Credit card fraud Analysis

Introduction:

As we know that credit card fraud is a wide-ranging term for theft and fraud committed using or involving a payment via credit card, it happens to be one of the major problems faced by any bank. Considering the number of transactions happening over credit / debit card, the transaction data is at a risk and therefore puts the personal information of individuals at a higher risk. This is a major problem in Europe where credit or debit cards are accepted almost at every corner and by every means – online/offline. The consumers are not informed via SMS or EMAIL when money is spent on their credit or debit card which leads to late reporting of card frauds.

Problem Definition:

As explained above, credit card fraud happens to be more in Europe. The purpose of this analysis is to highlight some factors that can be responsible for this type of fraud. This project will focus on:

- What are the factors that contribute to credit card fraud? Is there any repetitive pattern that can explain the major cause of this fraud in Europe?

Data Source:

There is data on German credit card fraud available on WEKA (format of data is arff) - http://weka.8497.n7.nabble.com/file/n23121/credit_fruad.arff

This data contains information about the card fraud which has information like age, personal status, overdraft limit and average credit balance. This data is in ARFF file format which can be extracted by Python using SCIPPY library.

File Description:

There are about 21 columns in this data which are described as below (along with its applicable values) –

S.No.	Column Name	Possible Values
1	Over draft	a. < 0 b. 0 <= X < 200 c. >=200 d. No Checking
2	Credit history	a. 'no credits/all paid' b. 'all paid' c. 'existing paid' d. 'delayed previously' e. 'critical/other existing credit’
3	Purpose	a. new car b. used car c. furniture/equipment d. radio/tv e. domestic appliance f. repairs

		g. education h. vacation i. retraining j. business k. other
4	Average Credit Balance	a. < 100 b. 100 <= X < 500 c. 500 <= X < 1000 d. X > 1000 e. no known savings
5	Employment	a. Unemployed b. < 1 c. 1 <= X < 4 d. 4 <= X < 7 e. > = 7
6	Location	Real Unique between 0 to 4
7	Personal Status	a. Male div/sep b. Female div/sep/mar c. Male single d. Male mar/wid e. Female Single
8	Residence Since	Real Unique between 0 to 4
9	CC Age	Between 0 to 67
10	Other Parties	a. None b. Co Applicant c. Guarantor
11	Property Magnitude	a. Real Estate b. Life Insurance c. Car d. No known property
12	Other Payment Plans	a. Bank b. Stores c. None
13	Housing	a. Rent b. Own c. For Free
14	Existing Credits	Unique between 0 to 2
15	Current Balance	Unique between 0 to 2
16	Job	a. Unemployed /unskilled non-resident b. Unskilled resident c. Skilled d. High qualified /self-employed /Management
17	Own Telephone	a. None b. Yes
18	Foreign Worker	a. No b. Yes
19	Class	a. Good b. Bad

20	Number of dependants	Unique between 0 to 1
21	Credit Usage	Unique between 0 to 6

Methods:

1. Load data using scippy library and clean data using data wrangling techniques.
2. Use matplot library to analyse data between attributes like - current balance vs credit usage.
3. Use algorithms to further to develop predictive modelling.

Deliverables:

1. Code in python for making the required analysis.
2. A final presentation stating the problem statement along with a solution, identified using data engineering practices.