

Master of Science

**Business Intelligence and Analytics**

Spring 2017

May 3, 2017

**SAKETH PATIBANDLA**

***CWID****: 10413333*

**KNOWLEDGE DISCOVERY IN DATABASES (MIS 637-A)**

Project Proposal

**Ethics Statement**

I pledge on my honor that I have not given or received any unauthorized assistance on this assignment/examination. I further pledge that I have not copied any material from a book, article, the Internet or any other source except where I have expressly cited the source.

Signature: Saketh Patibandla Date: 05/03/2017

# **Objective:** To detect fraudulent credit card transactions in Europe in September 2013 with real time transaction data that was taken from a partner bank with the ULB and predict accuracy to reduce credit card fraud.

# **Dataset**: Credit card transaction dataset.

# **Data Description:**

# The dataset has been collected during a research collaboration of Worldline and the Machine Learning Group of ULB (Université Libre de Bruxelles) of Brussels on big data mining and fraud detection.

* The dataset contains transactions made by credit cards in September 2013 that belong to cardholders across Europe.
* There are 284807 records that account to the same number of transactions.

## **The attributes that are present in the dataset are divided into the following:**

There are 32 attributes in the dataset out of which 30 are features for fraud detection, an ID and a target variable.

* **Transaction\_ID** – The unique identifier for each transaction that occurred.
* **Time**: The number of seconds elapsed between each transaction and the first selection of the dataset.
* **Amount**: The amount of money used by the cardholder in the respective transaction.
* V1 – V28: V1 to V28 are 28 principal components extracted on performing a comprehensive Principal component transformation/analysis (PCA). These are numerical data extracted from details of customers, transaction location, Business of the merchant and other details are kept confidential due to the university and credit card issuing authority policy.

**Target Variable:**

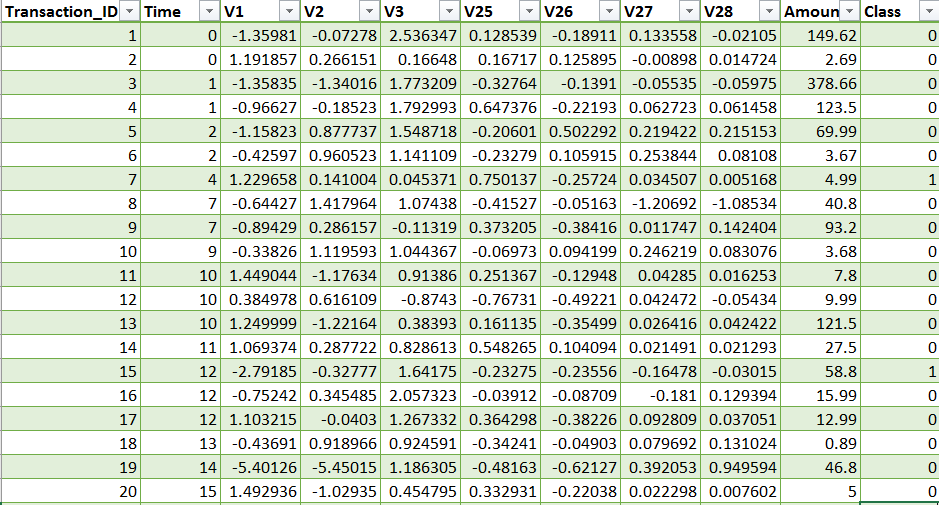
* Class: The attribute ‘Class’ is the responsible variable. It is binary and shows 1 in case of fraud and 0 otherwise.

**Technique and steps involved: CRISP Methodology**

Since the functionality, business requirement decisions, objective are clear, the steps from the other phases are stated below.

1. Data collection and data handling. The required data is arranged based on the requirement and the necessary variables are normalized and unnecessary variables are removed from the dataset.
2. Feature Engineering: PCA is performed on the data, attributes are divided into features and kept ready for machine learning.
3. Target and Test set distribution: The attribute ‘CLASS’ is segregated into training (target) and test sets and now, the algorithm can be run based on the Target from this step.
4. Machine learning Phase (Execution and classification): Three different algorithms are used on the dataset, CART, Random Forest (classification algorithm) and an advanced algorithm called XGBoost which enhances accuracy from C4.5 and Random Forest are performed and compared for the best accuracy to detect Fraud and to determine the best algorithm for Credit card transactions.
5. The three algorithms will be run in order with the input data and a confusion matrix is shown in Python individually to understand how many deviations existed from the actual ‘Class’ value and a comparative study between them.

**Sample Data:**



**Impact of Project:**

After implementing this project and predicting the ‘Class’ variable, the number of the fraudulent credit cards can be reduced by incorporating change management with the best algorithm that we obtain from the above analysis.

**Data source: (Link to data)**

<http://mlg.ulb.ac.be>