Fraud Detection

Fraud Detection using Machine Learning and Deep Learning

Pattern Recognition Project

Shiva Zeymaran - Fahime Vafi

Summer 2023

Overview

5

Introduction

Why Fraud Detection is important?

6

Datasets

Three datasets used in this paper in detail

7-15

German Dataset

Preprocess, Evaluation Metrics, Multiple Machine Learning Methods, Ensemble Classifier, Results, Compare with Article Results

16-19

Australian Dataset

Preprocess, Multiple Machine Learning Methods, Results, Compare with Article Results

20-23

European Dataset

Multiple Machine Learning Methods, Results, Compare with Article Results

24

Conclusion

25

References

Introduction

Credit card fraud is a form of fraud involving the use of fake or stolen credit card information and causing financial harm to account holders or merchants involved.

Frauds are known to be dynamic and have no patterns, hence they are not easy to identify.

Fraudsters somehow bypass security checks, leading to the loss of millions of dollars.

Detecting unusual activities using data mining techniques

Datasets

Large

European Dataset

- 284,807 instances
- 492 fraud instances
- 28 PCA transformed fields
- Time, Amount and Label

Small

Australian Dataset

- 690 instances
- 307 fraud instances
- 14 attributes + class label
- anonymized (no personal information)

Small

German Dataset

- 1000 instances
- 300 fraud instances
- 20 attributes + class label
- anonymized (no personal information)

GERMAN DATASET

German Dataset

7 Numerical

- Duration in month
- Credit amount
- Age
- Number of existing credits at this bank
- Etc.

13 Categorical

- Credit history
- Purpose
- Property
- Housing
- Job
- Etc.

	0	1	2	3	4	5	6	7	8	9	...	11	12	13	14	15	16	17	18	19	20
0	A11	6	A34	A43	1169	A65	A75	4	A93	A101	...	A121	67	A143	A152	2	A173	1	A192	A201	1
1	A12	48	A32	A43	5951	A61	A73	2	A92	A101	...	A121	22	A143	A152	1	A173	1	A191	A201	2
2	A14	12	A34	A46	2096	A61	A74	2	A93	A101	...	A121	49	A143	A152	1	A172	2	A191	A201	1
3	A11	42	A32	A42	7882	A61	A74	2	A93	A103	...	A122	45	A143	A153	1	A173	2	A191	A201	1
4	A11	24	A33	A40	4870	A61	A73	3	A93	A101	...	A124	53	A143	A153	2	A173	2	A191	A201	2

Preprocess

1
Feature/Label

Separate features and class labels

2
One Hot

Change categorical column values to one hot

3
Split Data

80% for train and 20% for test dataset

train dataset: (800, 61)
test dataset: (200, 61)

4
Missed Values

Check if there was any missed values and remove it

There was not any

5
Normalization

Normalize numerical column values using Standard Scalar (mean=0, std=1)

Methods

KNN

Accuracy: 0.78

Grid Search

- To Find best value of K
- Test for k = [30, 40]

```
Best: 0.739354 using {'n_neighbors': 33}
```

	precision	recall	f1-score	support
1	0.78	0.99	0.87	149
2	0.83	0.20	0.32	51
accuracy			0.79	200
macro avg	0.81	0.59	0.59	200
weighted avg	0.80	0.79	0.73	200

MCC

0.33

AUC

0.59

Cost of Failure

41,200

Methods

SVM

C = [1, 10, 100, 1000], kernel = linear *(OR)*
C = [1, 10, 100, 1000], kernel = rbf, gamma = [0.1, 0.01, 0.001, 0.0001]
Best C = 10, kernel = rbf, gamma = 0.001

RF

Random Forest

max_depth = [3, 5, None],
n_estimators = [3, 5, 10],
max_features = [5, 6, 7, 8]
Best Max_depth = 5, max_features = 8, n_estimators = 10

