DETECTING OUTLIERS USING MACHINE LEARNING

PRESENTED BY: REEM OMER

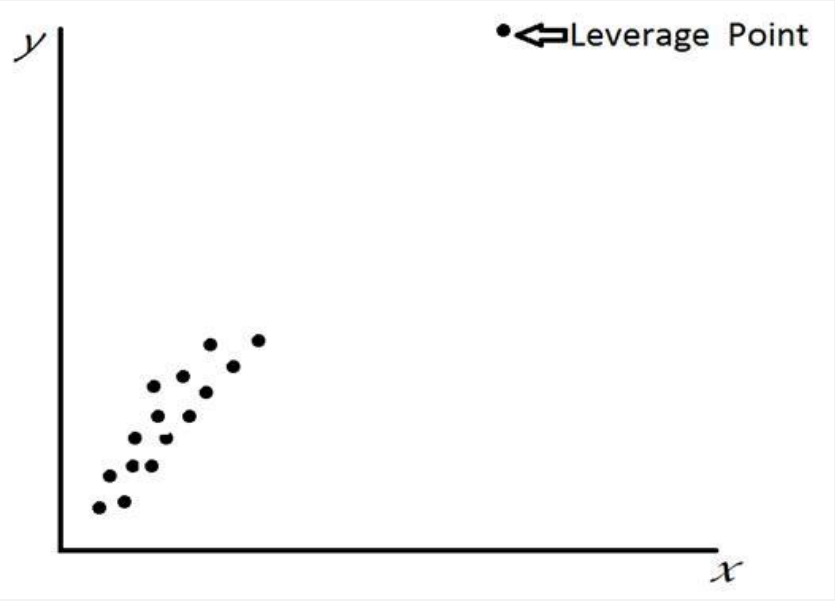
DEFINING OUTLIER

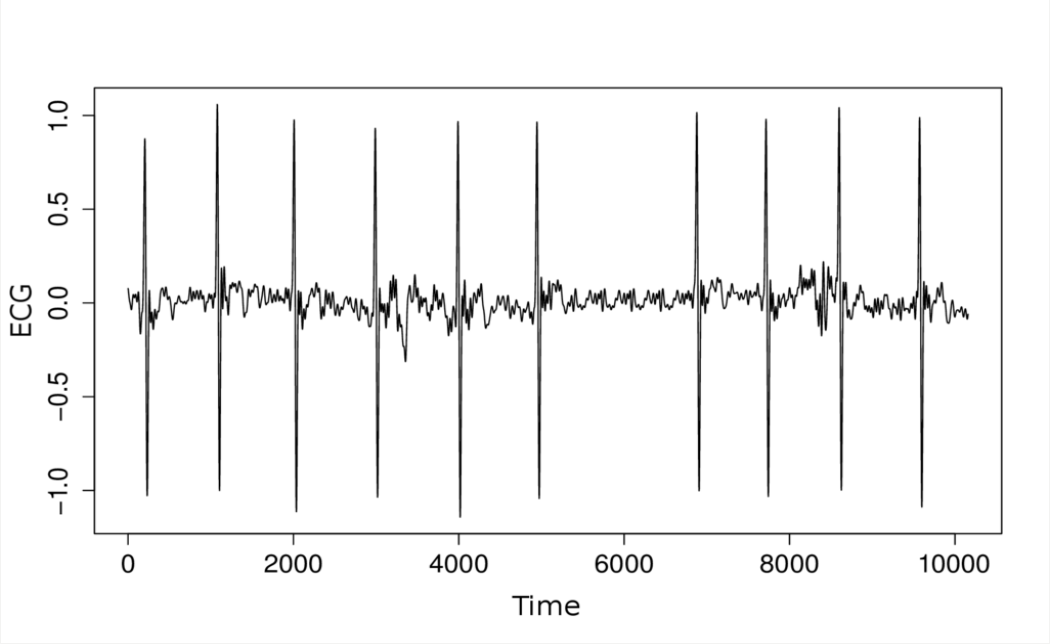
- An outlier is an observation (or subset of observations) which appears to be inconsistent with the dataset.
- Outliers were often treated as errors.

