Supervised model on Corporate Payments data

JIAYING GU, JESSIE YU, SIYU ZHANG, RUOYU SUN, ISABELLE ZHAO

Overview

Corporate payment data with label – fraud/nonfraud (4000 fraud)

- Goal: Detect as many fraud as possible!

Data Description

Entity Selection

Variable Construction

Methodology & Model Results

- Training, Testing, Validation
- Linear/Nonlinear Models

Our best model based on FDR @3%: KNN

Summary of Data

95271 records

8 fields

Fraud rate: 4.2%

- Dependent variables

Name	% Populated
Fraud label	100%

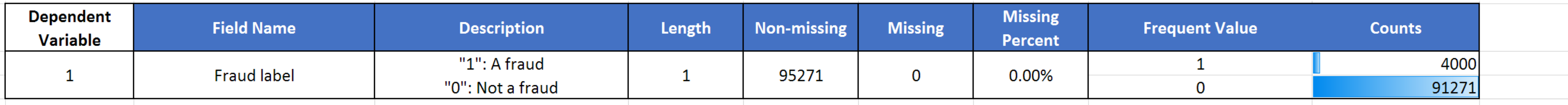
- Independent variables

Numerical Fields

Field Name	% Populated
AMOUNT	100%

Categorical Fields

Field Name	% Populated
CARDNUM	100%
MERCHNUM	96%
MERCHDESCRIPTION	100%
MERCHSTATE	99%
TRANSTYPE	100%
MERCHZIP	95%
DATE	100%



Dependent Variable	Field Name	Description	Length	Non-missing	Missing	Missing Percent	Frequent Value	Counts
1	Fraud label	"1": A fraud "0": Not a fraud	1	95271	0	0.00%	1	4000
							0	91271

Entities and Variables

Entities: CARDNUM, MERCHNUM, STATE

25 Variables

- $card\_scale\_trans\_N = (90/N) \cdot \frac{Number\ of\ transactions\ in\ the\ past\ N\ days\ on\ this\ card}{Number\ of\ transactions\ in\ the\ past\ 90\ days\ on\ this\ card}$

For N = 1, 2, 3, 7

- $card\_scale\_amount\_N = (90/N) \cdot \frac{Total\ transaction\ amount\ in\ the\ past\ N\ days\ on\ this\ card}{Total\ transaction\ amount\ in\ the\ past\ 90\ days\ on\ this\ card}$

For N = 1, 2, 3, 7

- $merch\_scale\_trans\_N = (90/N) \cdot \frac{Number\ of\ transactions\ in\ the\ past\ N\ days\ from\ merchant}{Number\ of\ transactions\ in\ the\ past\ 90\ days\ from\ merchant}$

For N = 1, 2, 3, 7

- $merch\_scale\_amount\_N = (90/N) \cdot \frac{Total\ trans\ amount\ in\ the\ past\ N\ days\ from\ merchant}{Total\ trans\ amount\ in\ the\ past\ 90\ days\ from\ merchant}$

For N = 1, 2, 3, 7

- $card\_scale\_dup\_N = 100 \cdot \frac{Number\ of\ trans\ in\ the\ past\ N\ days\ on\ this\ card\ with\ same\ amount}{Number\ of\ transactions\ in\ the\ past\ N\ days\ on\ this\ card}$

For N = 1, 2, 3, 7

- $card\_scale\_dup\_N = 100 \cdot \frac{Number\ of\ trans\ in\ the\ past\ N\ days\ from\ merch\ with\ same\ amount}{Number\ of\ transactions\ in\ the\ past\ N\ days\ from\ merchant}$

For N = 1, 2, 3, 7

- $card\_scale\_State\_N = 100 \cdot \frac{Number\ of\ trans\ in\ the\ past\ 1\ day\ on\ this\ card\ with\ same\ state}{Number\ of\ transactions\ in\ the\ past\ 1\ day\ on\ this\ card}$

For N = 1

Methodology

Linear models:

- Logistic regression
- LDA
- QDA

Non-linear models:

- Random forest
- SVM
- Neural network
- CART
- Boosted tree
- KNN

Training sets:

- 80:20 with 25 variables
- 80:20 with 16 variables
- Downsize 1:1 with 25 variables
- Downsize 3:1 with 25 variables
- Downsize 5:1 with 25 variables
- Downsize 7:1 with 25 variables
- Downsize 10:1 with 25 variables

Testing:

- 20% random selection

OOT Validation:

- Records starting from 9/1

Model Results - Linear

FDR@3%	Train	Test	Validate
Logistics	16.94%	19.15%	30.21%
LDA	3.90%	12.98%	29.58%
QDA	10.06%	25.11%	29.39%

Best linear model: Logistic

Low training score:

- LDA, QDA best model used downsized training data

Unresolved issue:

- High validation, low test & train

Model Results – Non-Linear

FDR@3%	Train	Test	Validate
KNN	29.41%	29.79%	29.01%
Neural Net	37.18%	36.81%	27.11%
Random Forest	65.09%	34.26%	26.36%
Boosted Tree	22.95%	36.60%	26.17%
CART - tree	14.90%	24.90%	24.70%
CART - rpart	14.90%	26.00%	21.90%
SVM	15.86%	32.98%	19.09%

