Fraud Detection

11 December 2024
Author: Darell Hendry (1darellhendry@gmail.com)

Background

Researchers from the IEEE Computational Intelligence Society are collaborating with Vesta Corporation to work together in fraud detection through advanced machine learning techniques. They intend to enhance the accuracy of fraud prevention systems, with reducing financial losses for businesses and minimizing embarrassing moments for consumers.
Reducing the false negatives.

Methodology

To predict whether a transaction is fraud or not.

Feature selection

Identify most relevant features that indicate fraud. Remove noise and irrelevant columns. Reduce computational complexity. Improve model performance
Using hypothesis testing: **Kolmogorov-Smirnov (numerical) & Chi-square (non-numerical)**. Kolmogorov-Smirnov has no assumption on the data.

Data Cleaning

- Dropped NaN values. Remove incomplete or missing data rows/columns. Ensures data quality and model reliability
- Filled missing values. Use statistical methods mean,median. Prevent information loss. Maintain dataset integrity

Model

Statistical Analysis

Naive Bayes Classification
Probabilistic machine learning algorithm. Calculates fraud probability based on feature characteristics. Assumes independence between features. **Lightweight and fast classification method.** Provides baseline fraud detection model. Works well with limited training data

Machine Learning Approaches

LightGBM (LGBM).
Advanced gradient boosting framework widely used in industry. Handles imbalanced fraud detection datasets

Key advantages:

- Faster training speed
- Lower memory usage
- Handles categorical features
- High prediction accuracy -> robust with imbalance data

K-means Clustering
Discovers hidden patterns in transaction data. Groups similar transactions. Identifies anomaly clusters

Model Evaluation

Performance Metric: Precision
Calculation: TP / TP+FP
Measures accuracy of positive fraud predictions. Focuses on minimizing false positives. Critical for financial fraud detection because the story of this study case tell how their customer failed to be classified as fraudster.

Result

Summary table

Model	Accuracy	**Precision**	Recall	F1-Score
Naive Bayes	0.1498	**0.26**	0.9906	0.1425
LGBM	0.9690	**0.9074**	0.6196	0.7363
K-means (clustering)	0.9080	**0.2795**	0.2002	0.2332

Modeling

Naive Bayes

Fold	Accuracy	Precision	Recall	F1-Score
1	0.1671	0.0773	0.9922	0.1434
2	0.1717	0.0765	0.9870	0.1421
3	0.1692	0.0784	0.9888	0.1453
4	0.1656	0.0762	0.9933	0.1415
5	0.0755	0.9920	0.9920	0.1404
Avg	0.1498	**0.26**	0.9906	0.1425

LGBM

Fold	Accuracy	Precision	Recall	F1-Score
1	0.9692	0.9126	0.6210	0.7391
2	0.9692	0.9141	0.6141	0.7346
3	0.9687	0.9078	0.6249	0.7402
4	0.9696	0.9124	0.6200	0.7383
5	0.9684	0.8904	0.6182	0.7297
Avg	0.9690	**0.9074**	0.6196	0.7363

K-means outlier detection

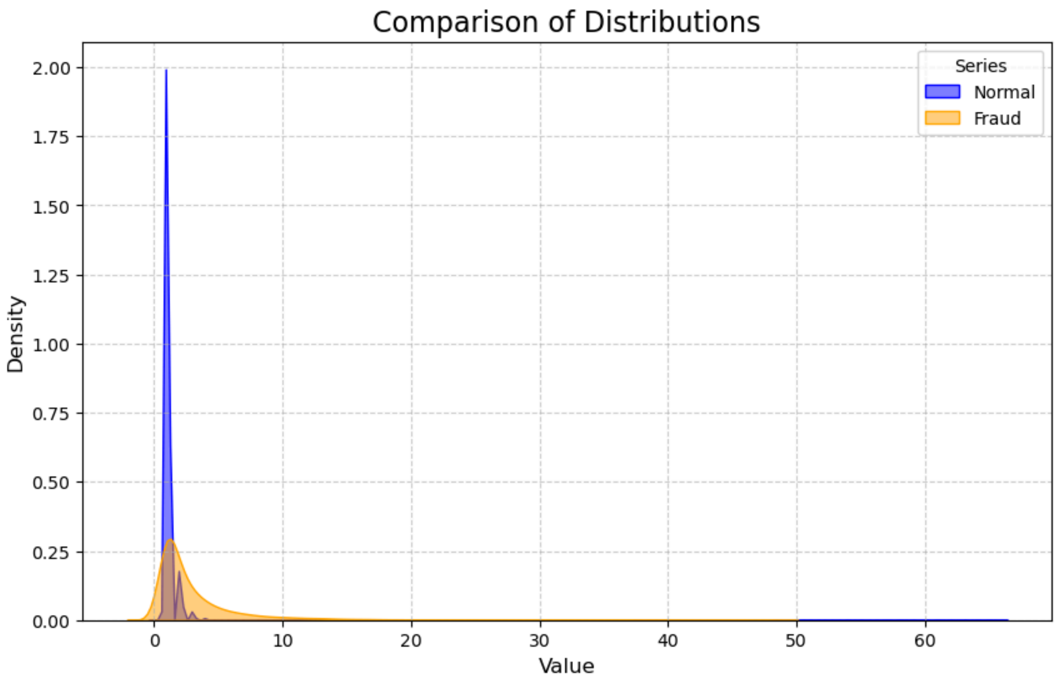
Fold	Accuracy	Precision	Recall	F1-Score
1	0.9075	0.2779	0.1979	0.2312
2	0.9087	0.2819	0.2028	0.2359
3	0.9051	0.2648	0.1855	0.2181
4	0.9074	0.2670	0.1928	0.2239
5	0.9117	0.3060	0.2220	0.2573
Avg	0.9080	**0.2795**	0.2002	0.2332

