ISTANBUL TECHNICAL UNIVERSITY

DEPARTMENT OF COMPUTER ENGINEERING

Credit Card Fraud Detection with NCA Dimensionality Reduction

Beyazıt Bestami YÜKSEL

yukselbe18@itu.edu.tr

SINCONF 2020

13th International Conference on

Security of Information and Networks

Nov 2020

Outline

- Introduction
- Fraud Detection and Machine Learning
- Proposed System
- Experimental Analysis
- Conclusion
- Q&A

Introduction

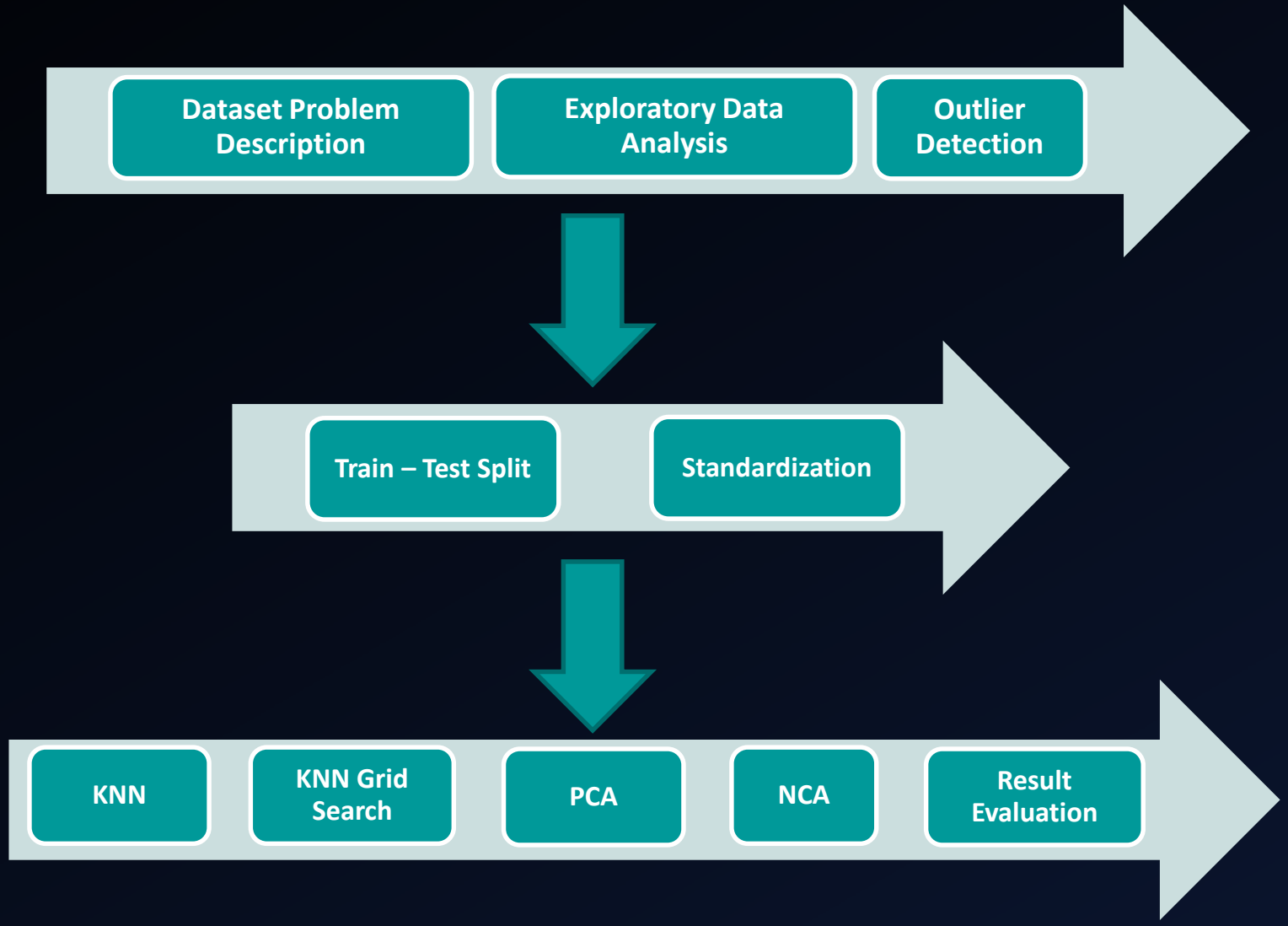
- Credit card has a significant role in today's economy. Therefore, we need to be very careful about the increased fraud activities.
- A proactive way is required to prevent credit card fraud.
- Machine Learning algorithms have a very important place in predictable system designs.
- Suggestion for performance enhancement for K-Nearest Neighbor (KNN) algorithm in Fraud Detection
- The success of the realized system
- Challenges

