# Title: Credit Card Fraud Detection

|  |  |  |
| --- | --- | --- |
| First name | Last Name | IIT Email |
| Vikas | Sanil | Vsanil1@hawk.iit.edu |
|  |  |  |
|  |  |  |

Important Notes:

* Each group must submit ONLY one copy by a single team member!

1. **Introduction**

Worldwide financial losses caused by credit card fraudulent activities are worth tens of billions of dollars. Financial Institutions keep upgrading their credit card fraud detection algorithm to early detect such incidents and reduce financial losses.

Analyzing credit card fraud detection datasets will provide a good opportunity to apply different algorithms learned in this course and understand their merits and demerits. This will also provide a platform to showcase my skills during my job search.

1. **Data Sets (or Applications if it is category 2)**

**Found data:** From Kaggle datasets**:** <https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud>

**Source:** Machine Learning Group – ULB

**Data Type:** Anonymized credit card transactions labeled as fraudulent or genuine

**Content:**

The dataset contains transactions made by credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) accounts for 0.172% of all transactions.

It contains only numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, the source cannot provide the original features and more background information about the data. There are 31 features. Features V1, V2, … V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. Feature 'Class' is the label and it takes value 1 in case of fraud and 0 otherwise.

1. **Research Problems**

* Use supervised learning algorithms such as Decision trees and SVC to come up with fraud prediction models.
* Verify whether unsupervised learning can provide any insight into the data by clustering and then feeding those results to Supervised learning algorithms.
* Use Ensemble methods such as Random Forest and Gradient Tree Boosting to feed different model predictions and see whether that improves the performance.
* Validate each model using hold-out validation.
* Evaluate each model under AUC ROC and Average precision.

1. **Potential Solutions (Solutions and Implementations if it is category 2)**

* Will find the best ensemble method model to predict credit card fraud.

1. **Evaluations**

* Will validate the models based on hold-out validation.
* Will evaluate each model performance using ROC AUC and Average Precision

1. **Expected Outcomes**

A better-performing credit card fraud detection model.