The main challenges currently facing researchers involved in detecting fraud in mobile payment transactions include:

* extreme class imbalance (only a small proportion of customers have fraudulent intentions),
* changing patterns of fraud over time (fraudsters are always looking for new ways to bypass systems and commit crimes),
* inadequate selection of performance metrics.

The second challenge usually leads to a decrease in the performance and efficiency of the detection model. Therefore, machine learning models must be constantly updated, otherwise they will not meet their objectives. Regarding the last challenge, in some cases the providers of mobile payment systems should prefer a higher false positive rate in exchange for a lower false negative rate and vice versa. But how to choose the right ratio between these two errors remains a challenging area in the feld of fraud detection in mobile payment transactions.

A relatively high detection accuracy was reported in earlier research by using both traditional supervised learning methods and deep learning-based methods. However, a major problem with this kind of application is the extreme class imbalance of transactions, with a considerable dominance of legitimate transactions in the data. This in turn leads to a poor classification performance on the minority class of fraudulent transactions. To address this issue, two approaches have been utilized.

* The first approach relies on under-sampling methods used to generate a balanced dataset. The main limitation of this approach is the loss of potentially important information stored in discarded legitimate transactions, which can reduce detection accuracy.
* Alternatively, an attempt has been made to isolate fraudulent transactions in an unsupervised fashion, inspired by outlier detection methods. Nevertheless, a comprehensive evaluation of machine learning methods is not yet available in the literature. Moreover, little is known about how the two approaches can be integrated to improve the detection performance.

To overcome the above problems, here we propose to enhance the performance XGBoost, a state-of-the-art machine learning method, by including a data sampling component addressing the issue of extreme class imbalance of mobile payment transactions.

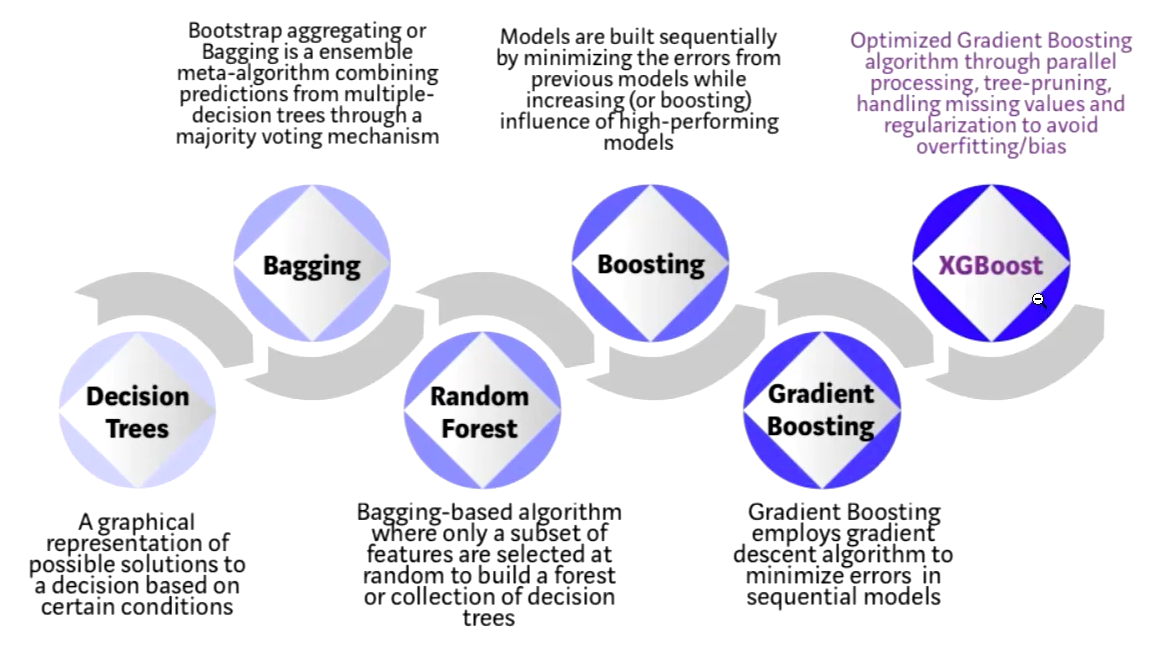
Moreover, several other challenges have been identified that make it the difficult to detect outliers (fraud) in the financial domain. First, efficient general purpose outlier detection methods are lacking because an outlier detection method in one fraud domain may not be appropriate for other scenarios, as legitimate and fraudulent behaviour is different from domain to domain. Second, unsupervised learning is preferred as sufficient labelled data for building models are rarely available. Third, legitimate behaviour may change over time, and fraudsters try to make their activities look legitimate. To take advantages of both supervised machine learning and outlier detection methods, for the first time, we propose ***a semi-supervised ensemble fraud detection model combining unsupervised outlier detection and supervised XGBoost methods*** that exploit all transactions contained in a large, highly imbalanced mobile payment transaction dataset.

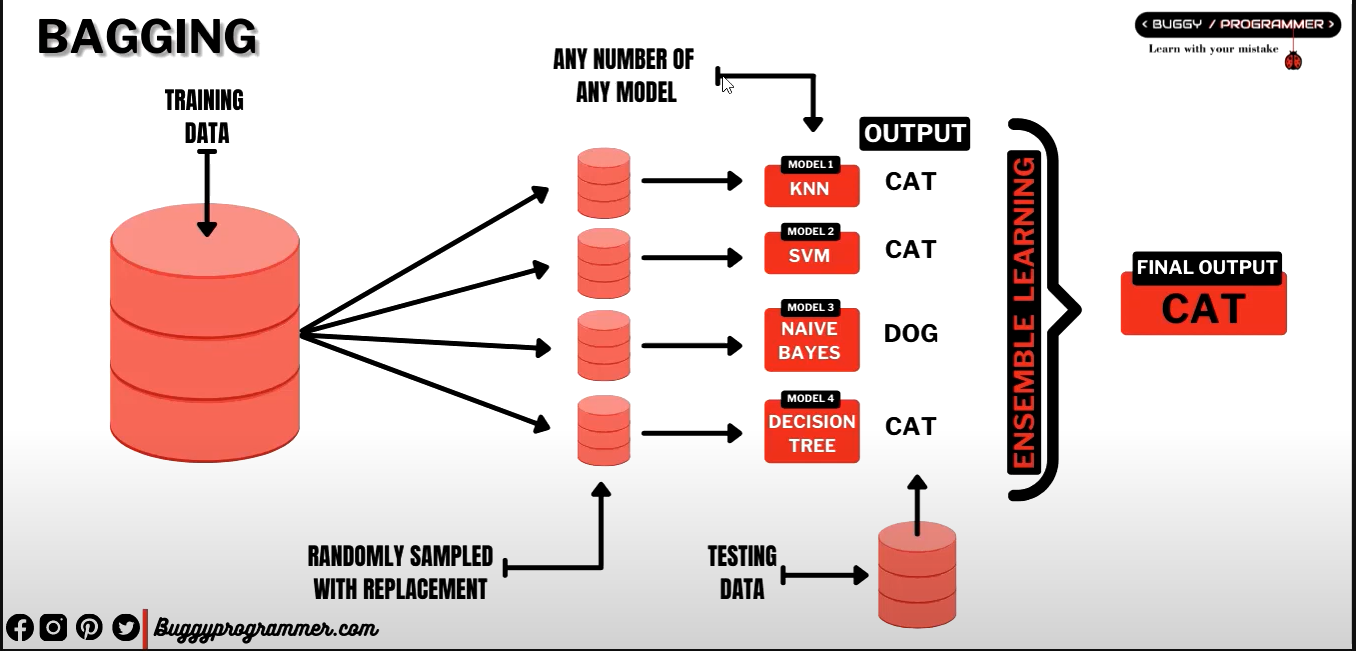
**First, *the eXtreme Gradient boosting (XGBoost) method, augmented with random under-sampling****,* is introduced to leverage both the supervised learning capability and robustness of XGBoost, a state-of-the-art machine learning method, and the data sampling component to overcome the class imbalance problem inherent in mobile payment transaction data*.* ***The second model exploits the extreme gradient boosting outlier detection (XGBOD) method, a semisupervised algorithm that improves the performance of the XGBoost method*** on highly imbalanced mobile payment transaction data by introducing outlier scores obtained from multiple unsupervised outlier detection methods.