Fraud Detection and Machine Learning

- Credit card fraud activities continue to be the subject of many studies as it causes huge financial losses and remains current.
- According to the literature, it is possible to group the studies on this subject in terms of the methods used in general

- Random Forest
- Decision Trees
- Support Vector Machine
- Spark
- Deep Learning Methods
- Adaboost
- Majority Vooting
- Online risk fraud scoring system and Offline fraud prediction systems
- KNN
- Artificial Neural Network
- Cost-Sensitive Approaches

Proposed System

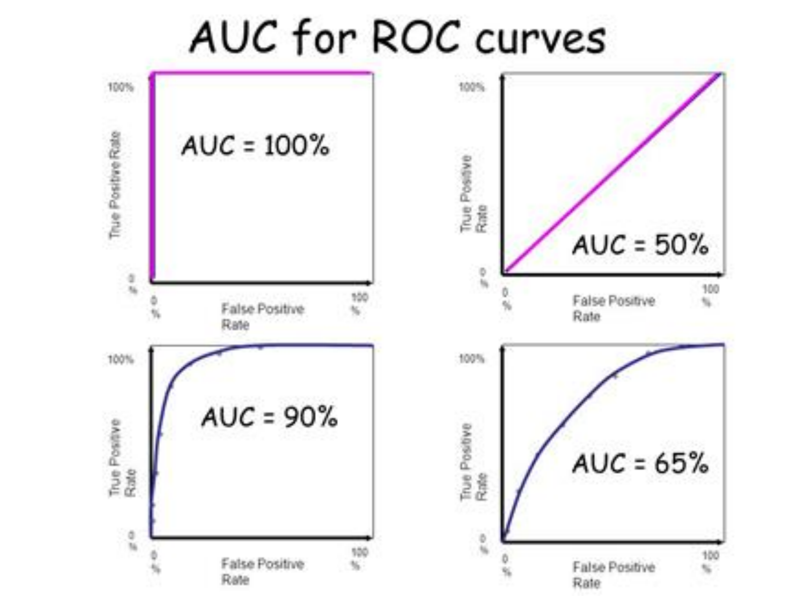


Dataset Description

- The Dataset consists of 284,807 records in total
- Only 492 of them are fraudulent cases
- 0.172% of fraud cases resulting in extremely imbalance.
- Features V1 to V28 are numerical values
- The attributes that are not changed by PCA are Time and Amount.
- 'time' contains the seconds gone by between each trade
- 'amount' is the total exchange Amount.
- Feature 'Class' is the response variable

Evaluation Metric

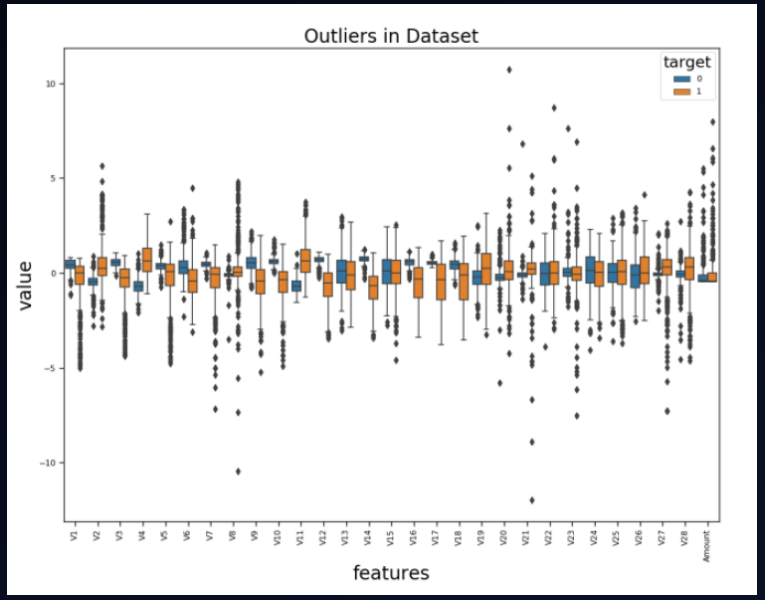
- It is reasonable to use Area under ROC Curve (AUROC) as a general-purpose metric in classification algorithms.
- AUROC represents the probability of distinguishing the model performed.
- AUC stands for "Area under the ROC Curve – Receiver Operating Characteristics".
- The scope of this area is AUC.
- The larger the area covered, the better the machine learning models are in distinguishing the given classes.
- The ideal value for AUC is 1.

