Supervised Learning on Corporate Card Transaction Data

Jiaying Gu

Ruoyu Sun

Jiaxi Yu

Siyu Zhang

Xinyi Zhao

Executive Summary

Project 3, Supervised Learning on Corporate Card Transaction Data, is an exploration of 95,272 credit card transactions made during the year of 2010 from corporate card dataset. For each transaction record, there's detailed information of card number, transaction date, merchant number, merchant state and zip, etc. In addition, each record is given a fraud label indicating whether or not the record is categorized as a fraud. There is a total of 4000 fraudulent transaction records. Through supervised learning method, we'll build several different linear and nonlinear models to compute a fraud score for each record, and use the fraud label to calculate our fraud detection rate. The goal is to catch as many fraudulent transactions as possible.

We started the project by conducting data quality report on the dataset, briefly explored the dataset and found interesting and unusual things about the data. After understanding what the data is about, we began to build variables and prepared the variables for model construction. The entity levels we chose are card number (CARDNUM), merchant number (MERCHNUM) and State(MERCHSTATE). Since we want to study the transaction behavior, things we focus on are the number of transactions, the amount of transaction and duplicated transaction amounts. For the entity STATE, we calculated the percentage of transactions on a certain card that happened in the same state as the current record. Under these guidelines, we built 25 variables for modeling.

After building the variables, we first separated the out of time data, which contains all the transactions from 9/1 and onward. Then, we separated the rest of the data to training and testing data. We used 80% of the records as training, and 20% of the records as testing. For the training data, to get better modeling results and higher fraud detection rates, we also down-sampled the goods and selected only a fraction of the nonfraud records with all the fraud records. The ratios between nonfraud-to-fraud we used to construct more training datasets are 1/1, 3/1, 5/1, 7/1, and 10/1. The model we tried include Logistics, LDA, QDA, random forest, CART, boosted tree, SVM, neural net and KNN. We compared our models based on their FDR@3%, and finally KNN model came up with the best result. Following sections are detailed discussion of the data, the process of selecting entities and building variables, model algorithm and our model results

including the model performance and the score distribution tables.

Summary of Data

The data contains credit card transaction records, along with the card number, merchant information, and transaction date and type. There is a total of 95271 records (We excluded the one record related to Mexico with a suspiciously high transaction amount, the detail record is listed below) with 10 fields (1 unique identifier, 1 dependent variable, 8 independent variables). For each record, fraud label of "1" means that the record is fraud and "0" means that the record is nonfraud. In the data set, the percentage of records with "Fraud label" =1 approximately equals to 4.2%. The timeframe was from 1/1/2010 to 12/31/2010 and the original format is Excel.

Looking at the data, we first separated the fields into numerical and categorical variables. The basic standard we follow is to put most fields with continuous numerical values as a numerical variable, and put fields with word descriptions as a categorical variable. However, for variables like CARDNUM, MERCHNUM and MERCHZIP which seem like numerical, we think it makes more sense to put those fields as categorical. We can conduct our analysis by selecting categorical variables to account for the difference between each level. Below is a summary of the field names and the percent populated in each field.

- 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%

From our basic exploration of the data, a few findings may help guide further analysis:

- The number of transactions associated with each card number varies greatly, with largest number over 1,000. The number of transactions for each merchant also varies greatly, with largest number over 9,000. It might be interesting to explore the high values within these two entities.
- The zip code with each merchant has the highest number of missing values. The zip codes listed also have different length and formats. Due to the unexplainable irregularity in this field, we might not choose it as an entity for our analysis.
-

CARDNUM	DATE	MERCHNUM	MERCHDESCRIPTION	MERCH	MERCHZIP	TRANSTYPE	AMOUNT	fraud?
5142189135	7/13/2010		INTERMEXICO			P	$3,102,045.53	

  This is the largest amount of payment in the dataset, and has a significantly higher value than other records. There are many missing information in this row and the information of Merchant description which is “INTERMEXICO” is very suspicious associated with this payment amount. The existence of this record might have an influence on our scoring of other records.

Entities and Variables

Entities:

We divided the data mainly based on two entity levels: CARDNUM and MERCHNUM. Observing anomalies on these two entity levels may help account for the user differences among different card holders and different merchants. We also included STATE as an entity level, and observed the frequency of location changes for a given card on a given date.

Variables:

We added a total of 25 variables to model our data. Our intention is to find anomalies based on the number of transactions and the transaction amounts during a time frame.

For entities CARDNUM and MERCHNUM, we calculated the number of transactions, the total transaction amount, and the number of duplicates in amount value in a given time on each entity level. Due to the usual patterns of credit card fraud, we selected the time frame to be in the past 1, 2, 3, or 7 days. Since we are assuming that we have no knowledge of records that happened after each existing record, we standardized the number of transactions and total transaction amount by setting the activity on each entity level in the past 90 days as normal. For variables

that convey information about duplicate transaction amounts, we standardized the variables by setting the total number of transactions in the given time frame on a certain entity level as base, and divided it by the number of transactions that have the same amount as the record at hand in that time frame and entity level. We multiplied this fraction by 100 to get the percentage value of transactions that have duplicate values as the current record. The higher this number is, the more likely duplicate amounts occurred.