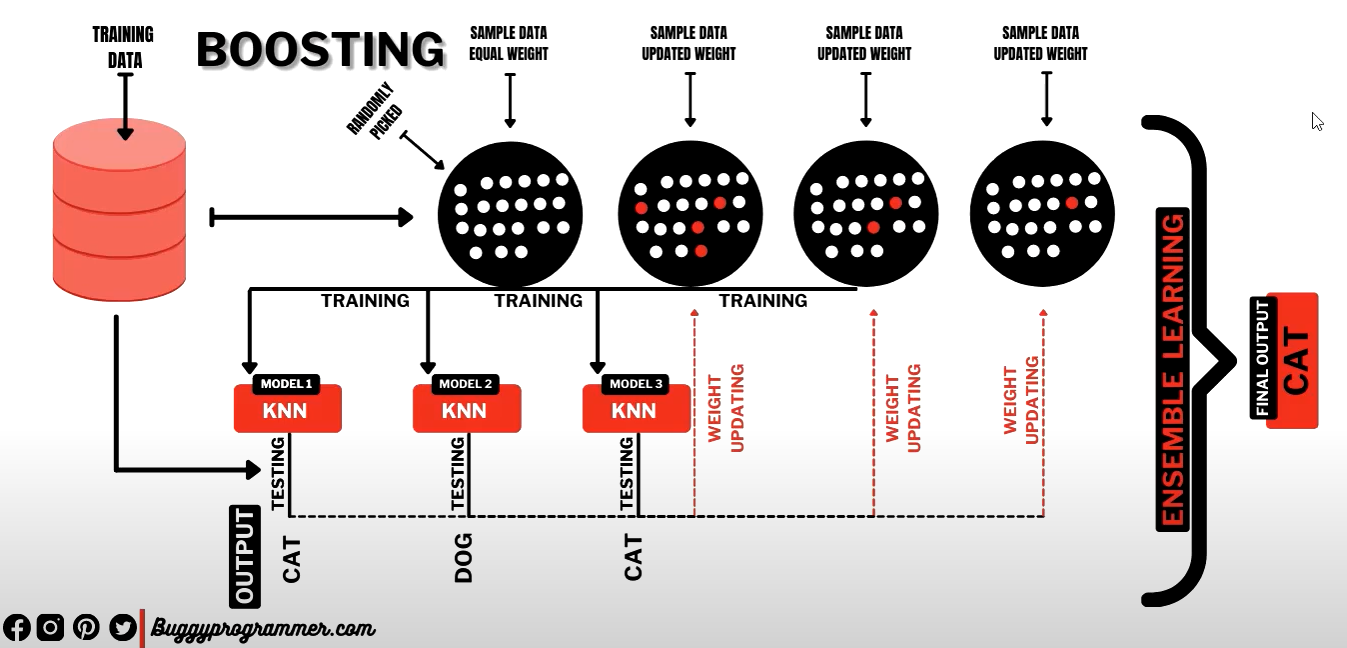
Using the benchmark PaySim dataset of more than 6 million mobile payment transactions, we demonstrate that the proposed fraud detection framework not only outperforms state-of-the-art fraud detection methods in terms of detection accuracy but also generates substantial financial savings to the providers of mobile payment systems.

From the data-level methods, over-sampling methods create artificial instances in the minority class to balance the training data. However, this can lead to problems of overfitting and overgeneralization as instances of the majority class are ignored. Moreover, given the gradual increase in data on financial fraud, undersampling methods should be a better choice than their oversampling counterpart.