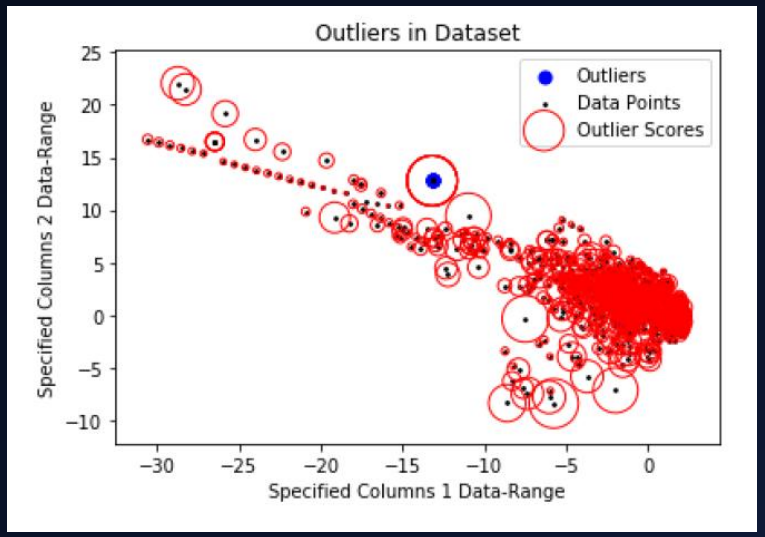


Reference for figure: https://medium.com/@bernatas/roc-e%C4%9Frisi-ve-e%C4%9Fri-alt%C4%B1nda-kalan-alan-auc-97b058e8e0cf

Outlier Detection

- Our data is skewed data
- Outliers detected and removed from the data set.
- Local Outlier Factor (LOF) method used for this process.
- LOF is an unsupervised Outlier Detection method.
- In order to calculate whether point *x* is outlier or inlier, the LOF value of that point is checked.
- If LOF(X) is greater than compare value (generally selected value 1) that point accepted as outlier.

