Fraud Detection

Fraud Detection using Machine Learning and Deep Learning

Pattern Recognition Project

Shiva Zeymaran

Spring 2023

Overview

5

Introduction

Why Fraud Detection is important?

6 - 7

Datasets

Three datasets used in this paper in detail

8 - 10

Methods

Using Multiple Machine Learning and Deep Learning Methods

11

Evaluation Metrics

MCC, AUC, and Cost of Failure

12 - 15

Results

Analyze the results of using multiple methods

16

Conclusion

17

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

7 Numerical

- Duration in month
- Credit amount
- Installment rate in percentage of disposable income
- Present residence since
- Age
- Number of existing credits at this bank
- Number of people being liable to provide maintenance for

13 Categorical

- Credit history
- Purpose (car, furniture/equipment, education, business, …)
- Present employment since
- Personal status and sex
- Property
- Housing
- Job
- Etc.

Methods

Machine Learning Methods

KNN

K-nearest neighbor
Using cross-validation
to find best value of K
(for each dataset)

SVM

Grid based search used
to find the best parameters
(C, kernel, gamma)

Random Forest

Grid based search used
to find the best parameters
(max depth, number of
estimators, max features)

Methods

Deep Learning Methods

Autoencoders

Train only on the
normal transaction.
A threshold is set.
if
reconstruction error > threshold
=> fraudulent

RBM

Restricted Boltzmann Machine
Produces free energy that tested against a threshold to determine normal vs. fraudulent

DBN

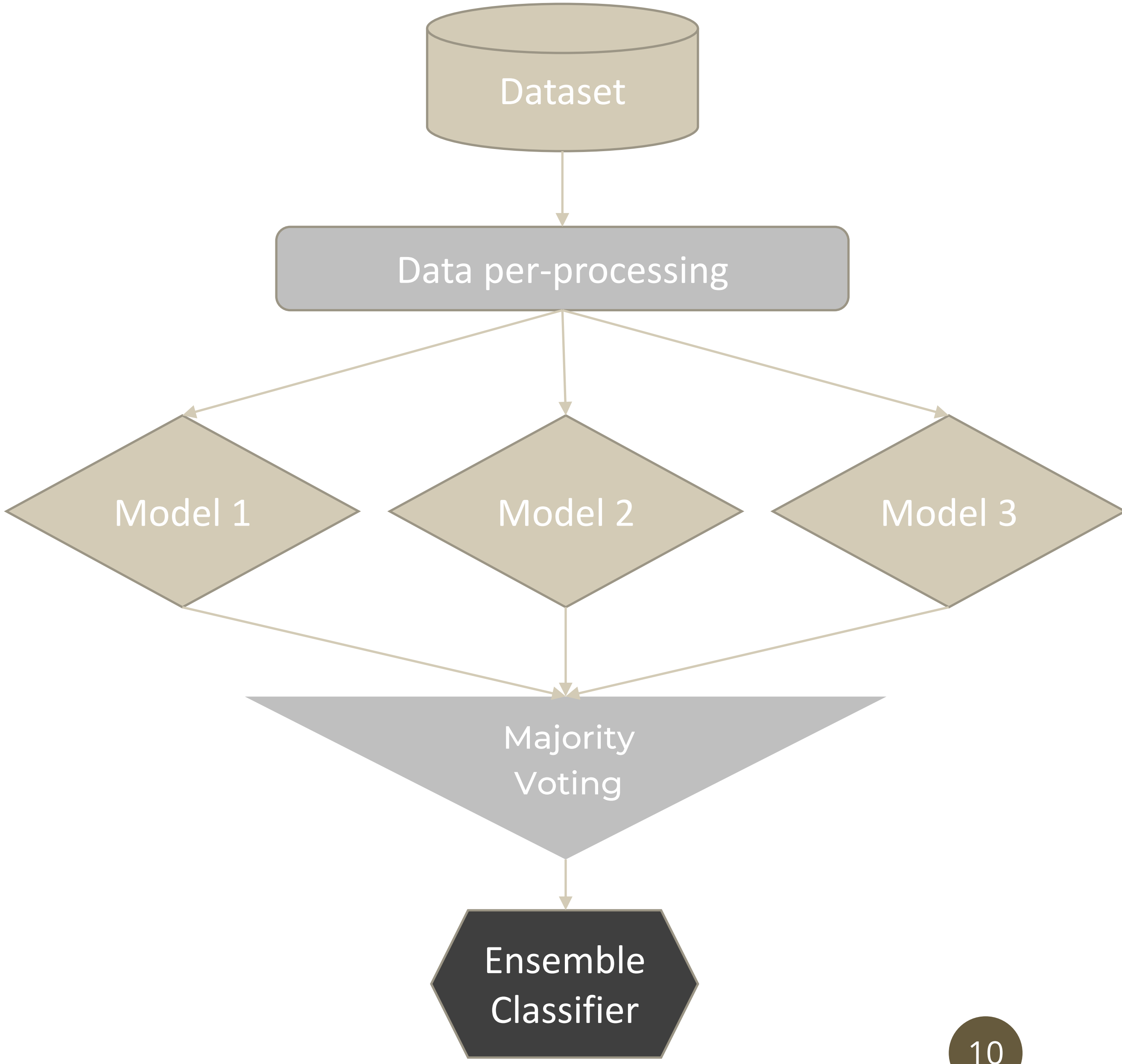
Deep Belief Network

CNN

Series of convolutional layers, max-pooling layers and a flatten layer.
Then data is classified at SoftMax layer