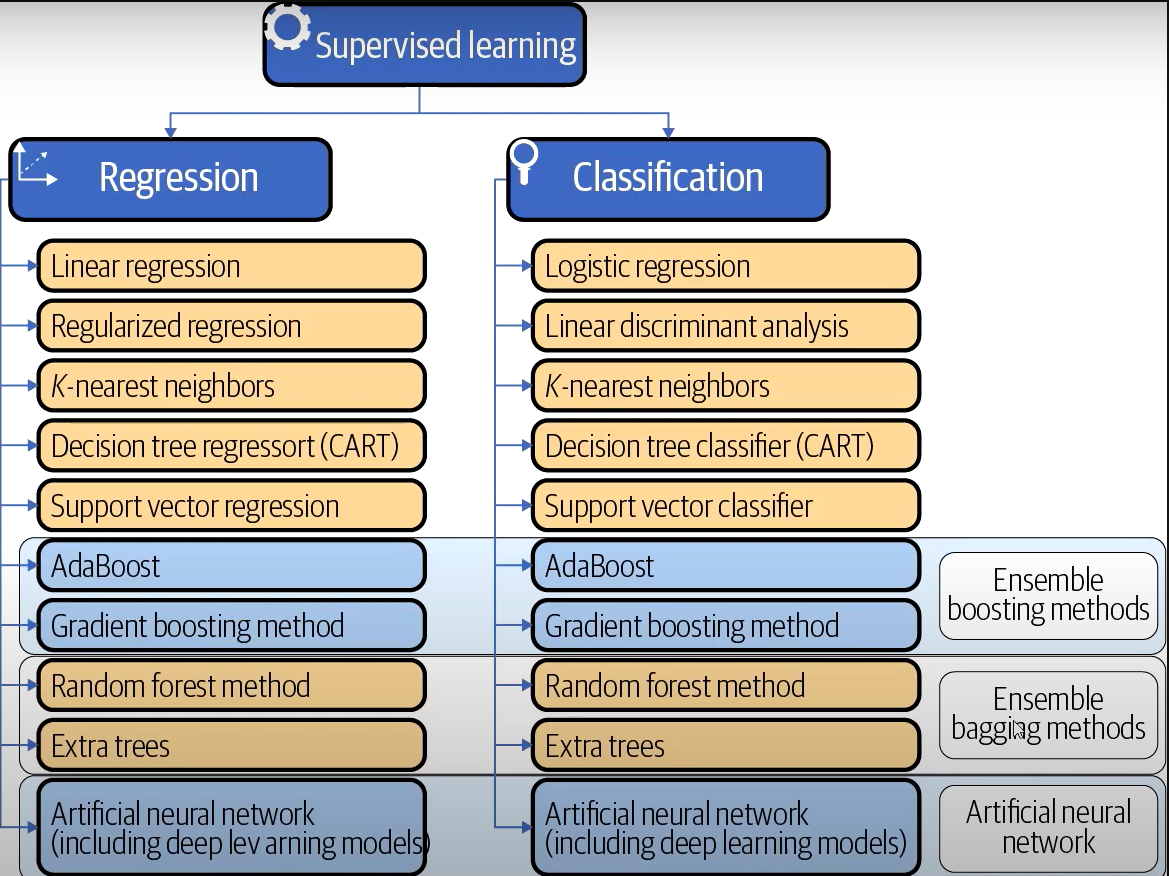
**RUS (Randomized Under-Sampling)** is a non-heuristic method that randomly selects a data subset from the majority class, which is computationally effective and enables sampling heterogeneous data.



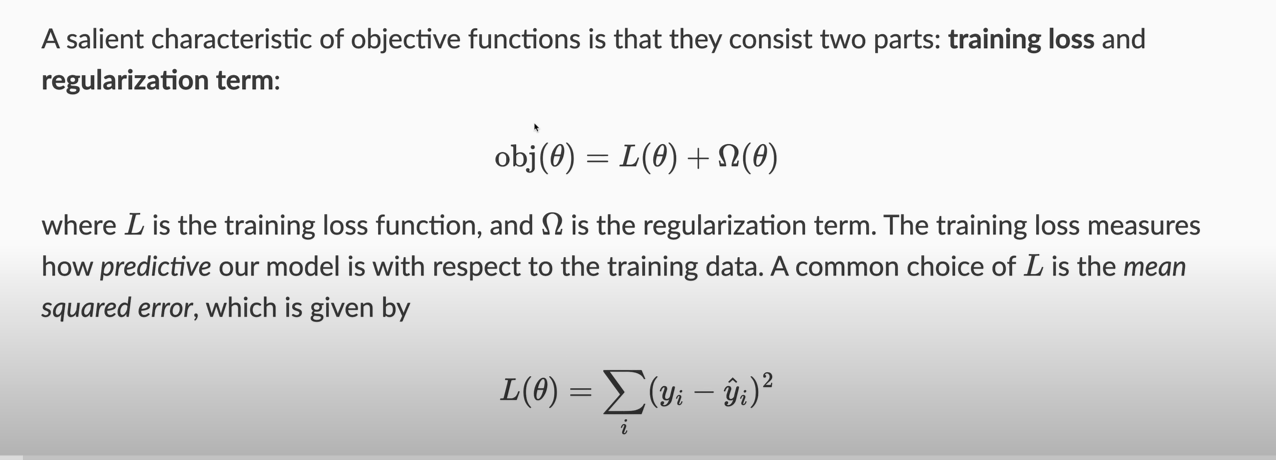


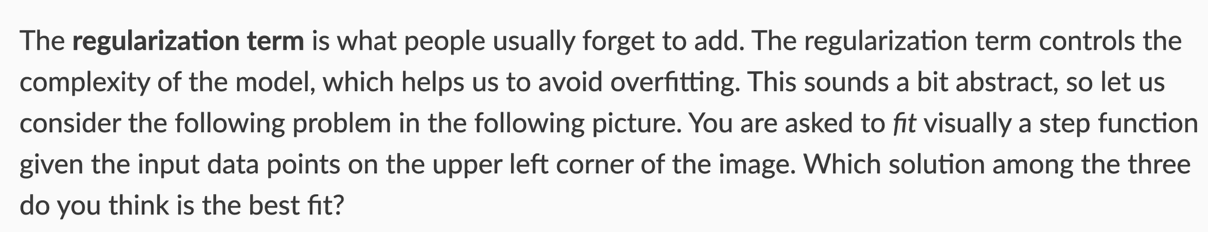


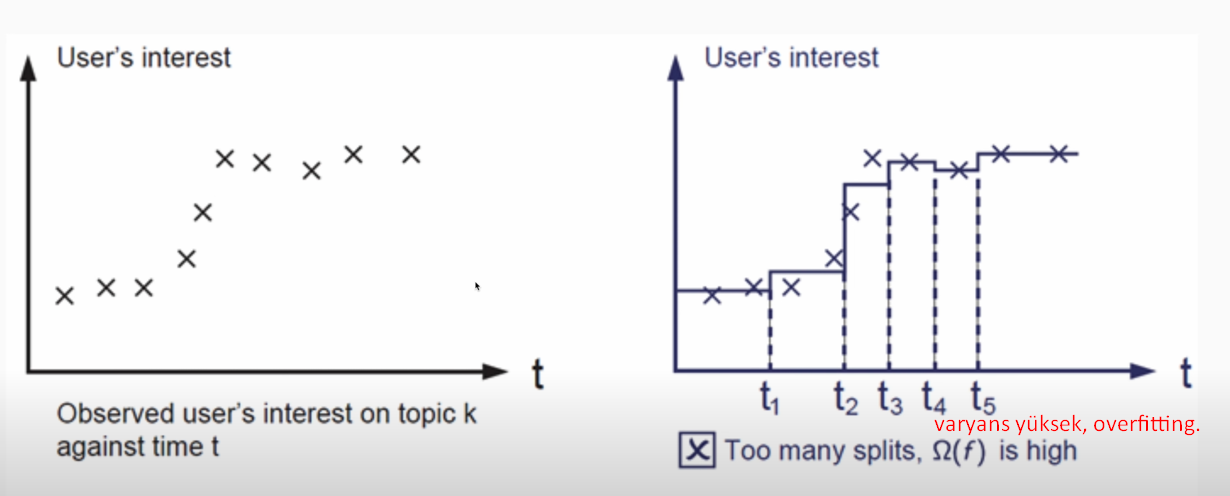
Bayes algoritmasının sonucunda biz bir başarı skoru alıyoruz. 100 gözlemden 60 tanesi doğru, 40 tanesi yanlış sınıflandırılmış olsun. Buradaki başarı oranımız %40. Bu tek bir model kullandığımızda aldığımız sonuçtur. Bayes’in yanlış sınıflandırdığı o 40 gözlemi farklı bir model doğru sınıflandırıyor ise ne olacak (KNN)? Eğitim data setini parçalara ayırarak oluşturduğumuz farklı model yapıları içerisindeki doğru ve hatalı tahmin yapılarına göre bir ağırlıklandırma işlemini gerçekleştirdiğimizde buna boosting diyoruz.