Exploratory Data Analysis

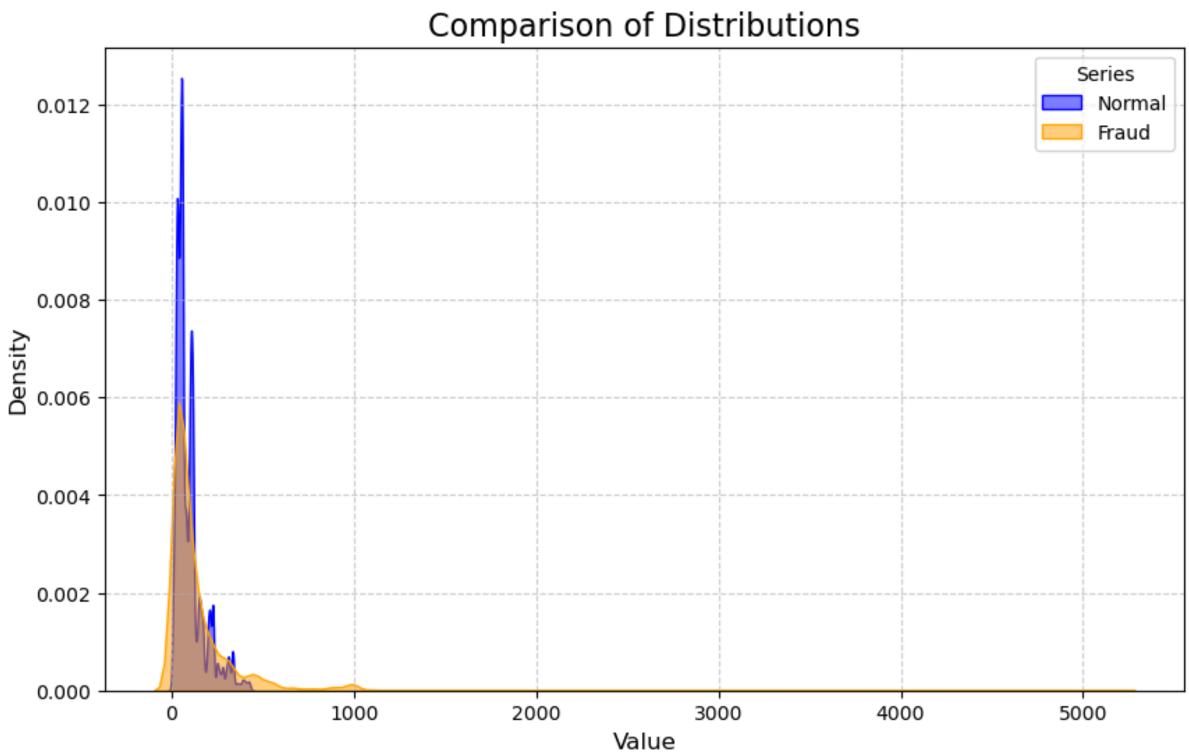
List of statistically has different distribution between normal and fraud transaction :

