Fraud Detection

Ming Ki Toby Cheng, Xinyue (Anna) Jin, Ling Jiang
03/10/2020

Agenda

Project Overview

Why Fraud Detection

Data Schema

Tasks

Identify Challenges

Expected Insights

Project Overview

- Large amount of transactions, some of which are fraudulent
- Use features to predict fraud based on transaction data
- Focus on getting a strong predictive model while trying to capture as many fraudulent cases as possible

Why Fraud Detection?

- Perceive and block fraudulent proactively
- Protect clients from credit card fraud
- Prevent global payment companies from money laundering

Influences:

- Improve banks' reputation and customer loyalty
- Decrease risks and improve profitability

Data Schema

- The dataset contains approximately 6,353,307 of financial records divided into the 5 types of categories: Cash-in, Cash-out, Debit, Payment and Transfer.

- Kaggle link: https://www.kaggle.com/ntnu-testimon/paysim1

Variable	Description
step	Maps a unit of time in real world (1 step is 1 hour of time)
type	Types of Transactions
amount	Amount of the transaction in local currency
nameOrig	Customer who started the transaction
oldbalance	Orginal balance before the transaction
newbalance	Original customer's balance after the transaction.
nameDest	Recipient ID of the transaction.
oldbalanceDest	Initial recipient balance before the transaction
newbalanceDest	Recipient's balance after the transaction
isFraud	Identifies a fraudulent transaction (1) and non fraudulent (0)
isFlaggedFraud	Flags illegal attempts to transfer more than 200.000 in a single transaction.

Initial Dataset

	step	type	amount	nameOrig	oldbalanceOrg	newbalanceOrig	nameDest	oldbalanceDest	newbalanceDest	isFraud	isFlaggedFraud
0	1	PAYMENT	9839.64	C1231006815	170136.00	160296.36	M1979787155	0.0	0.00	0	0
1	1	PAYMENT	1864.28	C1666544295	21249.00	19384.72	M2044282225	0.0	0.00	0	0
2	1	TRANSFER	181.00	C1305486145	181.00	0.00	C553264065	0.0	0.00	1	0
3	1	CASH_OUT	181.00	C840083671	181.00	0.00	C38997010	21182.0	0.00	1	0
4	1	PAYMENT	11668.14	C2048537720	41554.00	29885.86	M1230701703	0.0	0.00	0	0
5	1	PAYMENT	7817.71	C90045638	53860.00	46042.29	M573487274	0.0	0.00	0	0
6	1	PAYMENT	7107.77	C154988899	183195.00	176087.23	M408069119	0.0	0.00	0	0
7	1	PAYMENT	7861.64	C1912850431	176087.23	168225.59	M633326333	0.0	0.00	0	0
8	1	PAYMENT	4024.36	C1265012928	2671.00	0.00	M1176932104	0.0	0.00	0	0
9	1	DEBIT	5337.77	C712410124	41720.00	36382.23	C195600860	41898.0	40348.79	0	0

Tasks

Classification

Build classification models to predict whether a transaction shows fraudulent.

OR

Clustering

Cluster transactions into various aggregations by characteristics.

Identify Challenges

- Highly imbalanced data
- Feature selection
- Model selection

```
+-------+-------+
|isFraud|  count|
+-------+-------+
|      1|   8213|
|      0|6354407|
+-------+-------+
```

```
+--------+-------+
|    type|  count|
+--------+-------+
|TRANSFER| 532909|
| CASH_IN|1399284|
|CASH_OUT|2237500|
| PAYMENT|2151495|
|   DEBIT|  41432|
+--------+-------+
```

```
data.describe('amount').show()
```

```
+-------+-----------------+
|summary|           amount|
+-------+-----------------+
|  count|          6362620|
|   mean|179861.90354913412|
| stddev| 603858.2314629498|
|    min|              0.0|
|    max|    9.244551664E7|
+-------+-----------------+
```

```
+-------+-----------------+
|summary|             step|
+-------+-----------------+
|  count|          6362620|
|   mean|243.39724563151657|
| stddev|142.33197104912588|
|    min|                1|
|    max|              743|
+-------+-----------------+
```

Expected Insights

- Identify the key features of fraudulent transactions

- Use the model on future data to predict fraudulent activity

Thank you!