For the entity STATE, we calculated the percentage of transactions on a certain card that happened in the same state as the current record. We set the time frame to be in the past 1 day, since the change of location is time-sensitive. A longer time frame would be unnecessary, as it is possible and reasonable that travel occurred during a few days' time, and a change in state that happened in the past 1 day has a much higher probability of fraud. If this variable is 100, it means that the card has only been used in one state during the past 1 day. If the variable is small, it means that only a few records happened in a different location from most, and these are anomalies that we should put focus on.

Below are our 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

Model Algorithm

Model choosing

After creating the 25 variables as mentioned above, we ran 2 sets of models to see which performs better:

Logistic regression, LDA, QDA – Linear and simpler models. They should give us a baseline that all the nonlinear methods should improve over.

Random forest, SVM, neural network, CART, boosted tree, KNN – More sophisticated non-linear model. We expected these models to perform slightly better than logistic regression.

Data standardization

Experience as shown that neural network training is usually more efficient when numeric independent variables are scaled, or normalized, so that their magnitudes are relatively similar. Normalization also help SVM performs better in that all features have roughly the same magnitude (since we don't assume that some features are much more important than others). For this reason, we scaled the data we fed into the models. Noticed that we didn't apply data normalization to data to other models such as random forest because random forest is invariant to monotonic transformations of individual features so translations or per feature scaling will not change anything to the performance.

How to calculate FDA @3%

We calculate fraud detective rate for each model in order to know which one performs better. After we ran each model, we got a probability, which we used as a score, for each record.

	bt.preds	Record	label	card_scale_trans_1	card_scale_trans_2	card_scale_trans_3	card_scale_trans_7	merch_scale_trans_1	merch_scale_
1	0.009719414	68281	0	1.2857143	0.9642857	1.7142857	1.1020408	1.2872467	1
2	0.039239930	68282	1	2.6086957	4.5652174	3.4782609	1.8633540	45.0000000	22
3	0.056508043	68283	0	5.0000000	2.5000000	1.6666667	0.7142857	2.8125000	2
4	0.029403993	68284	0	2.6086957	4.5652174	3.4782609	1.8633540	10.0000000	5
5	0.019504135	68285	0	2.4324324	2.4324324	1.6216216	1.0424710	3.6460554	3
6	0.014748574	68286	0	3.2727273	2.4545455	2.1818182	0.9350649	4.5000000	5
7	0.103491530	68287	0	30.0000000	15.0000000	10.0000000	4.2857143	4.5000000	5
8	0.049018538	68288	0	10.9090909	6.8181818	4.5454545	3.1168831	45.0000000	22
9	0.011458561	68289	0	4.3902439	3.2926829	2.1951220	2.1951220	2.4545455	1
10	0.223139324	68290	0	11.2500000	5.6250000	3.7500000	1.6071429	0.9677419	1
11	0.032436955	68291	0	3.7500000	3.7500000	5.0000000	2.6785714	22.5000000	11
12	0.018188279	68292	0	1.8750000	0.9375000	0.6250000	1.3392857	3.4615385	1
13	0.019076443	68293	0	1.9459459	1.7027027	1.4594595	0.9729730	0.9677419	1
14	0.332456449	68294	0	22.5000000	11.2500000	7.5000000	3.2142857	90.0000000	45
15	0.050963259	68295	0	6.9230769	5.1923077	3.4615385	1.4835165	45.0000000	22
16	0.115087106	68296	1	90.0000000	45.0000000	30.0000000	12.8571429	90.0000000	45
17	0.036008329	68297	0	2.9032258	1.4516129	2.9032258	1.6589862	22.5000000	11
18	0.038591266	68298	0	5.0000000	5.0000000	3.3333333	1.4285714	30.0000000	30
19	0.088630467	68299	0	25.7142857	19.2857143	12.8571429	5.5102041	0.9407666	0
20	0.013385477	68300	0	1.5000000	3.0000000	4.5000000	2.1428571	8.1818182	4

We sorted the records by probability from high to low and chose top 3%.

	bt.preds	Record	label	card_scale_trans_1	card_scale_trans_2	card_scale_trans_3	card_scale_trans_7	merch_scale_trans_1
19288	0.9814075	87568	1	22.5000000	11.2500000	7.5000000	3.2142857	2.8125000
23472	0.9803146	91752	1	45.0000000	22.5000000	15.0000000	6.4285714	1.0344828
15224	0.9786295	83504	1	22.5000000	11.2500000	7.5000000	3.2142857	2.9032258
22043	0.9708119	90323	1	22.5000000	11.2500000	7.5000000	3.2142857	1.1688312
10347	0.9670025	78627	1	45.0000000	22.5000000	15.0000000	6.4285714	3.4615385
10515	0.9669711	78795	1	22.5000000	11.2500000	7.5000000	3.2142857	0.5960265
20027	0.9668874	88307	1	45.0000000	22.5000000	15.0000000	6.4285714	1.3432836
5986	0.9650111	74266	0	0.6569343	0.6569343	0.4379562	0.4692388	90.0000000
13126	0.9597665	81406	0	0.7086614	0.3543307	0.2362205	0.1012373	0.4017857
24306	0.9505489	92586	1	45.0000000	22.5000000	15.0000000	6.4285714	1.3705584
9620	0.9495970	77900	0	25.7142857	12.8571429	8.5714286	3.6734694	0.3600000
9684	0.9495970	77964	1	25.7142857	12.8571429	8.5714286	3.6734694	0.3600000
14916	0.9451182	83196	1	45.0000000	22.5000000	15.0000000	6.4285714	1.4400000
9298	0.9437336	77578	1	45.0000000	22.5000000	15.0000000	6.4285714	9.4736842
16039	0.9435496	84319	1	45.0000000	22.5000000	15.0000000	6.4285714	6.0000000
19382	0.9412072	87662	0	45.0000000	22.5000000	15.0000000	6.4285714	0.5027933
26812	0.9398193	95092	1	45.0000000	22.5000000	15.0000000	6.4285714	3.7500000
12023	0.9388383	80303	1	30.0000000	15.0000000	10.0000000	4.2857143	7.5000000
22463	0.9370430	90743	1	22.5000000	11.2500000	7.5000000	3.2142857	1.8198529
5926	0.9352995	74206	1	5.0000000	2.5000000	1.6666667	0.7142857	12.0000000

