CSE 4062 – Introduction to Data Science and Analytics
Spring 2021 - Delivery #1

Project Report - Group #8

Group Members:
Caner Dağdaş | 150716001 | canerdagdas@hotmail.com
Ceyhun Vardar | 150317022 | vardarceyhun13@gmail.com
Büşra Gökmen | 150116027 | busragokmen67@gmail.com
Cem Güleç | 150117828 | cem.ggulecc@gmail.com
Ömer Faruk Çakı | 150117821 | omerfarukcaki@gmail.com

Project Title: Fraud Detection on Financial Data

Lecturer: Assoc. Prof. Murat Can Ganiz

16.04.21

Project Explanation

The financial services industry and the industries that involve financial transactions are suffering from fraud-related losses and damages. The number of fraudulent customers has reached a high level in recent years. The reason for this is the money stolen from banks. The shift to the digital space opens new channels for financial service delivery. It also created a rich environment for scammers. As a consequence of this, the need for automatic systems which are able to detect and fight fraudsters has emerged.

Fraud detection is notably a challenging problem because;
Fraud strategies change in time, as well as customers' spending habits evolve.
Few examples of frauds available, so it is hard to create a model of fraudulent behavior.
Not all frauds are reported or reported with a large delay.
Few transactions can be timely investigated.

If earlier criminals had to counterfeit client IDs, now getting a person's account password may be all that's needed to steal money. With fraudsters becoming more adept at finding and exploiting loopholes in systems, fraud management has turned painful for the banking and finance industry. Customer loyalty and conversions are affected by fraudsters.

In order to maintain customer loyalty and conversions, financial services firm's need to detect fraud correctly and rapidly. Machine Learning and Deep Learning systems can detect changing strategies of fraudness quickly and correctly as needed.

Why we use Machine Learning and Deep Learning to detect fraud has 4 main reasons:

- Scalable
- Faster
- Efficient
- More accurate

There are 11 attributes in our data with approximately 6.3 million instances. Attributes are ;step, type, amount, nameOrig, oldbalanceOrg, newbalanceOrig, nameDest, oldbalanceDest, newbalanceDest, isFraud, isFlaggedFraud.

Our main purpose is to detect fraud activities in transactions between accounts.

Our project will take the financial sector one step further by identifying fraud, which is the bleeding wound of the financial sector, through machine learning by modeling in a way to detect fraud using the data set we have obtained.

Data Statistics

There are 11 columns and 6.362.622 rows in our dataset.

Explanation of Dataset Attributes:

step	type	amount	nameOrig	oldbalanceOrg	newbalanceOrig	nameDest	oldbalanceDest	newbalanceDest	isFraud	isFlaggedFraud
1	PAYMENT	9839.64	C1231006815	170136.0	160296.36	M1979787155	0.0	0.0	0	0
1	PAYMENT	1864.28	C1666544295	21249.0	19384.72	M2044282225	0.0	0.0	0	0
1	TRANSFER	181.0	C1305486145	181.0	0.0	C553264065	0.0	0.0	1	0
1	CASH_OUT	181.0	C840083671	181.0	0.0	C38997010	21182.0	0.0	1	0
1	PAYMENT	11668.14	C2048537720	41554.0	29885.86	M1230701703	0.0	0.0	0	0
1	PAYMENT	7817.71	C90045638	53860.0	46042.29	M573487274	0.0	0.0	0	0
1	PAYMENT	7107.77	C154988899	183195.0	176087.23	M408069119	0.0	0.0	0	0
1	PAYMENT	7861.64	C1912850431	176087.23	168225.59	M633326333	0.0	0.0	0	0
1	PAYMENT	4024.36	C1265012928	2671.0	0.0	M1176932104	0.0	0.0	0	0
1	DEBIT	5337.77	C712410124	41720.0	36382.23	C195600860	41898.0	40348.79	0	0
1	DEBIT	9644.94	C1900366749	4465.0	0.0	C997608398	10845.0	157982.12	0	0
1	PAYMENT	3099.97	C249177573	20771.0	17671.03	M2096539129	0.0	0.0	0	0
1	PAYMENT	2560.74	C1648232591	5070.0	2509.26	M972865270	0.0	0.0	0	0
1	PAYMENT	11633.76	C1716932897	10127.0	0.0	M801569151	0.0	0.0	0	0
1	PAYMENT	4098.78	C1026483832	503264.0	499165.22	M1635378213	0.0	0.0	0	0
1	CASH_OUT	229133.94	C905080434	15325.0	0.0	C476402209	5083.0	51513.44	0	0
1	PAYMENT	1563.82	C761750706	450.0	0.0	M1731217984	0.0	0.0	0	0
1	PAYMENT	1157.86	C1237762639	21156.0	19998.14	M1877062907	0.0	0.0	0	0
1	PAYMENT	671.64	C2033524545	15123.0	14451.36	M473053293	0.0	0.0	0	0
1	TRANSFER	215310.3	C1670993182	705.0	0.0	C1100439041	22425.0	0.0	0	0
1	PAYMENT	1373.43	C20804602	13854.0	12480.57	M1344519051	0.0	0.0	0	0
1	DEBIT	9302.79	C1566511282	11299.0	1996.21	C1973538135	29832.0	16896.7	0	0
1	DEBIT	1065.41	C1959239586	1817.0	751.59	C515132998	10330.0	0.0	0	0
1	PAYMENT	3876.41	C504336483	67852.0	63975.59	M1404932042	0.0	0.0	0	0
1	TRANSFER	311685.89	C1984094095	10835.0	0.0	C932583850	6267.0	2719172.89	0	0
1	PAYMENT	6061.13	C1043358826	443.0	0.0	M1558079303	0.0	0.0	0	0
1	PAYMENT	9478.39	C1671590089	116494.0	107015.61	M58488213	0.0	0.0	0	0
1	PAYMENT	8009.09	C1053967012	10968.0	2958.91	M295304806	0.0	0.0	0	0
1	PAYMENT	8901.99	C1632497828	2958.91	0.0	M33419717	0.0	0.0	0	0
1	PAYMENT	9920.52	C764826684	0.0	0.0	M1940055334	0.0	0.0	0	0
1	PAYMENT	3448.92	C2103763750	0.0	0.0	M335107734	0.0	0.0	0	0

step (numeric): maps a unit of time in the real world. In this case 1 step is 1 hour of time. Total steps 744 (30 days simulation).

type (text/nominal ?? bundan emin değilim): CASH-IN, CASH-OUT, DEBIT, PAYMENT and TRANSFER.

amount (numeric): amount of the transaction in local currency.

nameOrig (nominal): customer who started the transaction

oldbalanceOrg (numeric): initial balance before the transaction

newbalanceOrig (numeric): new balance after the transaction

nameDest (nominal): customer who is the recipient of the transaction

oldbalanceDest (numeric): initial balance recipient before the transaction. Note that there is not information for customers that start with M (Merchants).

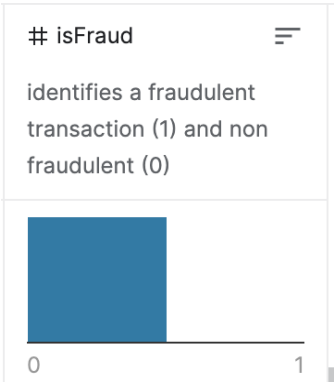
newbalanceDest (numeric): new balance recipient after the transaction. Note that there is not information for customers that start with M (Merchants).

isFraud (boolean): This is the transactions made by the fraudulent agents inside the simulation. In this specific dataset the fraudulent behavior of the agents aims to profit by taking control or customers accounts and trying to empty the funds by transferring to another account and then cashing out of the system.

isFlaggedFraud (boolean): The business model aims to control massive transfers from one account to another and flags illegal attempts. An illegal attempt in this dataset is an attempt to transfer more than 200.000 in a single transaction.

Target Attributes

- Our target attribute is isFraud column data.



0, represents non fraudulent transaction
1, represents fraudulent transaction

- We will use step, type, amount, nameOrig, oldbalanceOrg, newbalanceOrig, nameDest, oldbalanceDest, newbalanceDest, isFlaggedFraud parameters and target attribute(isFraud) for classification modeling.
- After having these attributes preprocessed and ready to be used in order to classify an instance - which will the future input to be predicted, we are planning to try several different classification algorithms such as Decision Trees (Gain Ratio and Gini index), Naive Bayes, Neural Networks (will be experimented with different number of hidden layers) and Support Vector Machines (SVM). These algorithms may require different types of hyperparameters and the experimentation phase also will include hypertuning these parameters and see which one is the best fitting one for our goal.