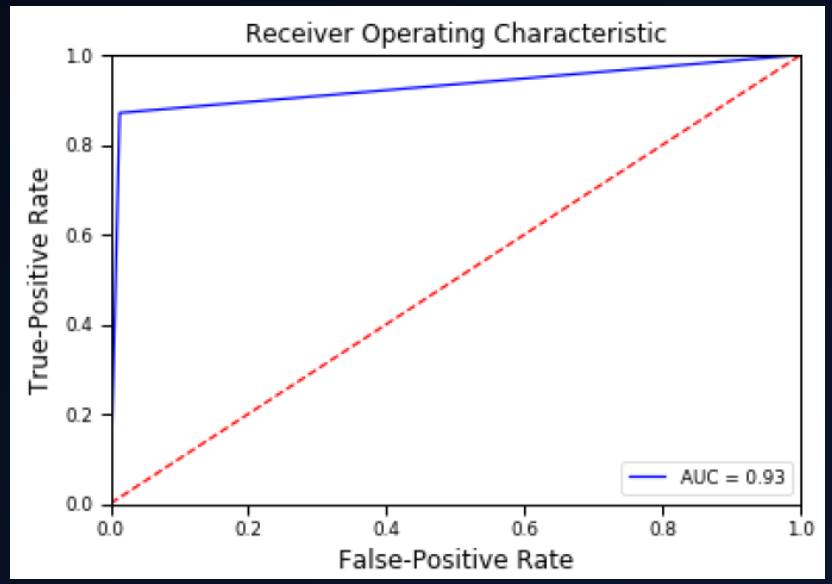




KNN

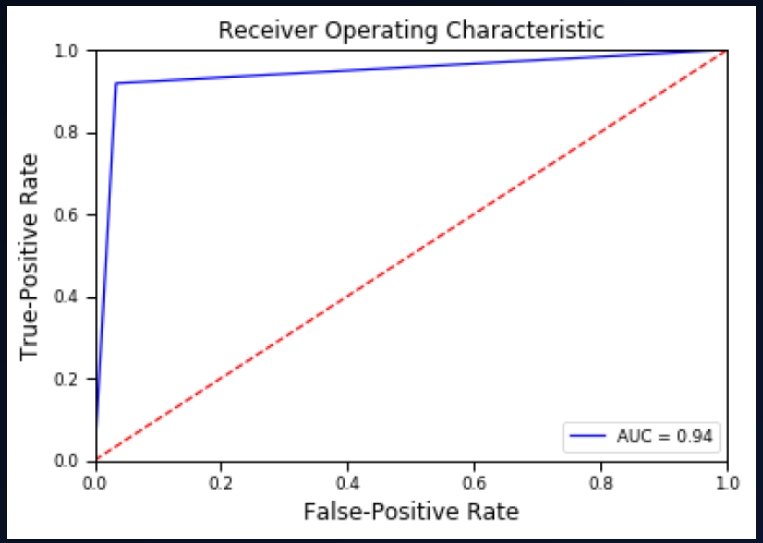
The main reasons for choosing the KNN algorithm

- We can list the training process is fast easy to implement
- Easy to tune because it has only k and distance parameters.
- It is sensitive against outliers
- Not very suitable for big data
- If there are too many features in the dataset it can be troublesome.



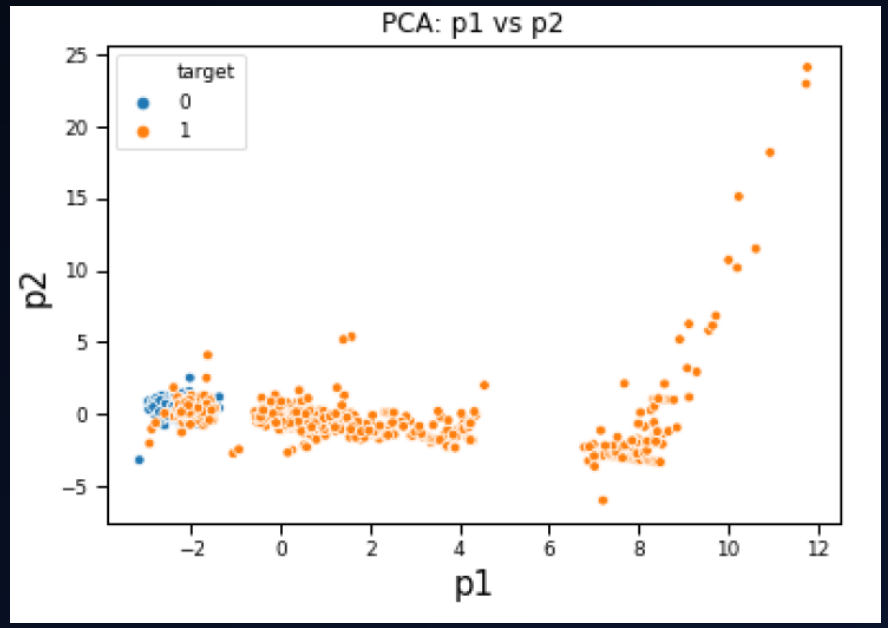
KNN with Grid Search

- The optimum parameters found in the KNN algorithm to achieve the higher accuracy
- For this, grid search method used.
- While finding the optimum parameters, the parameters that we tune;
  - n value to perform the classification process by looking at how many neighbors,
  - weight value that can take uniform and distance values
  - p value is the distance calculation method to be used when measuring distance between neighbors.