$$FDR@3\% = \frac{label = 1\ @3\%}{label = 1\ in\ Training/Testing/OOD}$$

We applied the same method to all the models and came up with the table below:

d1 ~ d10: down sample the goods from 1/1 goods-to-bads to 10/1 goods-to-bads.

Base: original training dataset, without down sample the goods, 25 independent variables

v2: dataset from project 2, 16 independent variables

Logistics	Base	v2	d1	d3	d5	d7	d10
Train	16.58%	16.94%	4.00%	6.16%	7.75%	9.34%	11.09%
Test	17.45%	19.15%	13.40%	14.68%	15.11%	14.68%	14.89%
Validate	28.63%	30.21%	29.71%	29.27%	29.58%	29.01%	29.20%

LDA	Base	v2	d1	d3	d5	d7	d10
Train	7.85%	7.91%	3.90%	3.90%	7.49%	8.47%	9.24%
Test	8.09%	8.51%	12.98%	12.98%	12.13%	12.13%	11.49%
Validate	14.29%	15.87%	29.58%	29.58%	31.73%	31.86%	31.92%

QDA	Base	v2	d1	d3	d5	d7	d10
Train	23.20%	22.54%	5.70%	10.06%	13.45%	15.81%	17.61%
Test	24.04%	23.40%	24.89%	25.11%	24.68%	24.68%	24.47%
Validate	30.09%	30.78%	29.46%	29.39%	30.28%	30.21%	30.03%

Random Forest	Base	v2	d1	d3	d5	d7	d10
Test	34.26%	34.26%	29.36%	34.04%	35.53%	34.26%	37.02%
Validate	25.03%	25.03%	20.99%	22.19%	24.40%	26.36%	25.66%

CART - rpart	Base	v2	d1	d3	d5	d7	d10
Train	0.20%	19.40%	5.10%	10.50%	14.90%	17.50%	19.10%
Test	12.30%	22.10%	0.90%	24.70%	26.00%	25.30%	21.50%
Validate	2.60%	19.70%	16.80%	20.40%	21.90%	24.60%	23.60%

CART - tree	Base	v2	d1	d3	d5	d7	d10
Train	19.40%	19.40%	5.80%	10.50%	14.90%	17.60%	19.10%
Test	21.60%	22.10%	14.00%	21.70%	24.90%	24.00%	21.50%
Validate	19.70%	19.70%	15.00%	26.20%	24.70%	26.40%	23.60%

Boosted Tree	Base	v2	d1	d3	d5	d7	d10
Train	43.43%	45.33%	6.06%	11.91%	17.56%	22.95%	29.26%
Test	37.87%	36.60%	31.70%	33.83%	34.89%	36.60%	36.81%
Validate	24.78%	26.74%	26.36%	26.30%	25.28%	26.17%	24.65%

SVM	Base	v2	d1	d3	d5	d7	d10
Train	38.24%	31.57%	27.62%	10.88%	15.86%	19.97%	23.56%
Test	26.17%	27.66%	9.57%	28.30%	32.98%	31.91%	31.70%
Validate	21.49%	23.51%	6.01%	11.44%	19.09%	19.53%	20.61%

Neural Net	Base	v2	d1	d3	d5	d7	d10
Train	37.18%	34.24%	31.43%	23.21%	29.12%	24.12%	25.13%
Test	36.81%	31.02%	26.35%	20.14%	21.12%	22.14%	23.14%
Validate	27.11%	13.12%	15.46%	12.14%	16.12%	13.12%	14.12%

KNN	Base	v2	d1	d3	d5	d7	d10
Train	29.40%	29.41%	2.10%	6.10%	9.80%	12.70%	17.80%
Test	29.80%	29.79%	3.20%	12.30%	12.80%	20.20%	28.90%
Validate	27.70%	29.01%	1.50%	9.80%	17.10%	24.80%	28.40%

We compared our models based on their FDR@3%, and did separate comparisons for our linear models and non-linear models. We compared them separately because our linear models have relatively low rates on training and testing, but scored very high in validation, but our non-linear models followed the expected pattern, with training and testing roughly the same and validation

slightly lower. Due to time limits, we haven't resolved the issue with our unexpectedly well-performed linear models, and we decided to go with our best performing non-linear model. Based on the comparison we made, KNN won the game.