['V25', 'V223', 'addr2', 'V26', 'V311', 'V286', 'V174', 'V173', 'V206', 'D13', 'V111', 'V108', 'V113', 'V110', 'V98', 'V112', 'V8', 'V109', 'V106', 'V7', 'V205', 'V299', 'V114', 'V285', 'V99', 'V19', 'V334', 'V325', 'V175', 'V6', 'V105', 'V207', 'V138', 'V116', 'V266', 'V300', 'addr1', 'V159', 'V169', 'V66', 'V4', 'V220', 'V284', 'V297', 'V319', 'V20', 'V226', 'V130', 'V2', 'V298', 'V296', 'V104', 'V135', 'V208', 'V209', 'V225', 'V301', 'V327', 'V115', 'V288', 'V210', 'V82', 'V326', 'V160', 'V100', 'V227', 'V310', 'V124', 'V61', 'V137', 'V321', 'V136', 'V336', 'V335', 'V330', 'dist2', 'V129', 'V3', 'V320', 'card2', 'V96', 'V268', 'V328', 'V9', 'V337', 'V250', 'V287', 'V255', 'V131', 'V197', 'V141', 'V161', 'V339', 'V224', 'V46', 'V329', 'V290', 'V163', 'V142', 'V289', 'V309', 'V123', 'V314', 'V196', 'V270', 'V267', 'V194', 'V162', 'V198', 'V272', 'V313', 'V271', 'V127', 'V256', 'TransactionAmt', 'V338', 'V67', 'V5', 'V251', 'D6', 'V252', 'V315', 'V191', 'V125', 'V23', 'V195', 'V83', 'V291', 'V312', 'D9', 'card5', 'V281', 'V55', 'V181', 'D11', 'V193', 'V62', 'V307', 'V247', 'V260', 'V56', 'V24', 'V184', 'V254', 'V238', 'V214', 'V151', 'V143', 'V185', 'V164', 'V192', 'V292', 'V183', 'V322', 'dist1', 'D12', 'V235', 'V216', 'V54', 'V215', 'V47', 'V239', 'V182', 'V180', 'V331', 'V293', 'V53', 'V253', 'V101', 'V249', 'D14', 'C6', 'V13', 'V75', 'V76', 'D4', 'V12', 'V152', 'V97', 'V144', 'V323', 'V276', 'V166', 'D1', 'V332', 'V324', 'V316', 'V237', 'V279', 'V165', 'V333', 'C13', 'V128', 'V278', 'V132', 'C14', 'V248', 'V234', 'V280', 'V308', 'V306', 'V103', 'V153', 'V95', 'V78', 'V77', 'V236', 'V282', 'V261', 'V126', 'V262', 'V36', 'V35', 'V37', 'V283', 'V145', 'V277', 'V134', 'C11', 'V102', 'D15', 'V133', 'V154', 'V295', 'C1', 'V318', 'V87', 'D10', 'D7', 'C2', 'V86', 'V294', 'V317', 'V148', 'C5', 'V155', 'V245', 'V38', 'V11', 'V45', 'card3', 'V10', 'C9', 'V44', 'V146', 'V63', 'V70', 'V30', 'V150', 'D8', 'V59', 'C7', 'V69', 'V71', 'V29', 'V72', 'V64', 'V149', 'V91', 'V60', 'V221', 'V188', 'V49', 'V57', 'V259', 'V90', 'V58', 'V156', 'V48', 'D2', 'V186', 'V147', 'V21', 'V22', 'V302', 'V17', 'V18', 'V31', 'V32', 'V222', 'V304', 'C12', 'V73', 'V84', 'V15', 'V16', 'V80', 'V303', 'V176', 'V189', 'V85', 'V157', 'C10', 'V81', 'C8', 'V92', 'V74', 'D3', 'V93', 'V158', 'V33', 'V139', 'V177', 'V79', 'V34', 'V190', 'V211', 'V42', 'C4', 'V167', 'V202', 'V94', 'V39', 'V43', 'V170', 'V50', 'V203', 'V40', 'V200', 'V171', 'V187', 'V178', 'V168', 'V242', 'V179', 'V213', 'V204', 'V244', 'V212', 'V140', 'V228', 'V199', 'V51', 'V201', 'V52', 'V231', 'V230', 'V246', 'V273', 'D5', 'V217', 'V229', 'V243', 'V233', 'V263', 'V275', 'V218', 'V257', 'V232', 'V264', 'V219', 'V274', 'V265', 'V258']

- V258 (Fraud group tend to has higher value)



- TransactionAmt (Fraud group has higher amount than Normal group)



List of cat features with different distributions between normal transaction and fraud:
['M7', 'M8', 'M6', 'M5', 'M9', 'card4', 'M2', 'M3', 'P_emaildomain', 'R_emaildomain', 'card6', 'M4', 'ProductCD']

	test_statistic	p_value	distributions_are_different	feature
11	11.256096	7.936196e-04	True	M7
17	55.145173	1.060176e-12	True	id_23
12	88.530215	5.006429e-21	True	M8
27	95.847507	1.240822e-22	True	id_37
14	119.092679	9.994885e-28	True	id_12
24	187.327111	2.306615e-40	True	id_34
26	194.553071	3.225181e-44	True	id_36
10	227.964137	1.657062e-51	True	M6
9	242.421692	1.165949e-54	True	M5
13	250.372504	2.153926e-56	True	M9
1	364.874139	8.969834e-79	True	card4
6	438.613213	2.169030e-97	True	M2
7	477.660569	6.899239e-106	True	M3
29	609.623764	1.350769e-134	True	DeviceType
28	701.740456	1.250925e-154	True	id_38
19	1216.975030	1.247881e-266	True	id_28
20	1351.992155	5.685467e-296	True	id_29
15	1399.954759	1.008525e-304	True	id_15
16	1553.785677	0.000000e+00	True	id_16
21	1937.280664	0.000000e+00	True	id_30
25	2888.471367	0.000000e+00	True	id_35
3	3497.812835	0.000000e+00	True	P_emaildomain
4	3670.552069	0.000000e+00	True	R_emaildomain
23	3959.447237	0.000000e+00	True	id_33
22	4583.114771	0.000000e+00	True	id_31
2	5957.032292	0.000000e+00	True	card6
8	6450.447980	0.000000e+00	True	M4
0	16742.171529	0.000000e+00	True	ProductCD
30	19231.894630	0.000000e+00	True	DeviceInfo