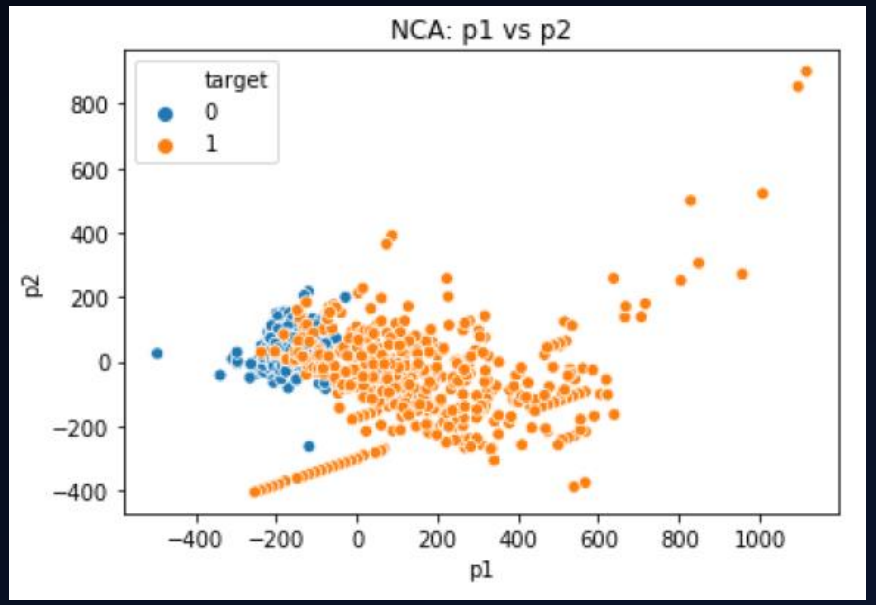
PCA (Principle Component Analysis)

- The dataset standardized again before applying PCA.
- Since PCA is an unsupervised algorithm, we have scaled all input data
- We did not split the dataset into train and test set.
- Based on the correlation matrix, the dataset reduced into two dimensions and its performance observed.
- For each point in the reduced dataset, the classification algorithm applied in the mesh grid created.
- We implemented the KNN algorithm again to the reduced data obtained after the dimension reduction process with PCA

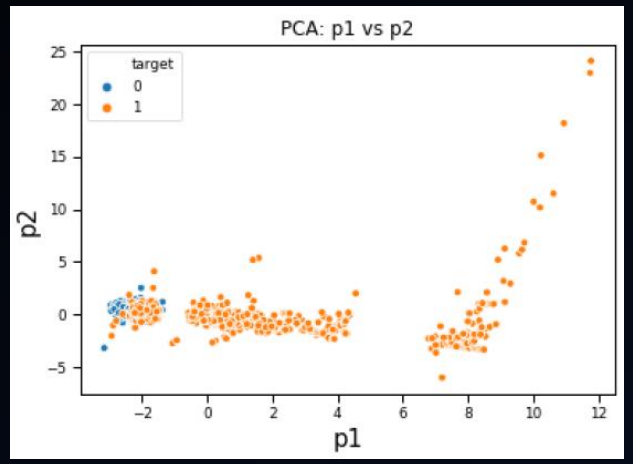
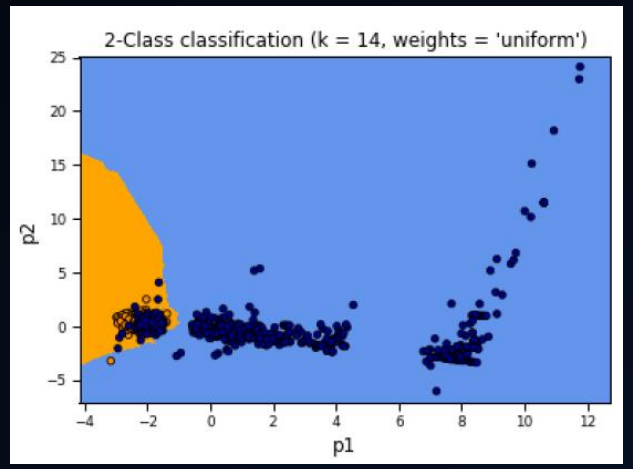
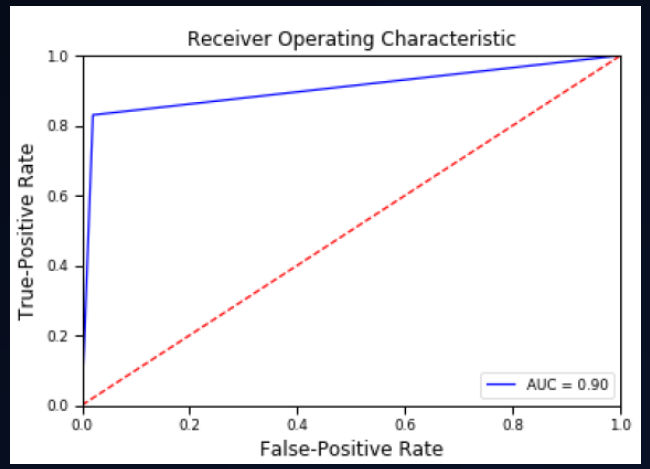
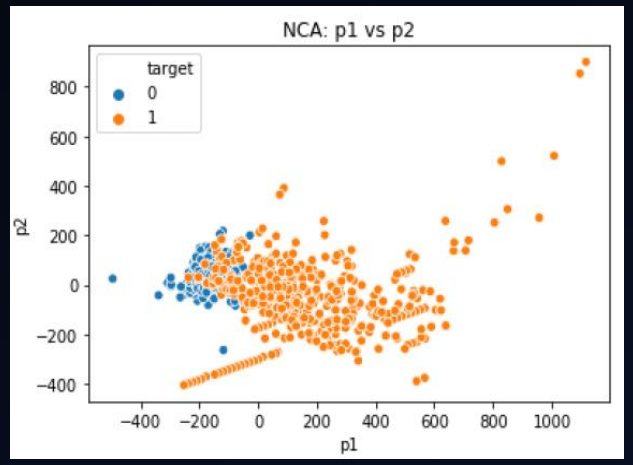
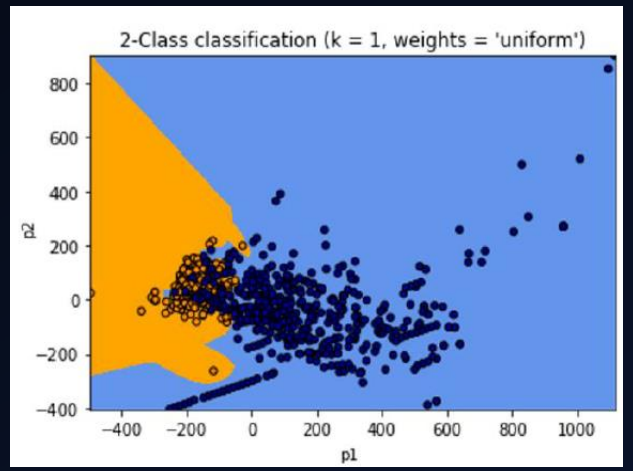
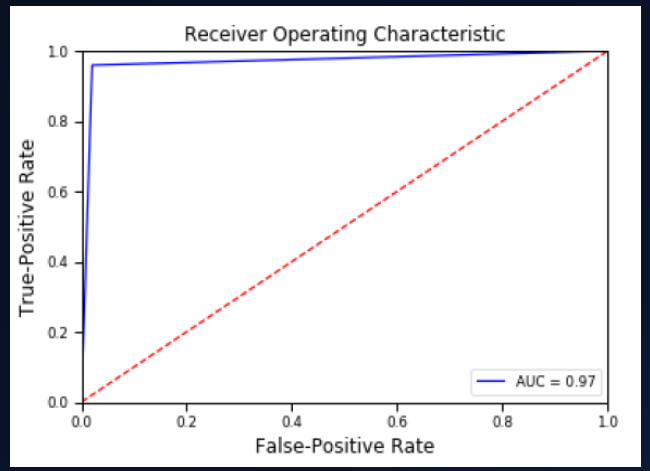


NCA (Neighborhood Component Analysis)

- In supervised learning, NCA is more successful than standard PCA methods
- For the optimization criterion, NCA takes advantage of stochastic neighbor assignments rather than simply using k nearest neighbors.
- Unlike PCA, NCA is not an unsupervised learning algorithm.
- It needs class information when performing fit operations.
- The p1 and p2 values in the x and y axes shown in the indicate the newly formed NCA principal components.



Evaluation

	Dimensionality Reduction	Classification	AUROC
PCA			
NCA			

Conclusion & Future Works

- Our main motivation for realizing this project was to identify transactions that could be described as credit card fraud in the light of historical data.
- Most of the machine learning algorithms applied in such studies yield over 90% accuracy results.

Results Comparison	Method			
	KNN	KNN with Grid Search	PCA	NCA
Accuracy	0.92	0.94	0.90	0.96
AUC	0.93	0.94	0.90	0.97

European Dataset Result

Future Works

- In this work, we presented a new method with higher accuracy than studies used KNN classification algorithms done to date.
- We obtained 0.97 AUC value
- The implemented method tested on different datasets which consisting of a small number of data, showed that it is possible to obtain high detection accuracy and low false negative.
- In future studies, we will be focused on optimization processes that will completely eliminate the error rate
- By working with different datasets, the performance of the applied method independent from the dataset will be observed.

THANK YOU

Q&A