KNN Model Result

We used different values for K in our K Nearest Neighbor model and found that K=120 lead to best result. Under KNN model, we have a FDR at 3% for training data of 29.41%, for testing data of 27.79%, and for validation data of 29.01%.

- Model performance statistics

KNN	FDR @3%
Train	29.41%
Test	29.79%
Validate	29.01%

- Score Distribution
  - 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
17.00%	546	510	36	7933	1351	6.59%	93.41%	69.35%	0.54
18.00%	546	527	19	8460	1370	3.48%	96.52%	70.33%	0.54
19.00%	547	523	24	8983	1394	4.39%	95.61%	71.56%	0.55
20.00%	546	516	30	9499	1424	5.49%	94.51%	73.10%	0.55
21.00%	546	519	27	10018	1451	4.95%	95.05%	74.49%	0.55
22.00%	546	524	22	10542	1473	4.03%	95.97%	75.62%	0.56
23.00%	546	523	23	11065	1496	4.21%	95.79%	76.80%	0.56
24.00%	546	532	14	11597	1510	2.56%	97.44%	77.52%	0.55
25.00%	546	523	23	12120	1533	4.21%	95.79%	78.70%	0.56
26.00%	546	537	9	12657	1542	1.65%	98.35%	79.16%	0.55
27.00%	546	524	22	13181	1564	4.03%	95.97%	80.29%	0.55
28.00%	546	534	12	13715	1576	2.20%	97.80%	80.90%	0.55
29.00%	546	533	13	14248	1589	2.38%	97.62%	81.57%	0.55
30.00%	546	522	24	14770	1613	4.40%	95.60%	82.80%	0.55
31.00%	547	535	12	15305	1625	2.19%	97.81%	83.42%	0.54
32.00%	546	538	8	15843	1633	1.47%	98.53%	83.83%	0.54
33.00%	546	541	5	16384	1638	0.92%	99.08%	84.09%	0.53
34.00%	546	530	16	16914	1654	2.93%	97.07%	84.91%	0.53
35.00%	546	533	13	17447	1667	2.38%	97.62%	85.57%	0.52
36.00%	547	533	14	17980	1681	2.56%	97.44%	86.29%	0.52
37.00%	546	527	19	18507	1700	3.48%	96.52%	87.27%	0.52
38.00%	546	534	12	19041	1712	2.20%	97.80%	87.89%	0.52
39.00%	546	534	12	19575	1724	2.20%	97.80%	88.50%	0.51
40.00%	546	532	14	20107	1738	2.56%	97.44%	89.22%	0.51
41.00%	546	533	13	20640	1751	2.38%	97.62%	89.89%	0.51
42.00%	547	537	10	21177	1761	1.83%	98.17%	90.40%	0.50
43.00%	546	536	10	21713	1771	1.83%	98.17%	90.91%	0.50
44.00%	546	530	16	22243	1787	2.93%	97.07%	91.74%	0.50
45.00%	546	536	10	22779	1797	1.83%	98.17%	92.25%	0.49
46.00%	546	538	8	23317	1805	1.47%	98.53%	92.66%	0.48
47.00%	546	537	9	23854	1814	1.65%	98.35%	93.12%	0.48
48.00%	546	540	6	24394	1820	1.10%	98.90%	93.43%	0.47
49.00%	547	535	12	24929	1832	2.19%	97.81%	94.05%	0.47
50.00%	546	537	9	25466	1841	1.65%	98.35%	94.51%	0.46
51.00%	546	539	7	26005	1848	1.28%	98.72%	94.87%	0.45
52.00%	546	533	13	26538	1861	2.38%	97.62%	95.53%	0.45