LR

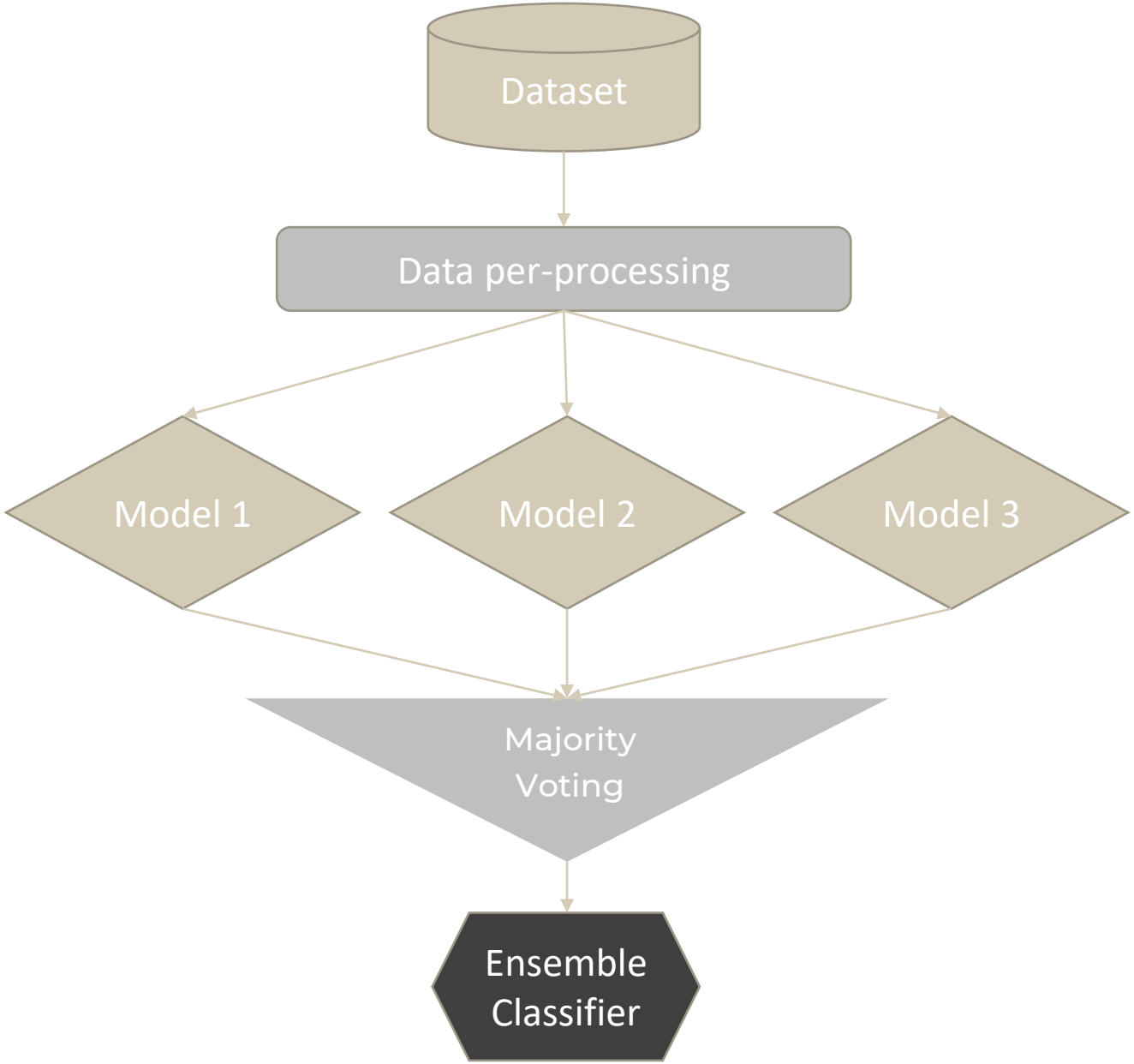
Logistic Regression

NB

Naïve Bayes

Ensemble Classifier

- Choose the Top 3 performing models
- Combine them using Majority Voting



Evaluation Metrics

MCC

Matthews Correlation Coefficient/
Phi coefficient

- Evaluate the quality of a binary classifier
- Consider all four values of the confusion matrix

- **from sklearn.metrics import matthews_corrcoef**

AUC

Area Under the Curve

- ROC curve helps in determining the precision because of the imbalance dataset

- **from sklearn.metrics import roc_auc_score**

Cost of Failure

- When similar AUC-> use Cost of failure
- Each of the FN have a cost of $1000 and FP have a cost of $100
- **(fn*1000) + (fp*100)**

Results

German Dataset

Methods	MCC	AUC	Cost of Failure
KNN	0.33	0.59	41200
SVM	0.53	0.75	21400
RF	0.29	0.59	39600
LR	0.44	0.67	30800
NB	0.49	0.77	14500
Ensemble (SVM, LR, NB)	**0.56**	**0.77**	**20300**

01

SVM, LR, and NB have the best performance in terms of MCC and AUC.
Top 3: SVM, LR and NB

02

The ensemble method performs better than SVM, LR, and NB individually.