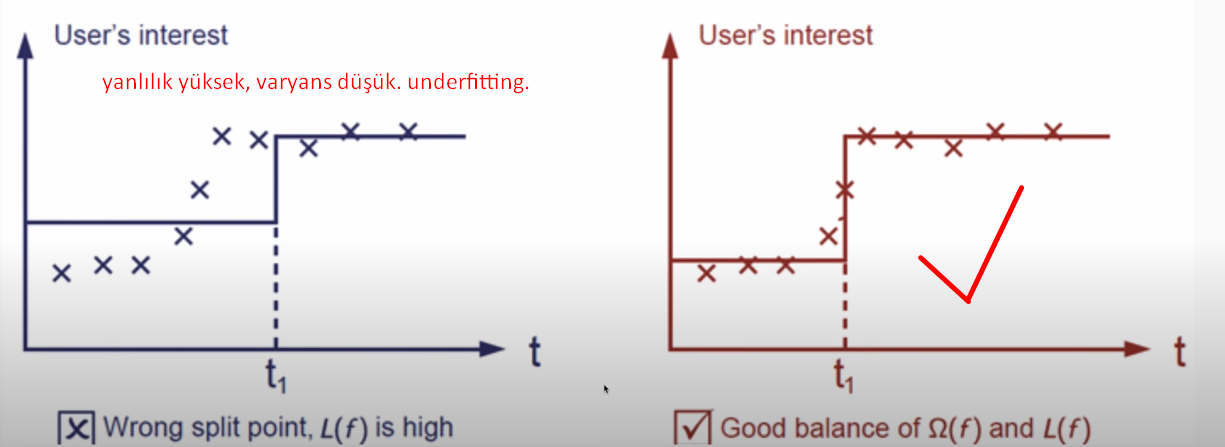


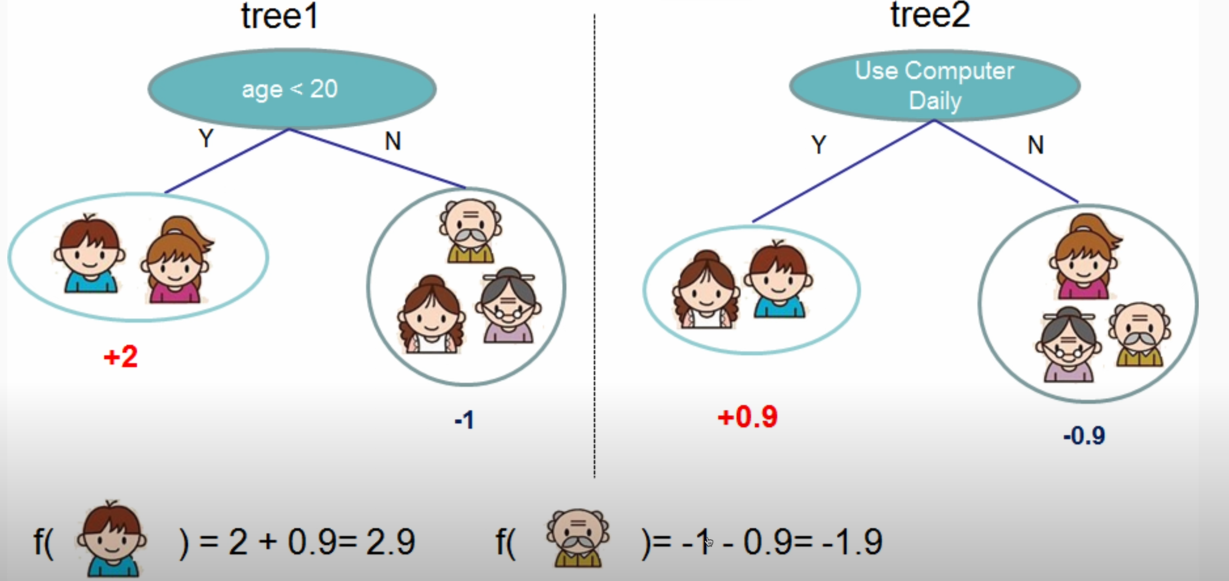
# XGBoost

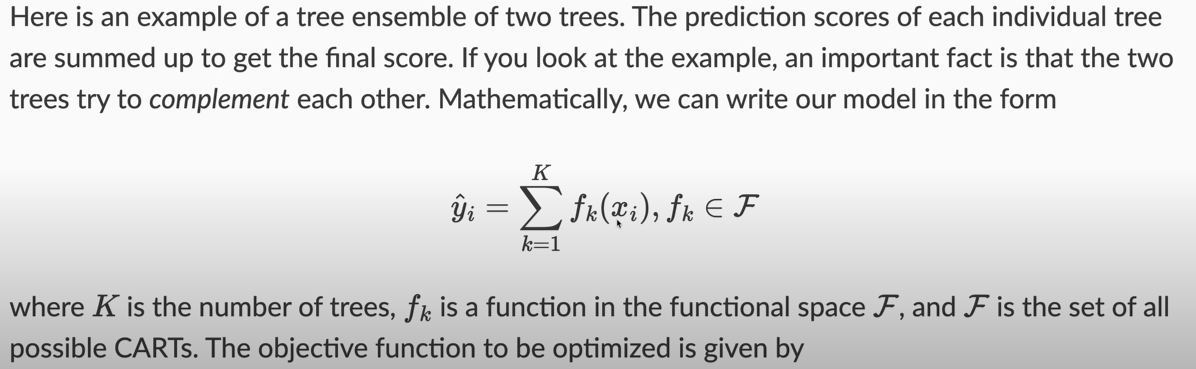