53.00%	547	543	4	27081	1865	0.73%	99.27%	95.74%	0.44
54.00%	546	545	1	27626	1866	0.18%	99.82%	95.79%	0.43
55.00%	546	541	5	28167	1871	0.92%	99.08%	96.05%	0.43
56.00%	546	542	4	28709	1875	0.73%	99.27%	96.25%	0.42
57.00%	547	543	4	29252	1879	0.73%	99.27%	96.46%	0.41
58.00%	546	542	4	29794	1883	0.73%	99.27%	96.66%	0.40
59.00%	547	542	5	30336	1888	0.91%	99.09%	96.92%	0.39
60.00%	546	542	4	30878	1892	0.73%	99.27%	97.13%	0.39
61.00%	546	535	11	31413	1903	2.01%	97.99%	97.69%	0.38
62.00%	546	544	2	31957	1905	0.37%	99.63%	97.79%	0.37
63.00%	547	542	5	32499	1910	0.91%	99.09%	98.05%	0.36
64.00%	546	541	5	33040	1915	0.92%	99.08%	98.31%	0.36
65.00%	547	543	4	33583	1919	0.73%	99.27%	98.51%	0.35
66.00%	546	544	2	34127	1921	0.37%	99.63%	98.61%	0.34
67.00%	546	543	3	34670	1924	0.55%	99.45%	98.77%	0.33
68.00%	546	544	2	35214	1926	0.37%	99.63%	98.87%	0.32
69.00%	547	539	8	35753	1934	1.46%	98.54%	99.28%	0.31
70.00%	546	543	3	36296	1937	0.55%	99.45%	99.44%	0.31
71.00%	547	542	5	36838	1942	0.91%	99.09%	99.69%	0.30
72.00%	546	541	5	37379	1947	0.92%	99.08%	99.95%	0.29
73.00%	547	546	1	37925	1948	0.18%	99.82%	100.00%	0.28
74.00%	546	546	0	38471	1948	0.00%	100.00%	100.00%	0.27
75.00%	546	546	0	39017	1948	0.00%	100.00%	100.00%	0.26
76.00%	547	547	0	39564	1948	0.00%	100.00%	100.00%	0.25
77.00%	546	546	0	40110	1948	0.00%	100.00%	100.00%	0.24
78.00%	547	547	0	40657	1948	0.00%	100.00%	100.00%	0.23
79.00%	546	546	0	41203	1948	0.00%	100.00%	100.00%	0.22
80.00%	546	546	0	41749	1948	0.00%	100.00%	100.00%	0.21
81.00%	546	546	0	42295	1948	0.00%	100.00%	100.00%	0.20
82.00%	547	547	0	42842	1948	0.00%	100.00%	100.00%	0.19
83.00%	546	546	0	43388	1948	0.00%	100.00%	100.00%	0.18
84.00%	547	547	0	43935	1948	0.00%	100.00%	100.00%	0.17
85.00%	546	546	0	44481	1948	0.00%	100.00%	100.00%	0.16
86.00%	547	547	0	45028	1948	0.00%	100.00%	100.00%	0.15
87.00%	546	546	0	45574	1948	0.00%	100.00%	100.00%	0.13
88.00%	546	546	0	46120	1948	0.00%	100.00%	100.00%	0.12
89.00%	546	546	0	46666	1948	0.00%	100.00%	100.00%	0.11
90.00%	547	547	0	47213	1948	0.00%	100.00%	100.00%	0.10

91.00%	546	546	0	47759	1948	0.00%	100.00%	100.00%	0.09
92.00%	546	546	0	48305	1948	0.00%	100.00%	100.00%	0.08
93.00%	546	546	0	48851	1948	0.00%	100.00%	100.00%	0.07
94.00%	547	547	0	49398	1948	0.00%	100.00%	100.00%	0.06
95.00%	546	546	0	49944	1948	0.00%	100.00%	100.00%	0.05
96.00%	547	547	0	50491	1948	0.00%	100.00%	100.00%	0.04
97.00%	546	546	0	51037	1948	0.00%	100.00%	100.00%	0.03
98.00%	546	546	0	51583	1948	0.00%	100.00%	100.00%	0.02
99.00%	546	546	0	52129	1948	0.00%	100.00%	100.00%	0.01
100.00%	546	546	0	52675	1948	0.00%	100.00%	100.00%	0.00

- Testing

Percentile	# of Records	# Goods	# Bads	Cumulative Goods	Cumulative Bads	% Bad	% Good	Cumulative Fraud Detection Rate	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
17.00%	137	132	5	2014	308	3.65%	96.35%	65.53%	0.50
18.00%	136	135	1	2149	309	0.74%	99.26%	65.74%	0.49
19.00%	137	128	9	2277	318	6.57%	93.43%	67.66%	0.50
20.00%	136	130	6	2407	324	4.41%	95.59%	68.94%	0.51
21.00%	137	137	0	2544	324	0.00%	100.00%	68.94%	0.50
22.00%	136	133	3	2677	327	2.21%	97.79%	69.57%	0.49
23.00%	137	131	6	2808	333	4.38%	95.62%	70.85%	0.50

24.00%	136	131	5	2939	338	3.68%	96.32%	71.91%	0.50
25.00%	137	128	9	3067	347	6.57%	93.43%	73.83%	0.51
26.00%	137	136	1	3203	348	0.73%	99.27%	74.04%	0.50
27.00%	136	135	1	3338	349	0.74%	99.26%	74.26%	0.49
28.00%	137	130	7	3468	356	5.11%	94.89%	75.74%	0.49
29.00%	136	136	0	3604	356	0.00%	100.00%	75.74%	0.48
30.00%	137	127	10	3731	366	7.30%	92.70%	77.87%	0.50
31.00%	136	131	5	3862	371	3.68%	96.32%	78.94%	0.50
32.00%	137	137	0	3999	371	0.00%	100.00%	78.94%	0.49
33.00%	136	135	1	4134	372	0.74%	99.26%	79.15%	0.48
34.00%	137	134	3	4268	375	2.19%	97.81%	79.79%	0.47
35.00%	137	131	6	4399	381	4.38%	95.62%	81.06%	0.48
36.00%	136	132	4	4531	385	2.94%	97.06%	81.91%	0.48
37.00%	137	133	4	4664	389	2.92%	97.08%	82.77%	0.47
38.00%	136	131	5	4795	394	3.68%	96.32%	83.83%	0.47
39.00%	137	135	2	4930	396	1.46%	98.54%	84.26%	0.47
40.00%	136	135	1	5065	397	0.74%	99.26%	84.47%	0.46
41.00%	137	137	0	5202	397	0.00%	100.00%	84.47%	0.45
42.00%	137	137	0	5339	397	0.00%	100.00%	84.47%	0.44
43.00%	136	133	3	5472	400	2.21%	97.79%	85.11%	0.44
44.00%	137	136	1	5608	401	0.73%	99.27%	85.32%	0.43
45.00%	136	132	4	5740	405	2.94%	97.06%	86.17%	0.43
46.00%	137	135	2	5875	407	1.46%	98.54%	86.60%	0.42
47.00%	136	133	3	6008	410	2.21%	97.79%	87.23%	0.42
48.00%	137	132	5	6140	415	3.65%	96.35%	88.30%	0.42
49.00%	136	133	3	6273	418	2.21%	97.79%	88.94%	0.41
50.00%	137	136	1	6409	419	0.73%	99.27%	89.15%	0.41
51.00%	137	136	1	6545	420	0.73%	99.27%	89.36%	0.40
52.00%	136	134	2	6679	422	1.47%	98.53%	89.79%	0.39
53.00%	137	136	1	6815	423	0.73%	99.27%	90.00%	0.38
54.00%	136	134	2	6949	425	1.47%	98.53%	90.43%	0.38
55.00%	137	137	0	7086	425	0.00%	100.00%	90.43%	0.37
56.00%	136	136	0	7222	425	0.00%	100.00%	90.43%	0.36
57.00%	137	136	1	7358	426	0.73%	99.27%	90.64%	0.35
58.00%	136	133	3	7491	429	2.21%	97.79%	91.28%	0.34
59.00%	137	137	0	7628	429	0.00%	100.00%	91.28%	0.33
60.00%	137	136	1	7764	430	0.73%	99.27%	91.49%	0.33
61.00%	136	133	3	7897	433	2.21%	97.79%	92.13%	0.32