Article Results

German Dataset

Method	*MCC*	*AUC*	*Cost of Failure*
RBM	0.0984	0.5524	**14160**
Autoencoders	0.139	0.5614	**22640**
KNN	0.2487	0.6047	**21100**
DBN	0.2725	0.5873	**23640**
Random Forest	0.2912	0.6437	**16970**
SVM	0.4038	0.6857	**16400**
CNN	0.4291	0.7056	**14220**
Ensemble (SVM, CNN, Random Forest)	**0.4439**	**0.7011**	**15620**

Recommendation:
choose Ensemble classifier

1. Same as the original results of the article the ensemble classifier works better
2. We reached higher MCC and AUC values which means our method works better than the article

AUSTRALIAN DATASET

Australian Dataset

Preprocess

Feature/Label	One Hot	Split Data	Missed Values	Normalization

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	1	22.08	11.46	2	4	4	1.585	0	0	0	1	2	100	1213	0
1	0	22.67	7.00	2	8	4	0.165	0	0	0	0	2	160	1	0
2	0	29.58	1.75	1	4	4	1.250	0	0	0	1	2	280	1	0
3	0	21.67	11.50	1	5	3	0.000	1	1	11	1	2	0	1	1
4	1	20.17	8.17	2	6	4	1.960	1	1	14	0	2	60	159	1

Results

01

SVM, LR, and RF have the best performance in terms of MCC and AUC.

The ensemble method performs better than all.

Australian Dataset

Methods	MCC	AUC	Cost of Failure
KNN	0.67	0.83	14800
SVM	0.68	0.84	6800
RF	0.72	0.85	10900
LR	0.68	0.84	9400
NB	0.64	0.82	13200
Ensemble (SVM, LR, RF)	**0.69**	**0.85**	**8400**

Article Results

Recommendation:
choose Ensemble classifier

1. Same as the original results of the article the ensemble classifier works better

Australian Dataset

Method	*MCC*	*AUC*	*Cost of Failure*
RBM	0.15	0.5546	**24600**
Autoencoders	0.2318	0.6174	**12220**
CNN	0.6408	0.8227	**6430**
Random Forest	0.684	0.8416	**4700**
KNN	0.6905	0.8425	**6460**
DBN	0.6999	0.8441	**6790**
SVM	0.7085	0.8551	**3380**
Ensemble1 (KNN, SVM, DBN)	**0.7144**	**0.8573**	**5290**
Ensemble2 (KNN, SVM, Random Forest)	**0.7281**	**0.8655**	**3470**

European DATASET

Results

European Dataset

Methods	MCC	AUC	Cost of Failure
KNN	0.82	0.88	23100
SVM	0.87	0.90	19500
RF	0.82	0.86	27500
LR	0.66	0.76	48000
NB	0.23	0.92	141200
Ensemble (SVM, KNN, RF)	**0.86**	**0.90**	**18700**

01

SVM, KNN, and RF have the best performance.

02

The ensemble method performs better than KNN and RF individually. However it has a similar AUC to SVM. But, SVM has a better MCC value.

Article Results

European Dataset

Method	*MCC*	*AUC*	*Cost of Failure*
RBM	0.176	0.9109	**227360**
Autoencoders	0.2315	0.8943	**127220**
Random Forest	0.7947	0.8507	**30340**
CNN	0.8096	0.8764	**25700**
SVM	0.8145	0.9004	**21220**
KNN	0.8354	0.8887	**22660**
Ensemble (KNN, SVM and CNN)	**0.8226**	**0.8964**	**21740**

Recommendation:
choose SVM instead of the Ensemble

1. Same as the original results of the article the SVM classifier is better to use
2. We reached higher MCC and AUC values and less cost in SVM; which means our method works better than the article

Results

Method	Number of times in Top 3
SVM	3 Times
RF	2 Times
LR	2 Times
KNN	1 Time
NB	1 Time