- Best nonlinear model: KNN
- KNN performed best in validation, consistent and stable results for train, test, and validate.
- Neural Net, Random Forest, Boosted tree all performed relatively well

Model Results – KNN Training

Percentile	# of Records	# Goods	# bads	Cumulative Goods	Cumulative Bads	% Bad	% Good	Cumulative fraud detection rate	Bin KS
1.00%	547	281	266	281	266	48.63%	51.37%	13.66%	0.13
2.00%	546	351	195	632	461	35.71%	64.29%	23.67%	0.22
3.00%	546	434	112	1066	573	20.51%	79.49%	29.41%	0.27
4.00%	546	456	90	1522	663	16.48%	83.52%	34.03%	0.31
5.00%	546	442	104	1964	767	19.05%	80.95%	39.37%	0.36
6.00%	546	469	77	2433	844	14.10%	85.90%	43.33%	0.39
7.00%	546	466	80	2899	924	14.65%	85.35%	47.43%	0.42
8.00%	547	479	68	3378	992	12.43%	87.57%	50.92%	0.45
9.00%	546	495	51	3873	1043	9.34%	90.66%	53.54%	0.46
10.00%	546	508	38	4381	1081	6.96%	93.04%	55.49%	0.47
11.00%	546	516	30	4897	1111	5.49%	94.51%	57.03%	0.48
12.00%	546	506	40	5403	1151	7.33%	92.67%	59.09%	0.49
13.00%	546	498	48	5901	1199	8.79%	91.21%	61.55%	0.50
14.00%	546	514	32	6415	1231	5.86%	94.14%	63.19%	0.51
15.00%	546	501	45	6916	1276	8.24%	91.76%	65.50%	0.52
16.00%	546	507	39	7423	1315	7.14%	92.86%	67.51%	0.53

Model Results – KNN Testing

Percentile	# of Records	# Goods	# Bads	Cumulative Goods	Cumulative Bads	% Bad	% Good	Cumulative Fraud Detection	Bin KS
1.00%	137	69	68	69	68	49.64%	50.36%	14.47%	0.14
2.00%	136	100	36	169	104	26.47%	73.53%	22.13%	0.21
3.00%	137	101	36	270	140	26.28%	73.72%	29.79%	0.28
4.00%	136	116	20	386	160	14.71%	85.29%	34.04%	0.31
5.00%	137	116	21	502	181	15.33%	84.67%	38.51%	0.35
6.00%	136	123	13	625	194	9.56%	90.44%	41.28%	0.37
7.00%	137	115	22	740	216	16.06%	83.94%	45.96%	0.40
8.00%	136	121	15	861	231	11.03%	88.97%	49.15%	0.43
9.00%	137	126	11	987	242	8.03%	91.97%	51.49%	0.44
10.00%	137	131	6	1118	248	4.38%	95.62%	52.77%	0.44
11.00%	136	131	5	1249	253	3.68%	96.32%	53.83%	0.44
12.00%	137	124	13	1373	266	9.49%	90.51%	56.60%	0.46
13.00%	136	133	3	1506	269	2.21%	97.79%	57.23%	0.46
14.00%	137	125	12	1631	281	8.76%	91.24%	59.79%	0.47
15.00%	136	127	9	1758	290	6.62%	93.38%	61.70%	0.48
16.00%	137	124	13	1882	303	9.49%	90.51%	64.47%	0.50

Model Results – KNN Validation

Percentile	# of Records	# Goods	# Bads	Cumulative Goods	Cumulative Bads	% Bad	% Good	Cumulative Fraud Detection Rate	Bin KS
1.00%	270	64	206	64	206	76.30%	23.70%	13.02%	0.13
2.00%	270	120	150	184	356	55.56%	44.44%	22.50%	0.22
3.00%	270	167	103	351	459	38.15%	61.85%	29.01%	0.28
4.00%	270	221	49	572	508	18.15%	81.85%	32.11%	0.30
5.00%	270	208	62	780	570	22.96%	77.04%	36.03%	0.33
6.00%	270	223	47	1003	617	17.41%	82.59%	39.00%	0.35
7.00%	269	235	34	1238	651	12.64%	87.36%	41.15%	0.36
8.00%	270	236	34	1474	685	12.59%	87.41%	43.30%	0.37
9.00%	270	242	28	1716	713	10.37%	89.63%	45.07%	0.38
10.00%	270	247	23	1963	736	8.52%	91.48%	46.52%	0.39
11.00%	270	246	24	2209	760	8.89%	91.11%	48.04%	0.39
12.00%	270	234	36	2443	796	13.33%	86.67%	50.32%	0.41
13.00%	270	242	28	2685	824	10.37%	89.63%	52.09%	0.42
14.00%	270	234	36	2919	860	13.33%	86.67%	54.36%	0.43
15.00%	270	249	21	3168	881	7.78%	92.22%	55.69%	0.43
16.00%	270	244	26	3412	907	9.63%	90.37%	57.33%	0.44