62.00%	137	134	3	8031	436	2.19%	97.81%	92.77%	0.32
63.00%	136	135	1	8166	437	0.74%	99.26%	92.98%	0.31
64.00%	137	135	2	8301	439	1.46%	98.54%	93.40%	0.30
65.00%	136	135	1	8436	440	0.74%	99.26%	93.62%	0.30
66.00%	137	134	3	8570	443	2.19%	97.81%	94.26%	0.29
67.00%	137	135	2	8705	445	1.46%	98.54%	94.68%	0.29
68.00%	136	136	0	8841	445	0.00%	100.00%	94.68%	0.28
69.00%	137	133	4	8974	449	2.92%	97.08%	95.53%	0.27
70.00%	136	136	0	9110	449	0.00%	100.00%	95.53%	0.26
71.00%	137	136	1	9246	450	0.73%	99.27%	95.74%	0.26
72.00%	136	134	2	9380	452	1.47%	98.53%	96.17%	0.25
73.00%	137	137	0	9517	452	0.00%	100.00%	96.17%	0.24
74.00%	136	136	0	9653	452	0.00%	100.00%	96.17%	0.23
75.00%	137	137	0	9790	452	0.00%	100.00%	96.17%	0.22
76.00%	137	137	0	9927	452	0.00%	100.00%	96.17%	0.21
77.00%	136	136	0	10063	452	0.00%	100.00%	96.17%	0.20
78.00%	137	137	0	10200	452	0.00%	100.00%	96.17%	0.19
79.00%	136	136	0	10336	452	0.00%	100.00%	96.17%	0.18
80.00%	137	137	0	10473	452	0.00%	100.00%	96.17%	0.17
81.00%	136	136	0	10609	452	0.00%	100.00%	96.17%	0.16
82.00%	137	137	0	10746	452	0.00%	100.00%	96.17%	0.15
83.00%	136	135	1	10881	453	0.74%	99.26%	96.38%	0.14
84.00%	137	136	1	11017	454	0.73%	99.27%	96.60%	0.13
85.00%	137	137	0	11154	454	0.00%	100.00%	96.60%	0.12
86.00%	136	135	1	11289	455	0.74%	99.26%	96.81%	0.11
87.00%	137	136	1	11425	456	0.73%	99.27%	97.02%	0.10
88.00%	136	136	0	11561	456	0.00%	100.00%	97.02%	0.09
89.00%	137	136	1	11697	457	0.73%	99.27%	97.23%	0.09
90.00%	136	133	3	11830	460	2.21%	97.79%	97.87%	0.08
91.00%	137	136	1	11966	461	0.73%	99.27%	98.09%	0.07
92.00%	137	135	2	12101	463	1.46%	98.54%	98.51%	0.07
93.00%	136	136	0	12237	463	0.00%	100.00%	98.51%	0.06
94.00%	137	136	1	12373	464	0.73%	99.27%	98.72%	0.05
95.00%	136	135	1	12508	465	0.74%	99.26%	98.94%	0.04
96.00%	137	135	2	12643	467	1.46%	98.54%	99.36%	0.03
97.00%	136	136	0	12779	467	0.00%	100.00%	99.36%	0.02
98.00%	137	136	1	12915	468	0.73%	99.27%	99.57%	0.02
99.00%	136	136	0	13051	468	0.00%	100.00%	99.57%	0.01

