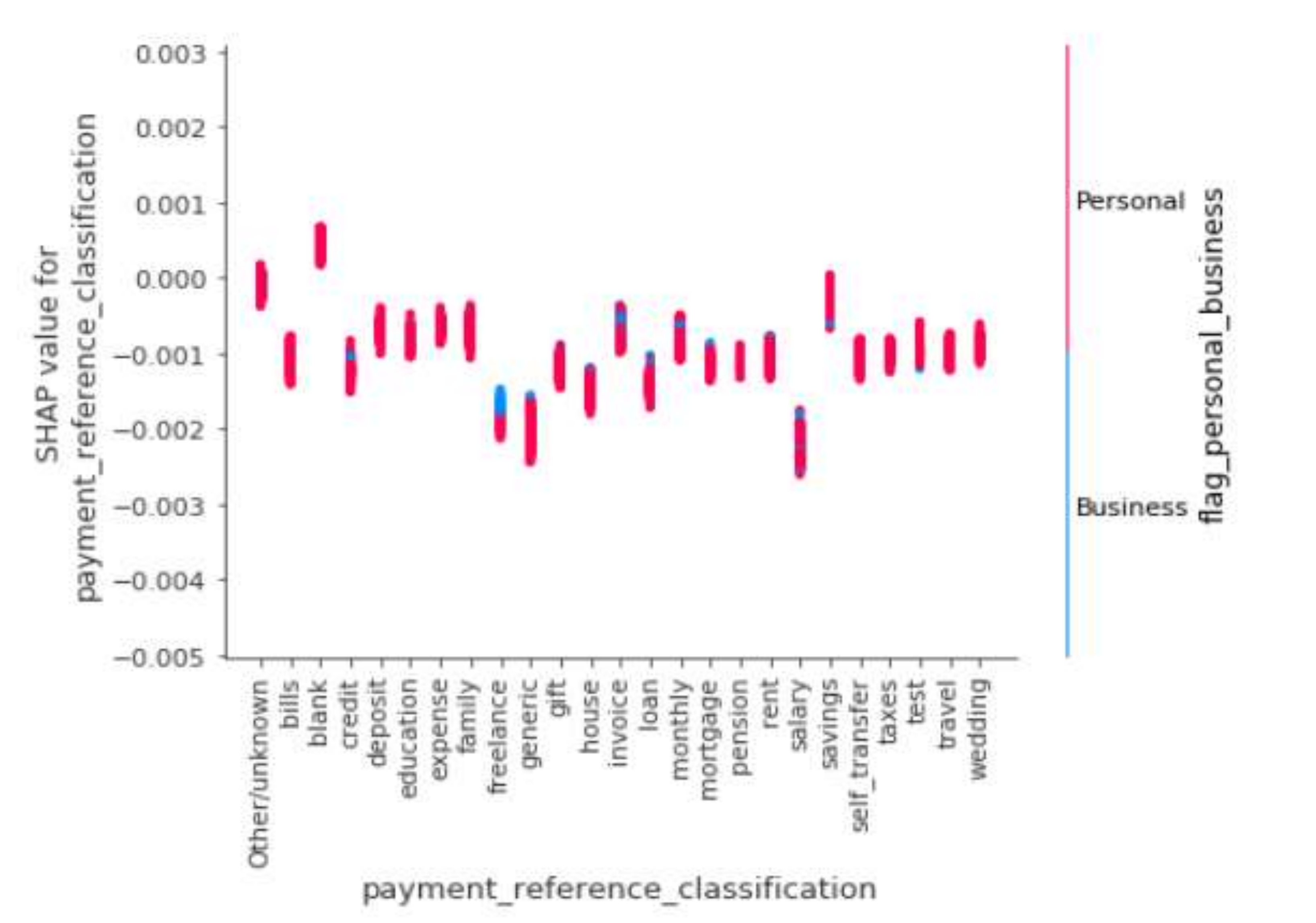
Anomaly Threshold
-0.0052

MODEL EXPLAINING