Ensemble Classifier

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



Evaluation Metrics

MCC

Matthews Correlation Coefficient/ Phi coefficient

- Measure to evaluate the quality of a two class/binary classifier
- Returns +1 for a perfect prediction, 0 for a random prediction
- much better than accuracy and F1 score (they do not consider all four values of the confusion matrix)

AUC

Area Under the Curve

- ROC curve helps in determining the precision of the model because of the imbalance in the dataset
- ROC curve is TPR on x and FPR on y

Cost of Failure

- Sometimes two ROC curves have the similar AUC, then we use the Cost of failure
- Each of the FN (Frauds detected as Normal) have a cost of $1000 and FP (Normal instances detected as fraud) have a cost of $100 to the company/entity

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**

01

RBM and AE have high false positives and perform poorly with respect to MCC and cost.
Random Forest has good AUC and MCC values but poor in terms of the cost

02

CNN, SVM, and KNN have the best performance in terms of MCC and AUC.
SVM has the least in terms of cost of failure.
Top 3: SVM, KNN and CNN

03

The ensemble method performs better than SVM and CNN individually, however it has a similar cost to SVM.
But, SVM has a better AUC value.

Recommendation:
choose SVM instead of the Ensemble

Results

Australian Dataset

01

Ensemble 1: KNN, DBN, and SVM
Improved the MCC and AUC performance compared to single SVM and others.
The cost is higher than the RF and SVM. (Because KNN and DBN have a high cost of failures)

02

Ensemble 2: using the models with the least cost of failures: KNN, SVM and Random Forest
Achieved a higher MCC, AUC, and lower cost value

Recommendation: choose Ensemble 2

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**

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**

01

RF, CNN and SVM also have a better cost of failure than other models.
Use an ensemble of these three models to build the majority voting classifier.

Results

Method	Number of times in Top 3
Support Vector Machines	3 Times
K-Nearest Neighbors	2 Times
Convolutional Neural Networks	2 Times
Random Forest	2 Times
Deep Belief Network	1 Time

01

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

02

Random Forest works best for smaller datasets

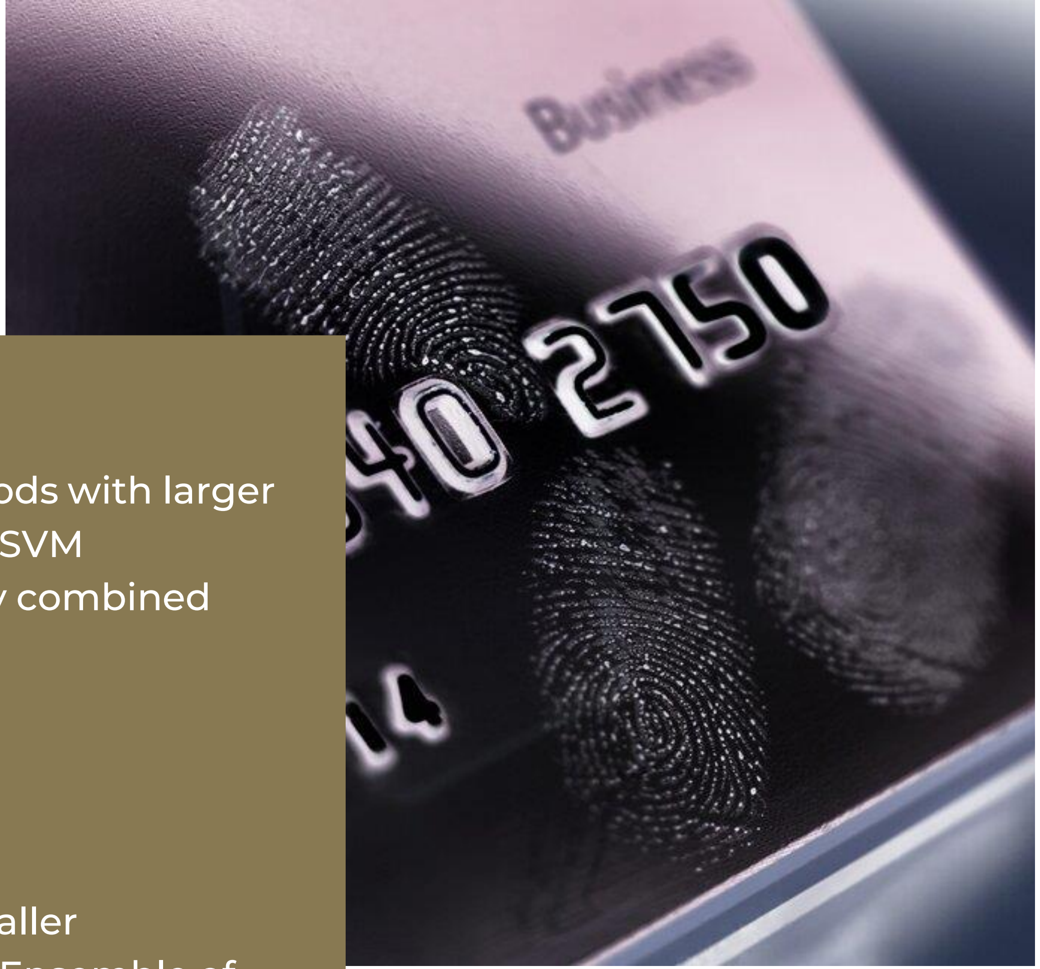
03

CNN was the best deep learning method

04

SVM was among the best performing models of all data sets. KNN 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 CNN)

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

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.