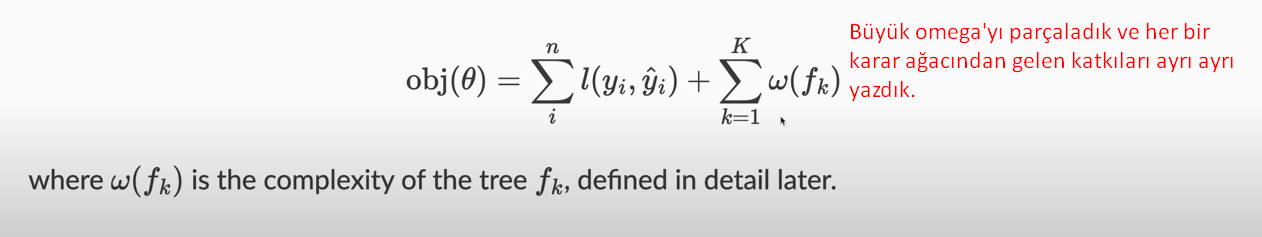




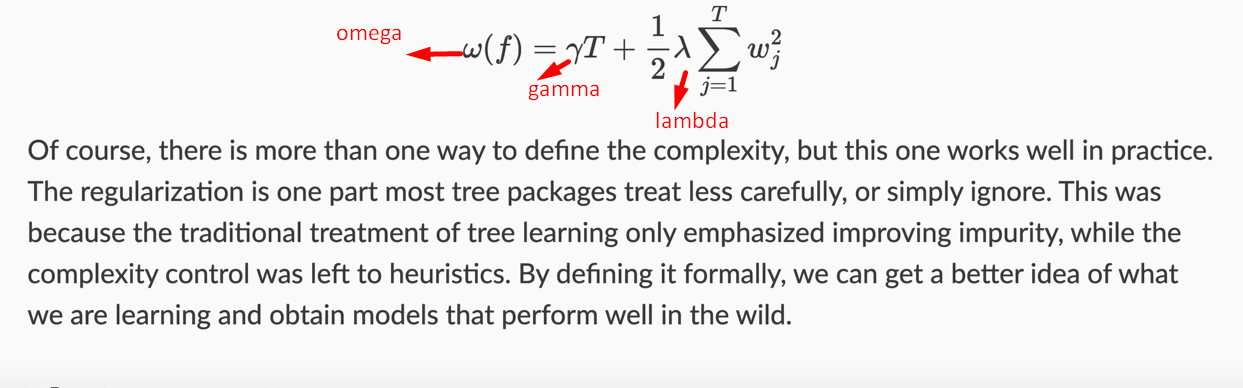


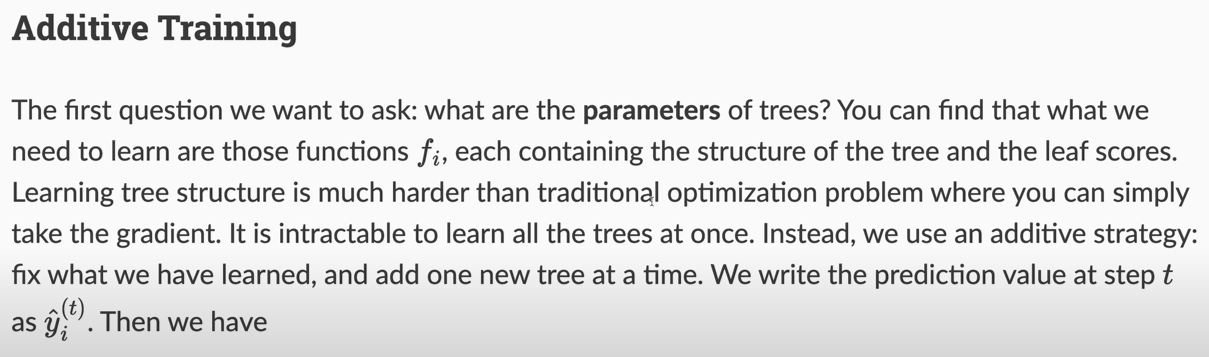
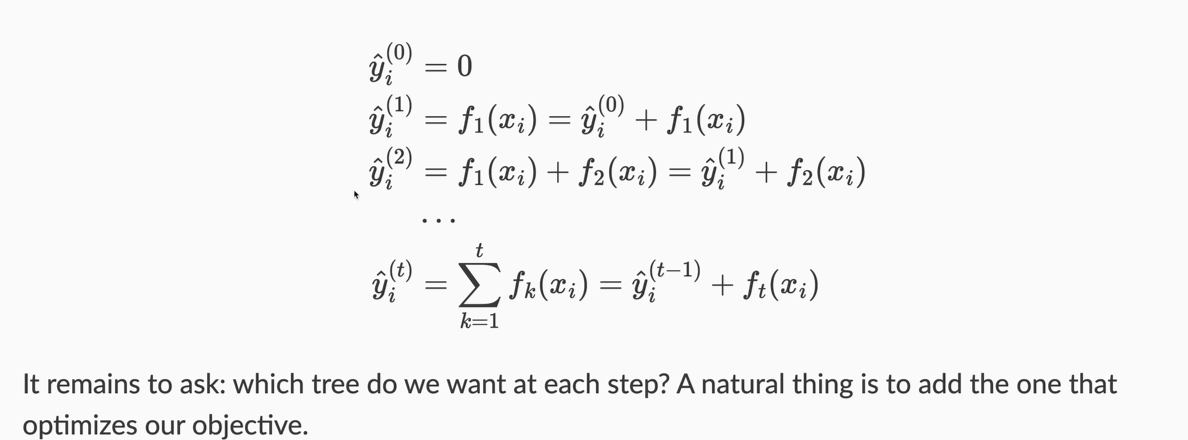