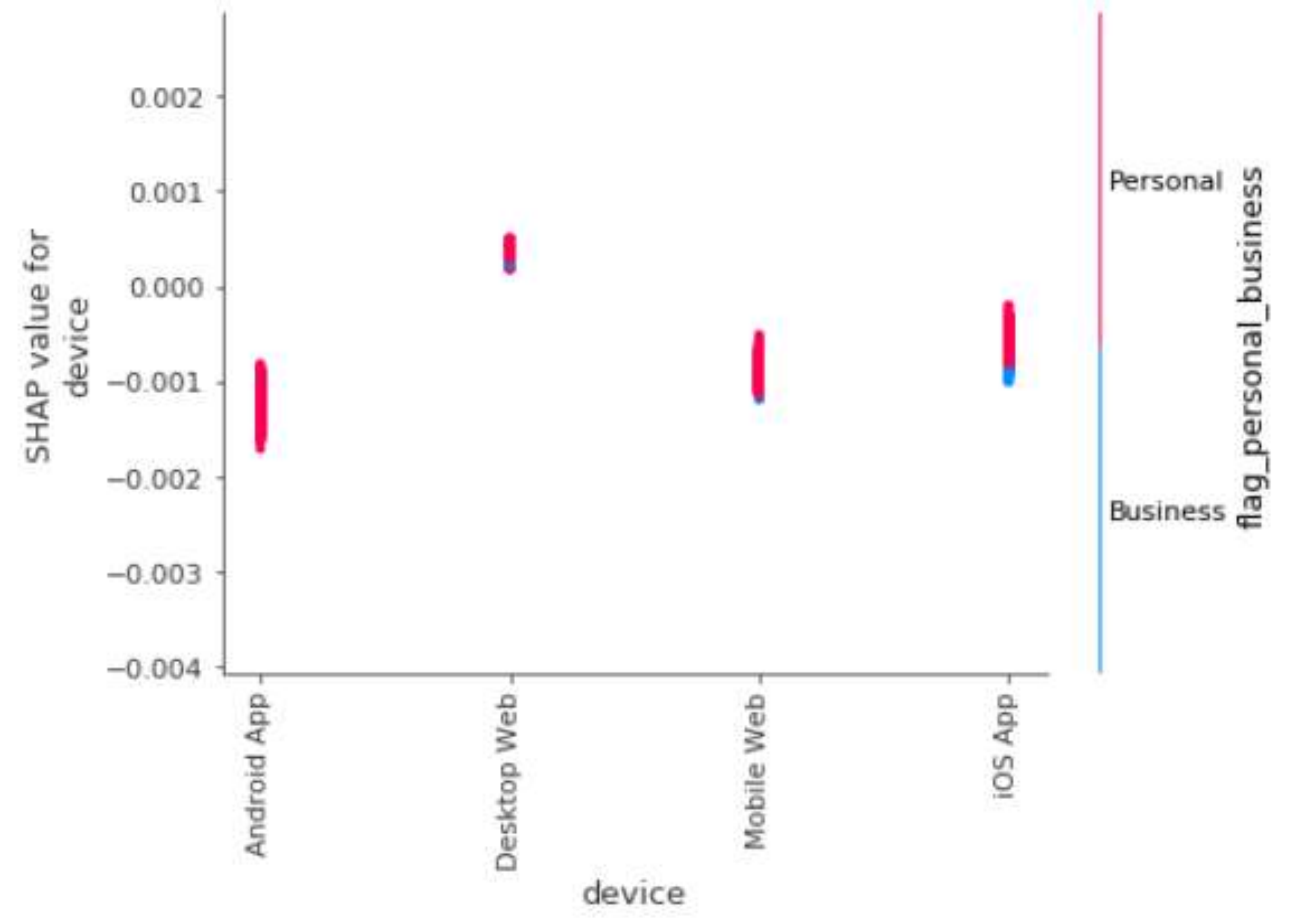
Anomaly Threshold
-0.0052

MODEL EXPLAINING

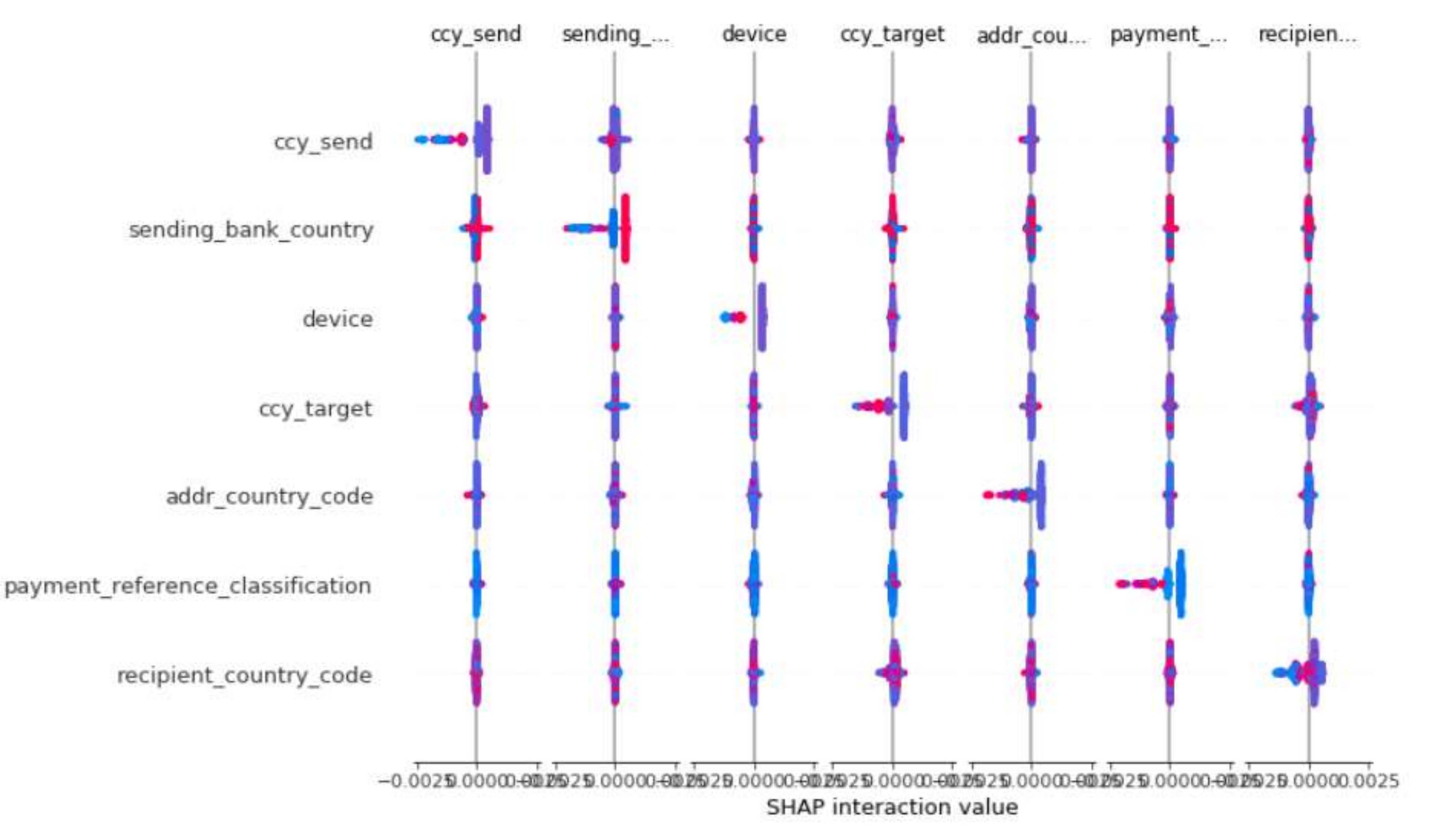


Anomaly Threshold
-0.0052

MODEL EXPLAINING



MODEL EXPLAINING



MODEL EXPLAINING