Masters Data Science and Analytics
Amir Ghaderi – 500794236
MRP Proposal

Credit Card Fraud Detection Using a Parallelized Bayesian Network

Introduction:
Bayesian networks (BNs) are a type of probabilistic graphical model that represents a set of random variables and their conditional probabilities in the form of a directed acyclic graph (DAG). Bayesian networks can be used to develop an inference model that is able to predict the probability of a specific event occurring. This probability is calculated using a Monte Carlo Simulation of the Bayesian network. A Monte Carlo Simulation refers to the process of repeating an algorithm a certain number of times in order to produce a distribution of outcomes. This distribution is then analyzed and a final probability is obtained. One of the major limitations to the Monte Carlo Simulation process is that it requires a high number of iterations to produce an accurate result. The issue is that a high number of iterations is extremely computationally heavy, especially when executed on a massive dataset.

There are two major objectives/goals for this project:

- The first is to develop a Bayesian network model that is able to accurately and precisely predict fraudulent credit card transactions. Ideally, this model will be able to predict credit card frauds with minimal errors and minimal false positive predictions.
- The second objective is to implement a parallelized version of this model, in order to compute a high number of iterations in a reasonable amount of time. I plan on executing this model in parallel by splitting it among each of my computers processing cores. The final probability will be calculated by aggregating the outputs from each of the processors.

Dataset Description:
The data set is title "*Anonymized credit card transactions labeled as fraudulent or genuine"* and can be found on kaggle (https://www.kaggle.com/dalpozz/creditcardfraud). The dataset contains actual credit card transaction made by European cardholders in September 2013. The dataset contains 284807 total transactions, 492 fraud cases (0.172%), and 284315 negative cases (99.828%). As you can see the dataset is highly class unbalanced, which is typical of a credit card dataset.

Technologies:
I plan on using R/Rstudio to construct and develop the Bayesian network model. I Initially plan on training the model on a subset of the dataset and then later extending it to the full dataset. In terms of visualizations I plan on using both R's ggplot2 package and a JavaScript library called D3.

Timeline:
This project will adhere to the following timeline:

May 15 – Project Proposal
June 12 – Literature Review and Exploratory Data Analysis
July 3 – Methodology and Experiments
July 24 – Results
August 14 – Written Report
August 17 – Oral Presentation
August 21 – Poster Presentation