100.00%	137	135	2	13186	470	1.46%	98.54%	100.00%	0.00

- 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
17.00%	270	236	34	3648	941	12.59%	87.41%	59.48%	0.45
18.00%	270	257	13	3905	954	4.81%	95.19%	60.30%	0.45
19.00%	269	244	25	4149	979	9.29%	90.71%	61.88%	0.46
20.00%	270	234	36	4383	1015	13.33%	86.67%	64.16%	0.47
21.00%	270	250	20	4633	1035	7.41%	92.59%	65.42%	0.47
22.00%	270	255	15	4888	1050	5.56%	94.44%	66.37%	0.47
23.00%	270	248	22	5136	1072	8.15%	91.85%	67.76%	0.48
24.00%	270	244	26	5380	1098	9.63%	90.37%	69.41%	0.48
25.00%	270	257	13	5637	1111	4.81%	95.19%	70.23%	0.48
26.00%	270	220	50	5857	1161	18.52%	81.48%	73.39%	0.50
27.00%	270	250	20	6107	1181	7.41%	92.59%	74.65%	0.51
28.00%	270	258	12	6365	1193	4.44%	95.56%	75.41%	0.50
29.00%	270	253	17	6618	1210	6.30%	93.70%	76.49%	0.50
30.00%	270	255	15	6873	1225	5.56%	94.44%	77.43%	0.50
31.00%	270	261	9	7134	1234	3.33%	96.67%	78.00%	0.50
32.00%	269	255	14	7389	1248	5.20%	94.80%	78.89%	0.50

33.00%	270	253	17	7642	1265	6.30%	93.70%	79.96%	0.50
34.00%	270	258	12	7900	1277	4.44%	95.56%	80.72%	0.50
35.00%	270	257	13	8157	1290	4.81%	95.19%	81.54%	0.49
36.00%	270	264	6	8421	1296	2.22%	97.78%	81.92%	0.49
37.00%	270	264	6	8685	1302	2.22%	97.78%	82.30%	0.48
38.00%	270	266	4	8951	1306	1.48%	98.52%	82.55%	0.47
39.00%	270	263	7	9214	1313	2.59%	97.41%	83.00%	0.47
40.00%	270	256	14	9470	1327	5.19%	94.81%	83.88%	0.47
41.00%	270	264	6	9734	1333	2.22%	97.78%	84.26%	0.46
42.00%	270	259	11	9993	1344	4.07%	95.93%	84.96%	0.46
43.00%	270	256	14	10249	1358	5.19%	94.81%	85.84%	0.46
44.00%	269	257	12	10506	1370	4.46%	95.54%	86.60%	0.45
45.00%	270	258	12	10764	1382	4.44%	95.56%	87.36%	0.45
46.00%	270	266	4	11030	1386	1.48%	98.52%	87.61%	0.44
47.00%	270	266	4	11296	1390	1.48%	98.52%	87.86%	0.43
48.00%	270	266	4	11562	1394	1.48%	98.52%	88.12%	0.43
49.00%	270	265	5	11827	1399	1.85%	98.15%	88.43%	0.42
50.00%	270	268	2	12095	1401	0.74%	99.26%	88.56%	0.41
51.00%	270	265	5	12360	1406	1.85%	98.15%	88.87%	0.40
52.00%	270	266	4	12626	1410	1.48%	98.52%	89.13%	0.39
53.00%	270	261	9	12887	1419	3.33%	96.67%	89.70%	0.39
54.00%	270	263	7	13150	1426	2.59%	97.41%	90.14%	0.38
55.00%	270	269	1	13419	1427	0.37%	99.63%	90.20%	0.37
56.00%	270	265	5	13684	1432	1.85%	98.15%	90.52%	0.37
57.00%	269	261	8	13945	1440	2.97%	97.03%	91.02%	0.36
58.00%	270	265	5	14210	1445	1.85%	98.15%	91.34%	0.35
59.00%	270	268	2	14478	1447	0.74%	99.26%	91.47%	0.34
60.00%	270	262	8	14740	1455	2.96%	97.04%	91.97%	0.34
61.00%	270	261	9	15001	1464	3.33%	96.67%	92.54%	0.34
62.00%	270	260	10	15261	1474	3.70%	96.30%	93.17%	0.33
63.00%	270	267	3	15528	1477	1.11%	98.89%	93.36%	0.32
64.00%	270	264	6	15792	1483	2.22%	97.78%	93.74%	0.32
65.00%	270	267	3	16059	1486	1.11%	98.89%	93.93%	0.31
66.00%	270	267	3	16326	1489	1.11%	98.89%	94.12%	0.30
67.00%	270	267	3	16593	1492	1.11%	98.89%	94.31%	0.29
68.00%	270	265	5	16858	1497	1.85%	98.15%	94.63%	0.28
69.00%	269	264	5	17122	1502	1.86%	98.14%	94.94%	0.28
70.00%	270	268	2	17390	1504	0.74%	99.26%	95.07%	0.27