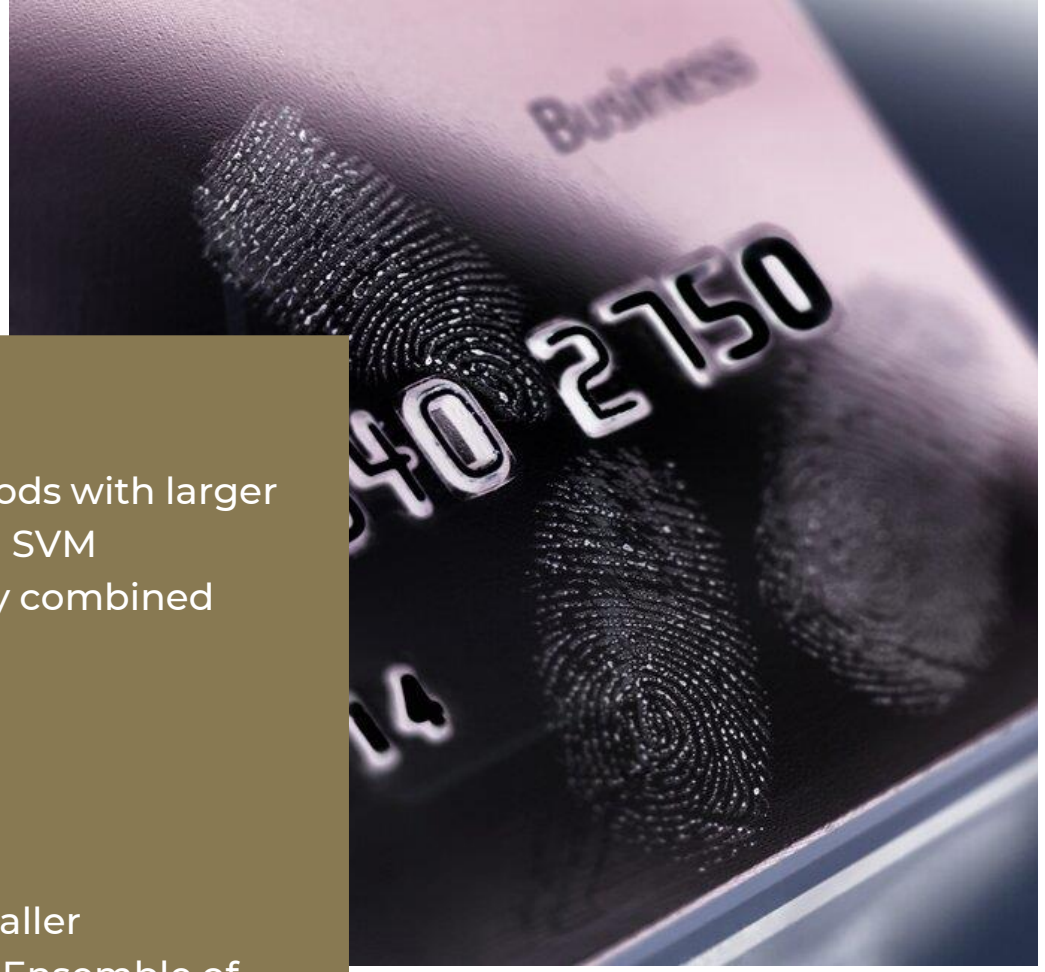
01

Ensemble works better for smaller datasets
(For the European dataset, SVM was better)

02

SVM was among the best performing models of all data sets.
RF also had good results with both large and smaller datasets.

Conclusion



I. The main aim of this study is to find which methods would best suitable for which type of datasets

II. This paper could help companies to better understand how different methods work on certain types of datasets

III. Best methods with larger datasets -> SVM (potentially combined with KNN)

IV. For the smaller datasets-> Ensemble of SVM, RF and LR

References

Raghavan P, El Gayar N. Fraud detection using machine learning and deep learning. In2019 international conference on computational intelligence and knowledge economy (ICCIKE) 2019 Dec 11 (pp. 334-339). IEEE.

Credit card fraud detection anonymized credit card transaction labeled as fraudulent or genuine [Online]. Available: https://www.kaggle.com/mlg-ulb/creditcardfraud.

Dheeru Dua and Casey Graff. UCI Machine Learning Repository. 2017 [Online]. Available: http://archive.ics.uci.edu/ml/datasets/

THANKS

end.