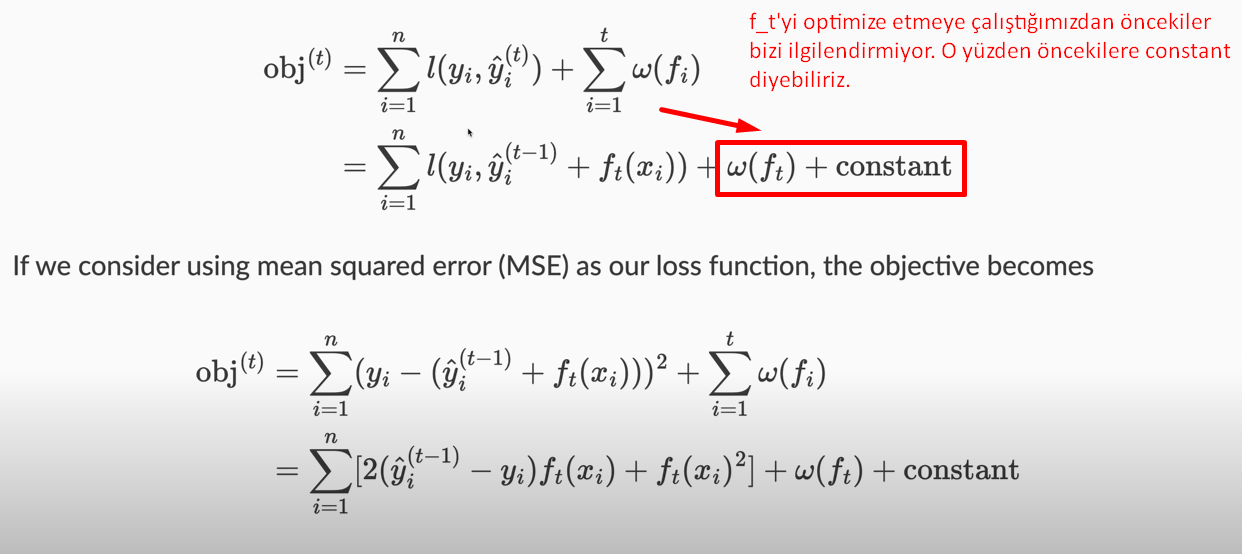


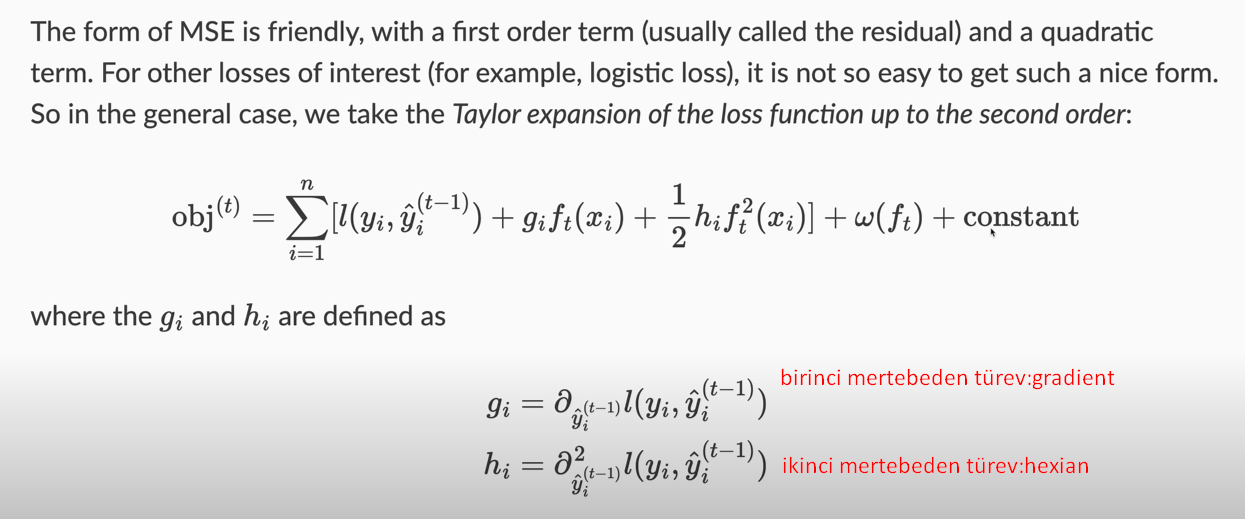


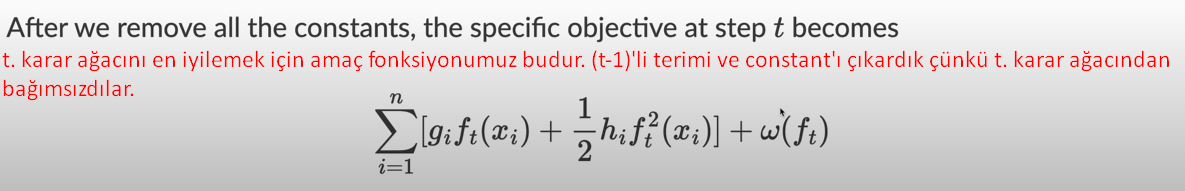
ω () karar ağacındaki yaprak sayısını yani karmaşıklığı ifade eder.

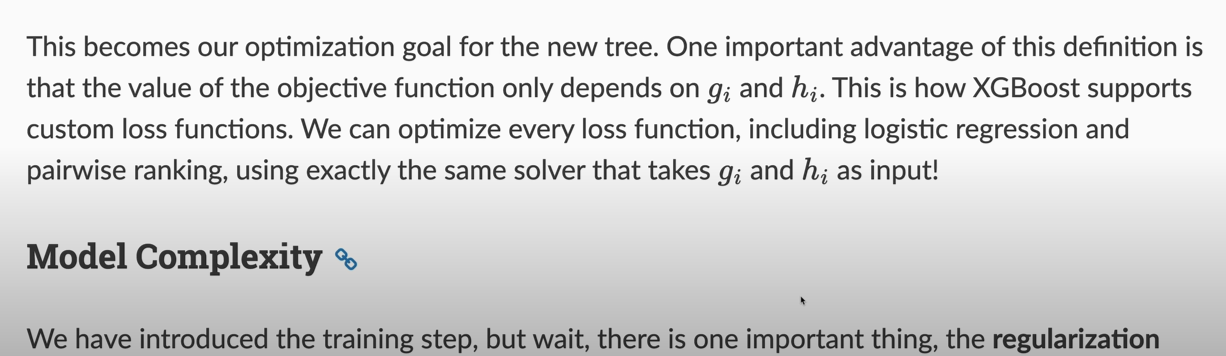


  
Amacımız yaprak sayısını yani ()yi kontrol etmek.





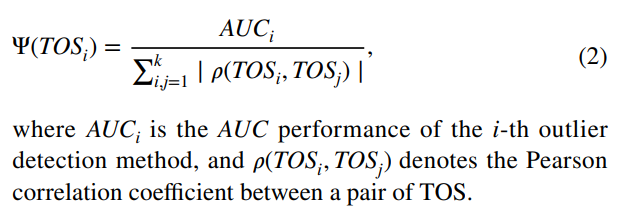




# XGBOD

The XGBOD method is a semisupervised ensemble algorithm integrating multiple unsupervised outlier detection algorithms and an XGBoost classifier. First, unsupervised methods are used to obtain data representations in terms of **transformed outlier scores (TOS).** Second, a feature selection method is used to reduce the TOS feature space so that only relevant TOS are retained. Then, the outlier score matrix is combined with the original features to produce a combined feature space. An improved feature space is thus generated, and the XGBoost classifer is used in this feature space to produce the final outlier scores for each mobile payment transaction. The advantage of this approach is its good predictive ability, which is due to its robustness to overfitting and data imbalance.