71.00%	270	267	3	17657	1507	1.11%	98.89%	95.26%	0.26
72.00%	270	264	6	17921	1513	2.22%	97.78%	95.64%	0.25
73.00%	270	266	4	18187	1517	1.48%	98.52%	95.89%	0.24
74.00%	270	265	5	18452	1522	1.85%	98.15%	96.21%	0.24
75.00%	270	267	3	18719	1525	1.11%	98.89%	96.40%	0.23
76.00%	270	268	2	18987	1527	0.74%	99.26%	96.52%	0.22
77.00%	270	264	6	19251	1533	2.22%	97.78%	96.90%	0.21
78.00%	270	266	4	19517	1537	1.48%	98.52%	97.16%	0.20
79.00%	270	267	3	19784	1540	1.11%	98.89%	97.35%	0.19
80.00%	270	267	3	20051	1543	1.11%	98.89%	97.53%	0.19
81.00%	270	269	1	20320	1544	0.37%	99.63%	97.60%	0.18
82.00%	269	268	1	20588	1545	0.37%	99.63%	97.66%	0.17
83.00%	270	269	1	20857	1546	0.37%	99.63%	97.72%	0.16
84.00%	270	268	2	21125	1548	0.74%	99.26%	97.85%	0.15
85.00%	270	266	4	21391	1552	1.48%	98.52%	98.10%	0.14
86.00%	270	266	4	21657	1556	1.48%	98.52%	98.36%	0.13
87.00%	270	266	4	21923	1560	1.48%	98.52%	98.61%	0.12
88.00%	270	268	2	22191	1562	0.74%	99.26%	98.74%	0.11
89.00%	270	268	2	22459	1564	0.74%	99.26%	98.86%	0.10
90.00%	270	265	5	22724	1569	1.85%	98.15%	99.18%	0.10
91.00%	270	269	1	22993	1570	0.37%	99.63%	99.24%	0.09
92.00%	270	269	1	23262	1571	0.37%	99.63%	99.30%	0.08
93.00%	270	269	1	23531	1572	0.37%	99.63%	99.37%	0.07
94.00%	269	269	0	23800	1572	0.00%	100.00%	99.37%	0.06
95.00%	270	267	3	24067	1575	1.11%	98.89%	99.56%	0.05
96.00%	270	269	1	24336	1576	0.37%	99.63%	99.62%	0.04
97.00%	270	269	1	24605	1577	0.37%	99.63%	99.68%	0.03
98.00%	270	268	2	24873	1579	0.74%	99.26%	99.81%	0.02
99.00%	270	267	3	25140	1582	1.11%	98.89%	100.00%	0.01
100.00%	270	270	0	25410	1582	0.00%	100.00%	100.00%	0.00

Results and Summary

Through the process of exploring data, building variables and running models, the supervised learning method gave us a good practice of fraud detection. To better train the model, we used different training data sets with different nonfraud-to-fraud ratios. We also tried different linear and nonlinear models including logistics, LDA, QDA, SVM, random forest, etc. Although the FDR @3% for training, testing and validating datasets vary a lot under different models, our best

result comes from KNN method. FDR @3% is just one aspect of testing the effectiveness of the models. To make further improvements, we probably can study further on the records that are labeled fraud, trying to find the hidden pattern behind them and to come up with better variables for model testing.

Appendix-Data quality report

Dataset Name	Payments
File format	Excel
# of records	95271
# of fields	1 Unique Identifier 8 Independent Variables 1 Dependent Variable (fraud?)

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 0	4000 91271

Independent Variables

Numerical Fields	Field Name	Description	Length	Non-missing	Missing	Missing Percent	Cumulative Distribution		Standard Deviation	Mean	Min	Max
1	AMOUNT	Payment amount	7	95271	0	0.00%	p1 p5 p10 p25 p50 p75 p90 p95 p99	3.57 3.62 4.37 33.2 136.31 420.09 1054.63 1721.00 2486.58	595.99	380.78	0.01	47900

Categorical Fields	Field Name	Description	Length	Non-missing	Missing	Missing Percent	Frequent Value	Counts	Unique
1	CARDNUM	Card number	10	95271	0	0.00%	5142148452 5142184598 5142189108 5142297710 5142223373 5142187452 5142299634 5142189945 5142149691 5142190147	1192 921 663 583 577 526 515 512 497 488	1636
2	MERCHNUM	Merchant number	13	91921	3350	3.52%	930090121224 5509006296254 9900020006406 602608969534 410000971343 9918000409955 4353000719908 5725000466504 9108234610000 602608969138	9157 2131 1713 1091 981 953 940 868 784 782	6625
3	MERCHDESCRIPTION	Merchant description	25	95271	0	0.00%	GSA-FSS-ADV SIGMA-ALDRICH STAPLES #941 FISHER SCI ATL MWI*MICRO WAREHOUSE CDW*GOVERNMENT INC DELL MARKETING L.P. FISHER SCI CHI OFFICE DEPOT #1082 AMAZON.COM*F SUPERSTOR	1663 1632 1131 1092 955 868 804 782 744 723	13122
4	MERCHSTATE	State of merchant's location	5	94081	1190	1.25%	TN VA CA IL MD GA PA NJ TX WA	11834 7674 6677 6485 5343 4930 4844 3904 3748 3215	228
5	TRANSTYPE	Transaction Type	1	95271	0	0.00%	P A D Y	94916 181 173 1	4
6	MERCHZIP	Zip Code of merchant's location	5	90638	4633	5.11%	38118 63103 8701 60061 22202 17201 98101 30091 60143 60069	11669 1647 1263 1217 1197 1136 1117 1091 940 825	4568
7	DATE	Date of merchant	8	95271	0	0.00%	5/17/10 7/4/10 6/24/10 9/16/10 5/28/10 5/27/10 9/15/10 7/5/10 5/29/10 9/17/10	30 19 18 13 11 9 9 8 7 5	67