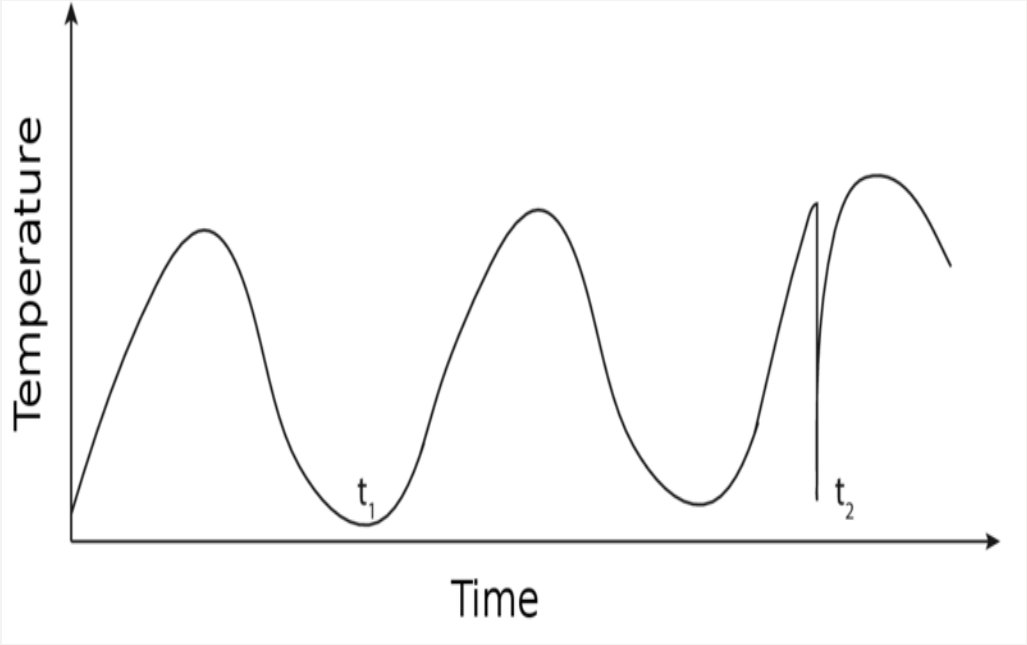
- Outliers indicate something significant and as such may be useful.

TYPES OF OUTLIERS

- Point Outliers
- Contextual Outliers
- Collective Outliers







OUTLIER DETECTION

- Outlier detection is the process of identifying the unusual observations in the datasets.
- **Why to detect outliers?**

Mobile Phone Fraud Detection

Medical and Public Health Outlier Detection

Intrusion Detection

Fraud Detection

Insider Trading Detection

Sensor Networks

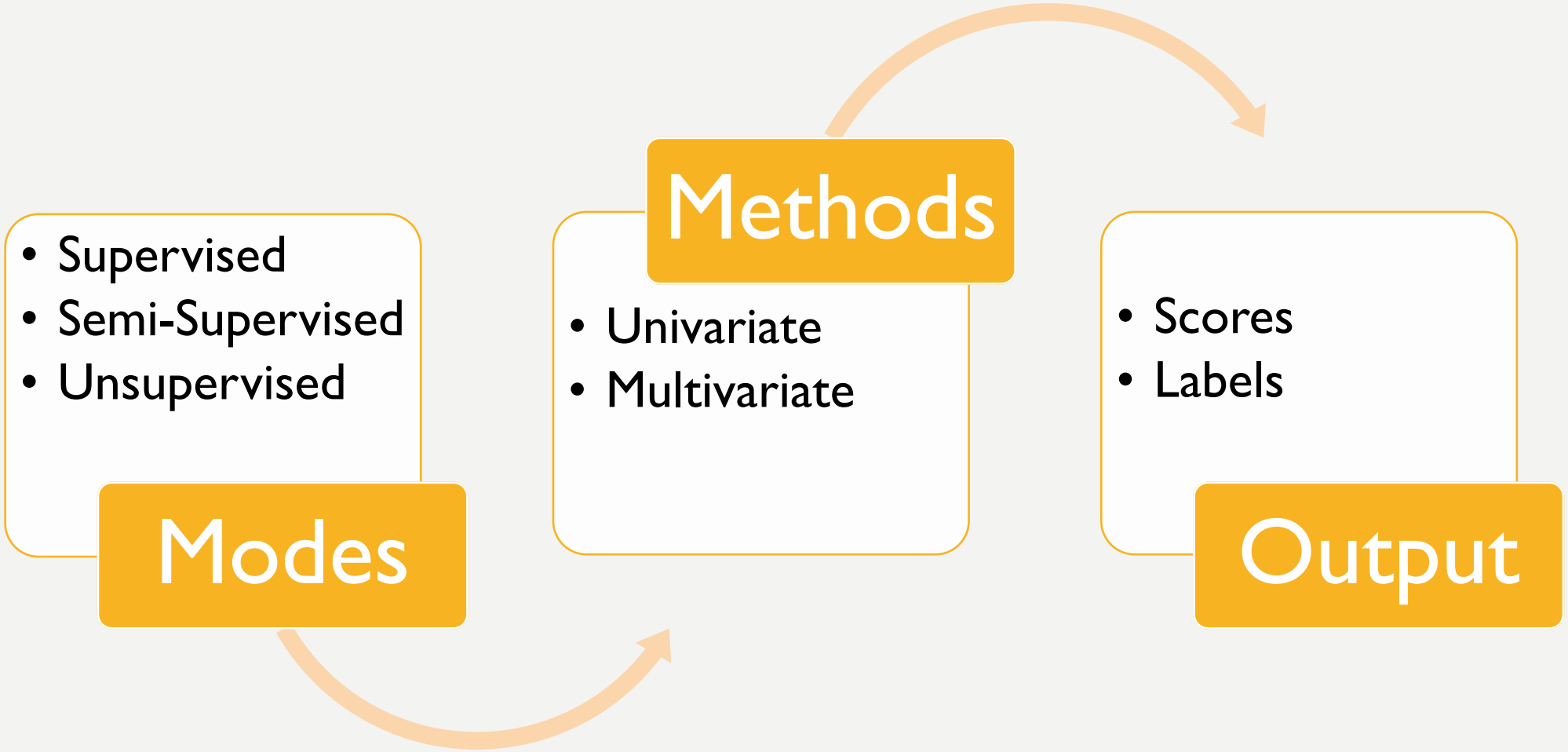
Image Processing

Insurance Claim Fraud Detection

Outlier Detection in Text Data

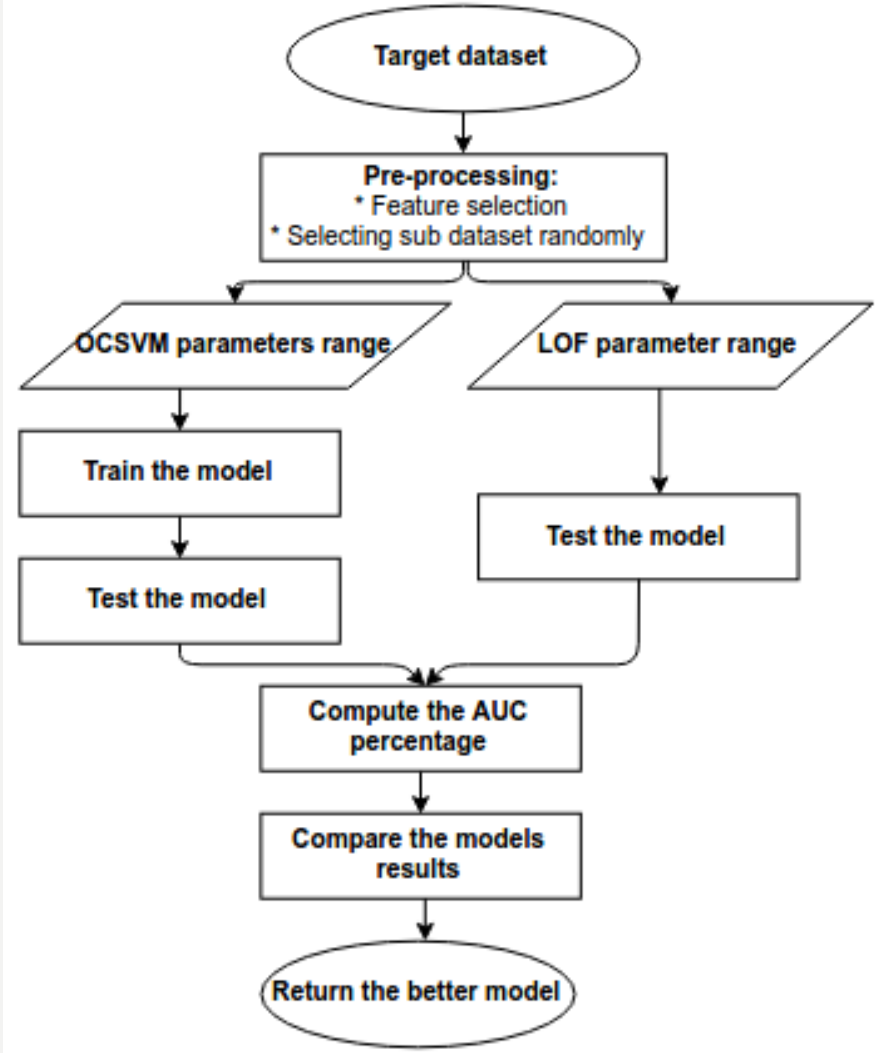
Industrial Damage Detection

HOW TO DETECT OUTLIERS?



EXPERIMENT

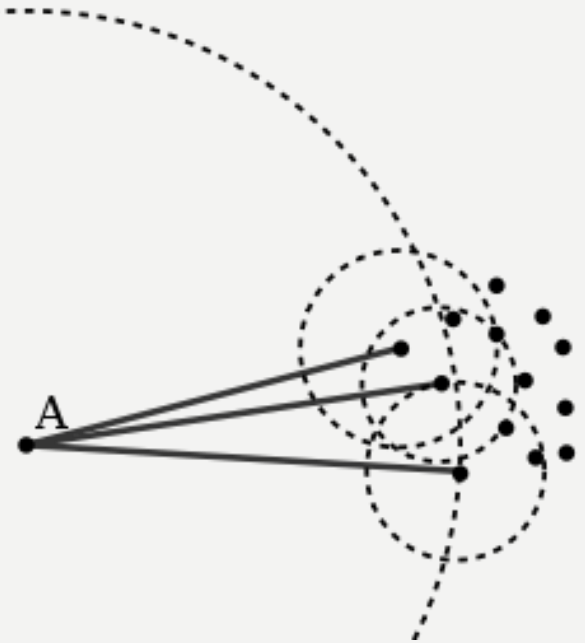
- Which one of Local Outlier Factor and One Class Support Vector Machine is better in detecting outliers in a highly unbalanced dataset?



LOCAL OUTLIER FACTOR

- The idea of the LOF method is to assign a probability for each object/observation of being an outlier.

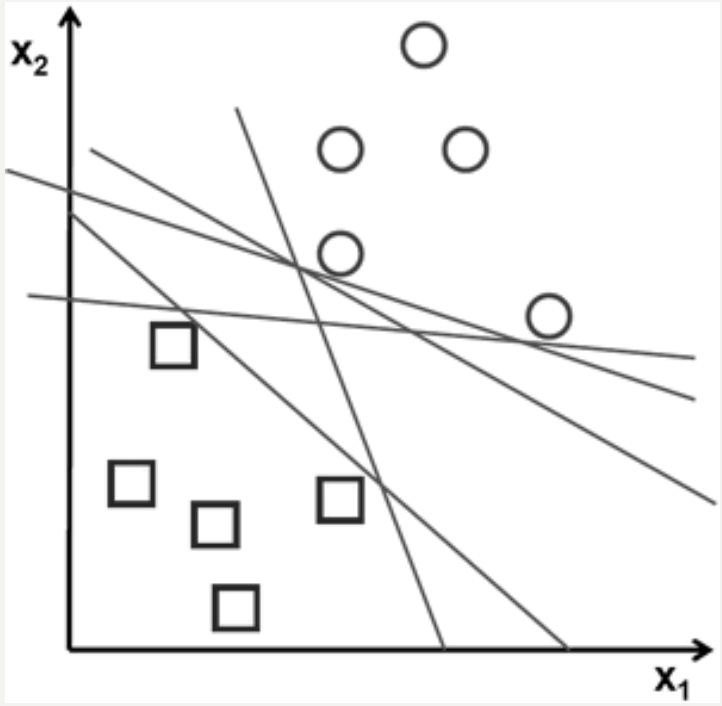
- Point A has a much lower density than its neighbours.

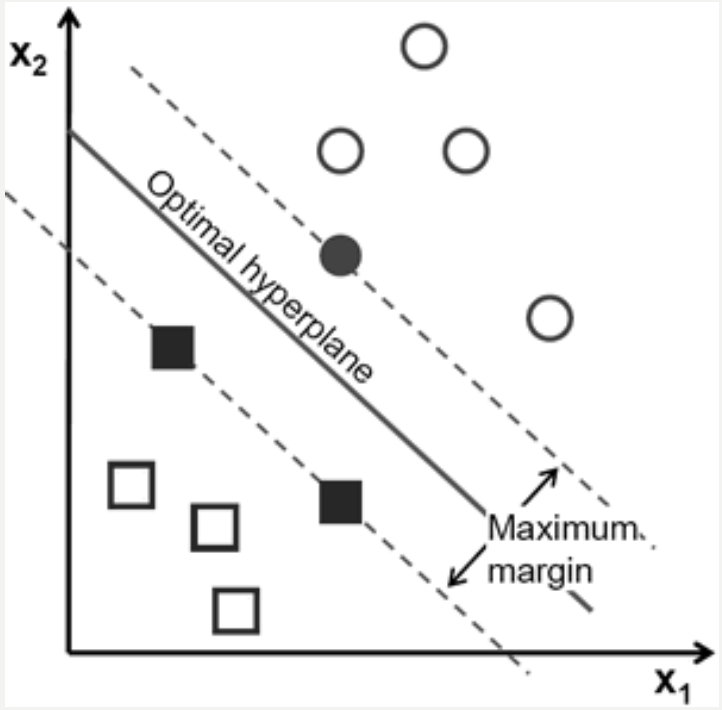


- In LOF we have to define the k nearest neighbours with respect to each point in the dataset.

SUPPORT VECTOR MACHINES

- The main idea of SVMs is to find the optimal hyperplane in feature space that best separates classes.





- OCSVM requires two parameters: γ defines how far the influence of an observation reaches, and C is the proportion of outliers expected in the data.

INPUT DATA

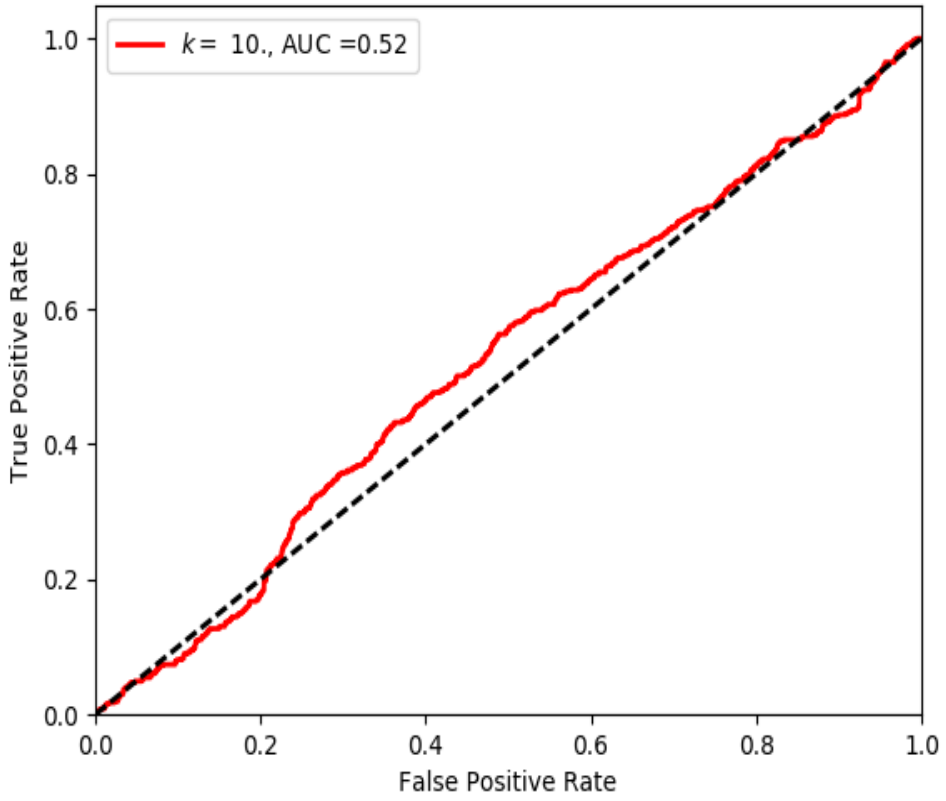
- A credit card transactions dataset from Kaggle data science repository is used to evaluate the performance of the algorithms.

Algorithm	Training	Testing
LOF	40492	40492
OCSVM	30000	10492

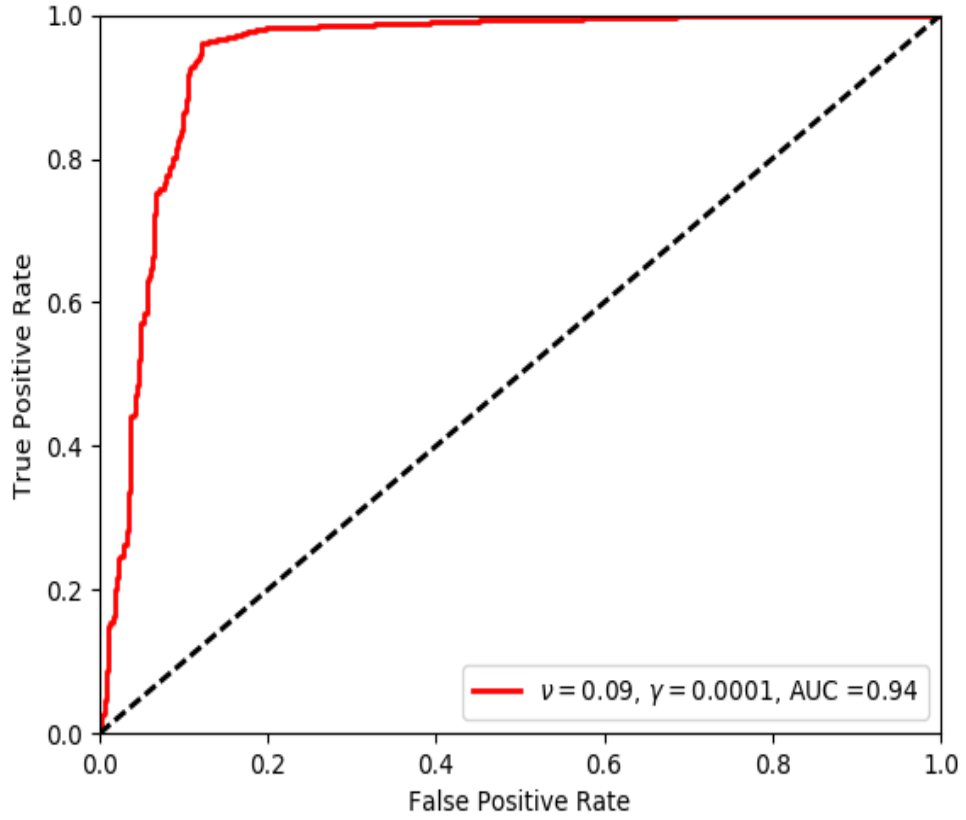
EVALUATION METRICS

		Predicted Observations	
		+	-
Actual Observations	+	True Positive	False Negative
	-	False Positive	True Negative

ANALYSIS AND RESULTS



- LOF when parameter value k=10



- OCSVM parameters C=0.09 and γ=0.0001

THANK YOU!

QUESTIONS ARE WELCOME.