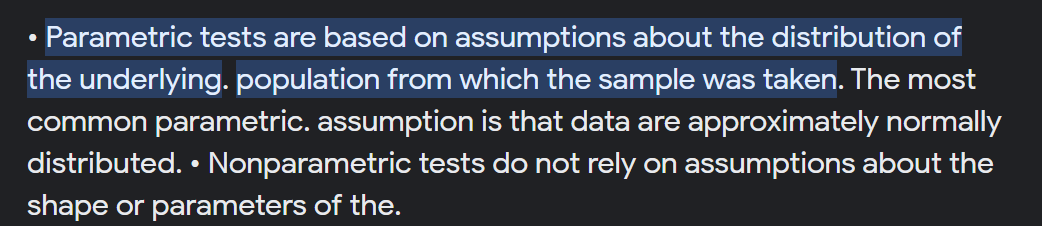
In the proposed XGBOD-based fraud detection model, a variety of unsupervised outlier detection methods (presented in Section 3.2.2) are used to produce the TOS features. To maintain the balance between their diversity and accuracy, the balance selection algorithm (Zhao & Hryniewicki, 2018) is used to perform TOS selection. This algorithm applies a discounted accuracy function Ψ(TOSi) to pick the subset of p most relevant TOS. The function is defined as follows:



AUC – area under ROC curve

# Supervised Learning Methods for Imbalanced Data

## k‑nearest Neighbour Classifier



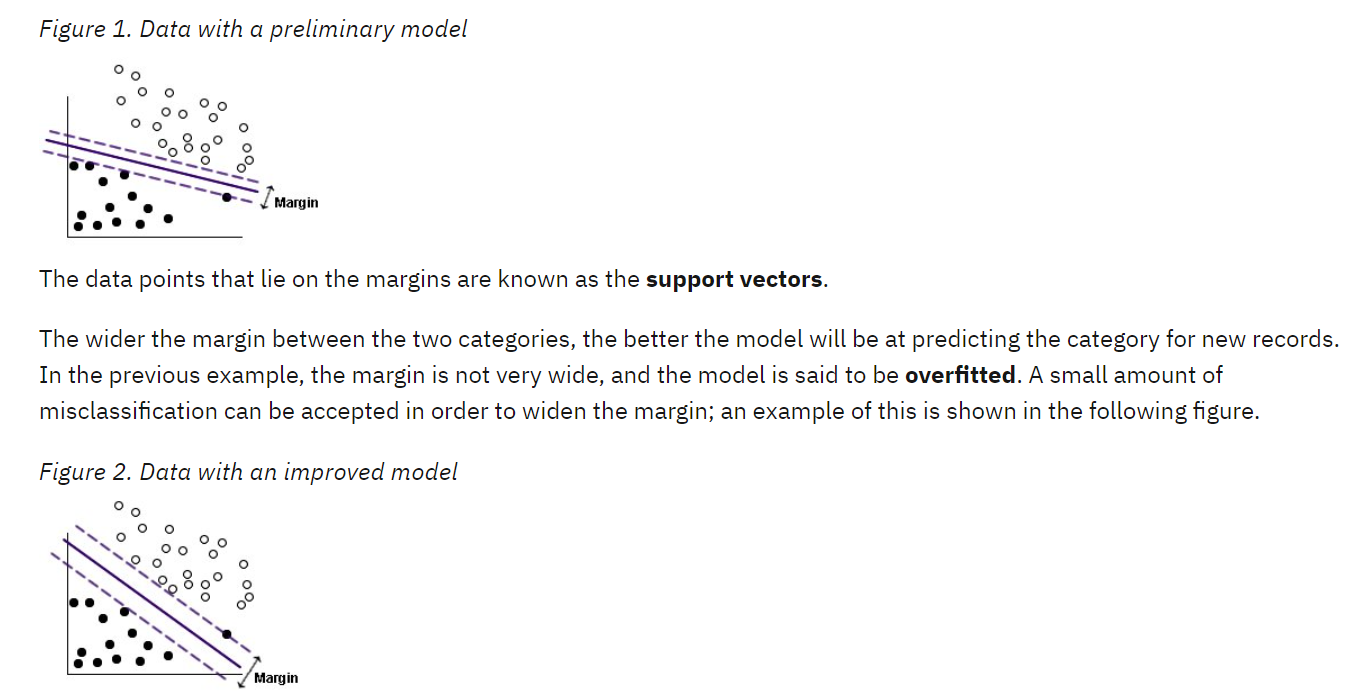
The k-nearest neighbour (k-NN) method is an instance-based non-parametric classifier that uses training instances for comparison purpose. An instance is classified considering its k most-similar instances (typically in terms of Euclidean distance) using a majority vote. This simple approach proved to be accurate in a comparative analysis of machine learning methods for highly imbalanced credit card fraud detection (Awoyemi et al., 2017). In financial fraud detection, it is assumed that fraud instances are far from the samples of the legitimate class. Therefore, k-NN can be effectively used even in unsupervised outlier detection mode.

## Support Vector Machine

SVM is a particularly effective classifier for financial fraud detection due to its capacity to deal with high-dimensional data. The SVM algorithm aims to find the optimal separating hyperplane that maximizes the margin between instances from different classes. The decision boundary is represented by a subset of the data known as support vectors. Finding the parameters of the hyperplane is an optimization problem that takes into consideration both, minimizing the training error and maximizing the margin. To handle nonlinear relationships in the data, kernel functions (e.g., linear, polynomial or radial basis functions) are employed to map the classiffication problem from the original feature space to a new feature space of higher dimension where linear separation is possible.

Support Vector Machine (SVM) is a robust classification and regression technique that maximizes the predictive accuracy of a model without overfitting the training data. SVM is particularly suited to analyzing data with very large numbers (for example, thousands) of predictor fields.

SVM has applications in many disciplines, including customer relationship management (CRM), facial and other image recognition, bioinformatics, text mining concept extraction, intrusion detection, protein structure prediction, and voice and speech recognition.



**In a case like this, the goal is to find the optimum balance between a wide margin and a small number of misclassified data points.** The kernel function has a regularization parameter (known as C) which controls the trade-off between these two values. You will probably need to experiment with different values of this and other kernel parameters in order to find the best model.

## Random Forest

Random Forest (RF) integrates multiple decision tree predictors trained independently on different data samples. This allows to generate a number of trees, ensuring that the generalization error converges to a certain limit. Another major advantage of RF is its non-differentiable decision boundary. In addition, random feature selection is used to split the nodes in each tree, making the RF classiffier more robust to noise. **The application of RF in financial fraud detection is particularly effective when the class distribution is imbalanced because its hierarchical structure enables learning patterns from both classes** (Nami & Shajari, 2018). These advantages explain the good performance of RF on financial fraud detection tasks (Zhou et al., 2018).

# Outlier Detection Methods

Outlier detection is typically conducted using unsupervised machine learning methods. The methods presented in this section are trained to represent the legitimate data using clusters of similar data observations. Then, an unseen instance is assigned a score that is compared to a threshold representing the decision boundary separating legitimate instances from outliers.

## Proximity‑Based Methods

An important advantage of proximity-based methods is their independence of the data distribution. In other words, no a priori knowledge about the data distribution is required. To detect outliers, proximity-based methods investigate the neighbourhood of each data instance.

For example, the local outlier factor (LOF) method (Breunig et al., 2000) uses the Euclidean distance between the data instance and its closest neighbour to obtain an outlier score. In the k-NN method (KNN) (Ramaswamy et al., 2000), a partition-based algorithm is first used to identify candidate partitions containing outliers, and then the distances of instances from these partitions are calculated to detect outliers. However, these methods usually do not scale well for high-dimensional data. To reduce the sensitivity of LOF to the curse of dimensionality, the cluster-based local outlier factor (CBLOF) method (He et al., 2003) replaces closest neighbours with closest clusters, and the angle-based outlier detection (ABOD) method (Kriegel et al., 2008) replaces distances with the angular radius and variance of each data vector. The histogram-based outlier detection (HBOS) method assumes independence of features to score instances in linear time and is thus computationally more efficient compared to nearest-neighbour-based methods. However, HBOS fails in detecting local outliers because the density estimation produced by histograms does not allow modelling local outliers.

## Linear Model-Based Methods

### SVM

Linear model-based methods rely on the construction of decision boundary separating instances in the legitimate class from the rest of the input data space. The one-class SVM (OCSVM) method (Schölkopf et al., 2000) constructs a separating hyperplane in high-dimensional space by minimizing the structural risk to capture regions of data belonging to the legitimate class. **To prevent overfitting, this method allows a certain percentage of data instances (regularization parameter) to fall outside the separation boundary.**

### The Minimum Covariance Determinant (MCD)

The minimum covariance determinant (MCD) method (Hardin & Rocke, 2004) combine a multivariate location and scale estimator with a robust clustering algorithm so that the determinant of the covariance matrix is minimized for each cluster. This method is first trained to fit a minimum covariance determinant model and then the outlier score is calculated using the Mahalanobis distance. However, problems can arise when clusters overlap significantly, leading to poor convergence of the algorithm.

## Ensembling Methods

### Isolation Forest

Isolation Forest (Liu et al., 2008) aims to separate outliers from the rest of the data samples. To calculate an isolation score for the data instances, random forest is employed. The method assumes that outliers are susceptible (duyarlı) to isolation and, therefore, can be isolated closer to the root of the tree. Isolation trees are thus able to build submodels on diferent data samples while maintaining low computational complexity and the ability to scale to handle large volumes of data and high-dimensional problems.

### LODA

Similarly, **lightweight on-line detector of anomalies (LODA)** comprises a collection of weak learners represented by one-dimensional histograms approximating probabilities of random data projections. The use of sparse (seyrek) projections makes LODA robust to both the large number of samples and missing data, allowing the detection of anomalous samples in real-time (Pevny, 2016).

## Neural Network-Based Methods

Neural network-based methods utilize feature learning to reduce dimensionality.

### Autoencoder

An autoencoder is an unsupervised neural network capable of nonlinear dimensionality reduction and reproducing input data vectors. Sakurada and Yairi (2014) showed that autoencoder (AE) can be successfully applied to outlier detection. To detect outliers in financial fraud, AEs can be trained to learn legitimate behaviour and compute a reconstruction error (autoencoder’ın loss function’ı) representing the outlier score (Sakurada & Yairi, 2014).

### Variational Autoencoder (VAE)

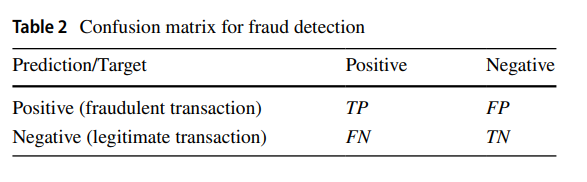
To achieve robustness in learning disentangled (çözülmemiş) representations, variational autoencoder (VAE) was proposed that utilizes both the joint data distribution and their latent (örtülü) generative factors (Burgess et al., 2018). VAE represents a probabilistic graphical model whose posterior distribution is estimated using a neural network. The outlier score of VAE is calculated as the reconstruction probability.

### Generative Adversial Networks (GANs)

Recently, generative adversial networks (GANs) have been deployed to unsupervised outlier detection. Specifically, multi-objective generative adversarial active learning (MO-GAAL) uses GANs to sample informative potential outliers following a mini-max game between a discriminator and a generator (Liu et al., 2019). Thus, GANs assist the discriminative algorithm in finding a boundary that can effectively separate fraudulent outliers from legitimate normal data. This has been exploited in several studies on financial fraud (Sethia et al., 2018; Delecourt & Guo, 2019).

# Performance Evaluation

In many related studies (Du et al., 2018; Misra et al., 2020; Mubalaike & Adali, 2018), the ratio of correctly classified transactions to the total number of transactions (i.e., accuracy) has been used as the evaluation measure. However, in the scenario of class-imbalanced data, this measure fails to detect well the model performance for the minority (fraud) class. In the absence of an adequate measure of fraud detection performance, existing fraud detection approaches rely on traditional measures of classification performance.



In reality, financial institutions try to reduce the risk of fraud while trying to comply with regulations, but Recall is difficult to estimate in the real world because FN is unknown (hidden fraud). Therefore, financial institutions can only calculate Precision (i.e., the number of transactions correctly identified as fraudulent as a percentage of all transactions that are expected to be fraudulent).



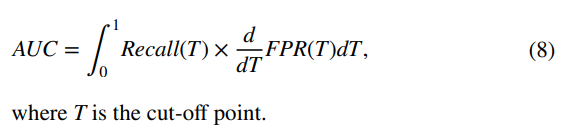
TP: Fraud transaction’lar doğru tahmin edilmiş.

FP: Fraud transaction’lar legitimate olarak tahmin edilmiş.

TN: Legitimate transaction’lar doğru tahmin edilmiş.

FN: Legitimate transaction’lar fraud olarak tahmin edilmiş.

The area under the receiver operating characteristic curve (AUC) has also been used as a more appropriate measure for fraud detection in mobile payment transactions due to its robustness to imbalanced data (Buschjäger et al., 2021; Mendelson & Lerner, 2020). AUC can be defined as the probability that a fraud detection model ranks a randomly selected fraudulent transaction higher than a randomly selected legitimate transaction, as follows:,



# Cost Saving Measures

The proposed cost savings measure was inspired by profit-based loan default prediction systems, considering potential returns and losses (Papouskova & Hajek, 2019; Ye et al., 2018). On the one hand, correct detection of a fraudulent transaction leads to the following cost savings:

