PROJECT REPORT

on

Fraud Detection System using graphs (in Neo4j)

(Btech AI&DS IV Semester Mini project)

2020-24

Submitted to:

Department of computer Applications

Submitted by:

Ayush Rawat(05)

University Roll Number: 2017639

Under the guidance of ***Ms. Garima Sharma***

DEPARTMENT OF COMPUTER SCIENCE AND INFORMATION TECHNOLOGY

GRAPHIC ERA DEEMED TO BE UNIVERSITY, DEHRADUN

DECLARATION

I, ***Ayush Rawat*** *student of* ***B-tech, Semester 4****, Department of Computer Science and Engineering, Graphic Era Deemed To Be University, Dehradun, declare that the technical project work entitled "Fraud Detection System using graphs" has been carried out by me and submitted in partial fulfillment of the course requirements for the award of degree in B-tech of* ***Graphic Era Deemed To Be University*** *during the academic year* ***2021-22****. The matter embodied in this synopsis has not been submitted to any other university or institution for the award of any other degree or diploma.*

ACKNOWLEDGEMENT

I would like to take this opportunity to express my gratitude to entire faculty at Department of Computer Science and Information Technology, Graphic Era Deemed To Be University, Dehradun who evaluated the project from time to time and gave me the valuable suggestions as to how to improve the project.

I am grateful to **Ms. Garima Sharma**, Graphic Era Deemed To Be University, for her supervision, encouragement, inspiration, and guidance. Working under her is being an enriched experience. In all, I found congenial work environment in Graphic Era University, Dehradun and this project completion will mark a new beginning for me in the coming days.

I am highly indebted to Graphic Era University for providing me the required infrastructure and facilities to accomplish the given task.

Ayush Rawat

Btech AI&DS

2020-24

Graphic Era University

INSURANCE CLAIM FRAUD DETECTION

➢ Problem Statement

The aim of this project is to make a model that can detect a fraud in vehicle insurance claims. Frauds are unethical and losses to a company. By building a model that can detect insurance fraud, a company can get rid of the losses due to fraud which inturn will maximize their economy and earnings.
This project deals with insurance claim fraud detection using neo4j graph representation.

➢ Motivation:

- India is one of the biggest markets for the insurance companies across the world.
- It is estimated that the Indian Insurance companies loses close to $6 billion due to fraud per year.
- Insurance fraud can be submitting claims for injuries or damage that never happens and false reports of stolen vehicles.
- Hence the insurance companies has urgent need to have a model that can detect fraud with high accuracy.

➢ Tools Used:

- **Neo4j**- Neo4j is one of the popular Graph Database Management system that uses Cypher Query Language (CQL). Neo4j is written in Java Language. Neo4j creates graph in the form of nodes from the dataset and helps user to easily understand relationship between different nodes.

- **Google colab**- Google colab or colaboratory allows anybody to write and execute python code through the browser, and is mainly used to create machine learning models, data analysis and education.

➢ ***Header Files Used***

1) **Numpy**- NumPy is a general purpose Python library used for working with arrays, linear algebra, and matrices.

2) **Pandas**- Pandas is an open-source Python package used mainly for working with relational or labeled data easily.

3) **Matplotlib.pyplot**- Pyplot is a Matplotlib module used for creating animated and visualization in python. The various plots we can do using Pyplot are Line Plot, Histogram, Scatter, 3D Plot, Image, Contour, and Polar.

4) **Seaborn**- It is a python library basically used for data visualization and analysis. Seaborn is used to customize the graphs created and work in dataframes.

➢ ***Fraud in Automobile Insurance***

Fraud in Automobile can be of various forms like:

- Fraudsters issue fake insurance claims
- Making fake reports of stolen vehicles
- Claiming money for pre-existing damages

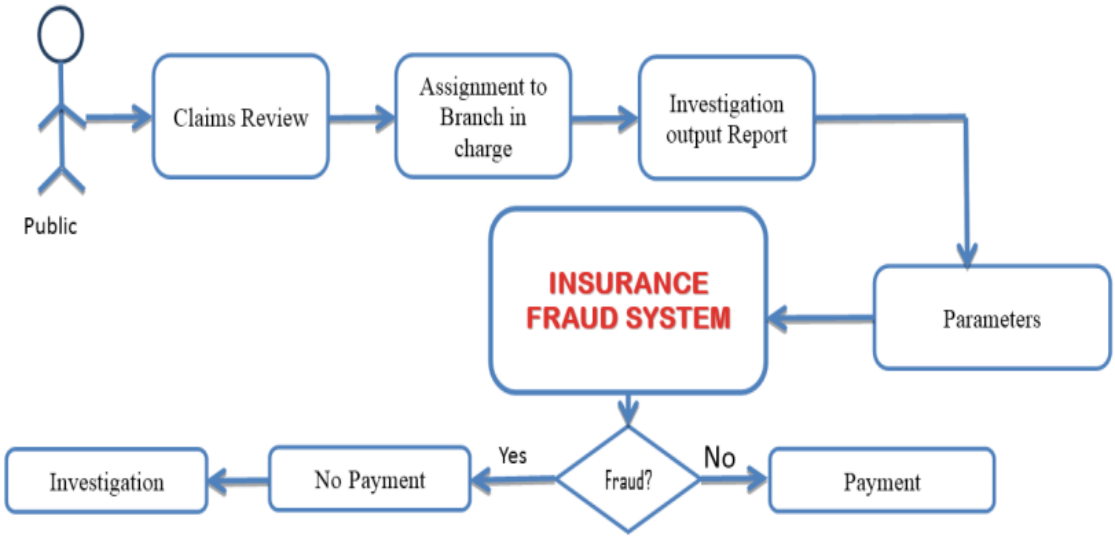
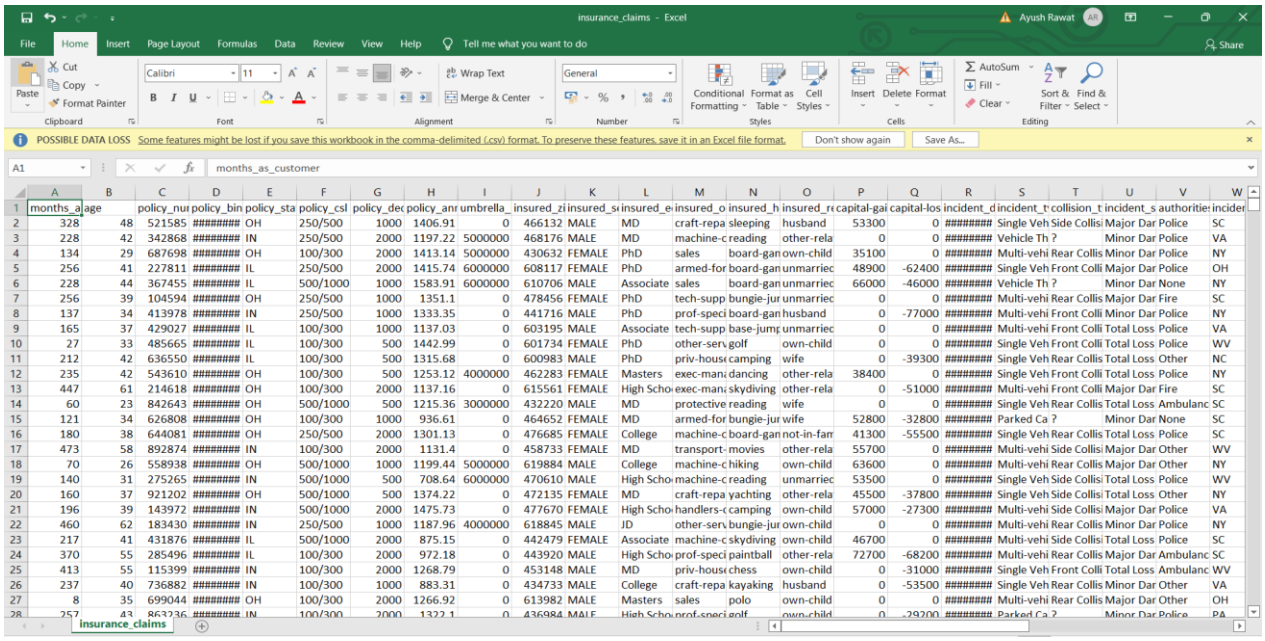


Figure 1: Architectural design of the proposed system

➢ ***Methodology***

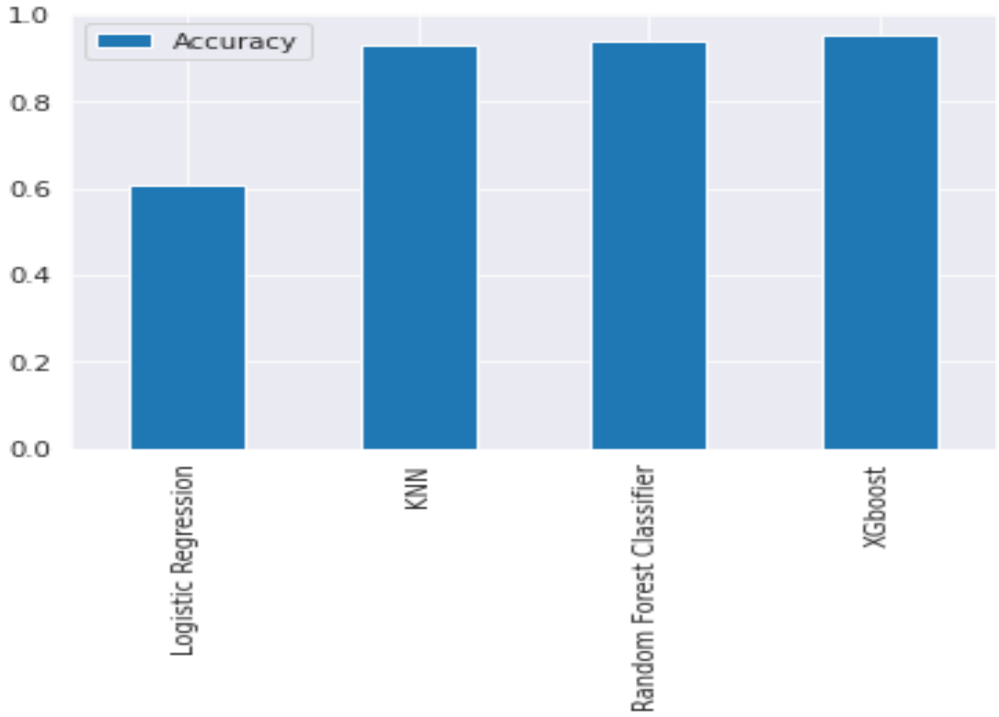
1) Data import from Dataset: The data stored in the CSV file need to be imported in our model.

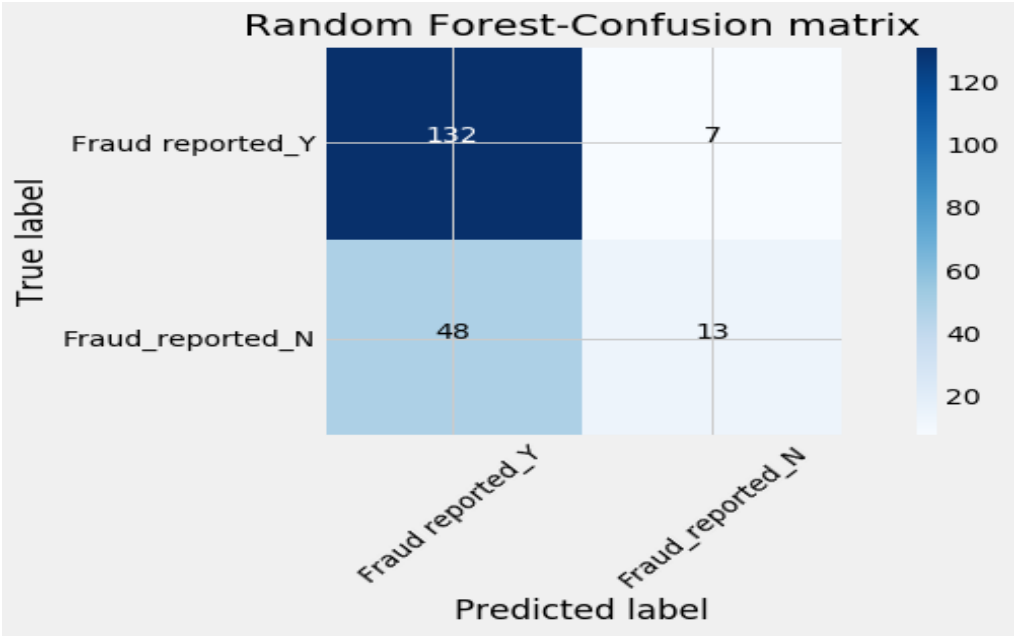


2) Data Preprocessing:
   a) Drop the column not required for prediction.
   b) In the dataset, their may be NULL values or '?' in the colomns . These need to be replaced by NaN values.
   c) Check for the categorical values and if present replace them with numeric values.
   d) Merge the pre-esxisting numeric columns and converted numeric columns to the main dataset.

3) Clustering: KMeans algorithm is used to create clusters in the preprocessed data. The KMeans model is trained over this data, and the model is used for further prediction.

4) Model Selection: After the clusters have been created, we find the best model for each cluster by checking the accuracy given by different algorithms like K nearest neighbour, Logistic Regression, XGboost. In this, I have used "Random Forest" classifier for my model.


Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
Logistic Regression
KNN
Random Forest Classifier
XGboost


Random Forest-Confusion matrix
Fraud reported_Y
Fraud_reported_N
True label
132
7
48
13
120
100
80
60
40
20
Fraud reported_Y
Fraud_reported_N
Predicted label

➢ ***How a graph database can fit into fraud detection***

If we consider a fraud that we are taking into account, different subject and the relationship between them like how they are linked together can give us the valuable insights out of our data. For representing these subjects and their relationships, the best suited is graphs.

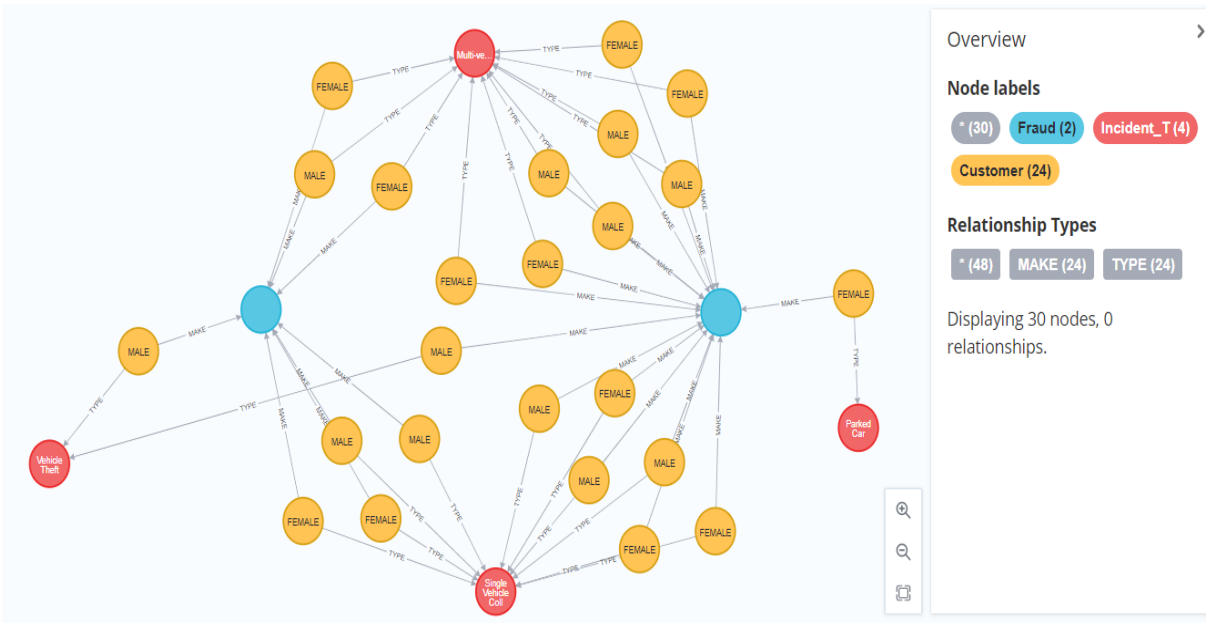
For building a fraud detection system that detects and prevents fraud as they happen, real-time traversal of a complex and highly interconnected dataset is essential.

Another important aspect to take in account is that traditional fraud systems looks for outliers(noise) but fraudsters try to act normal to avoid detection. Therefore, we need to detect fraudulent by linking different points and analysing normal behaviours.

➢ ***Why Neo4j?***

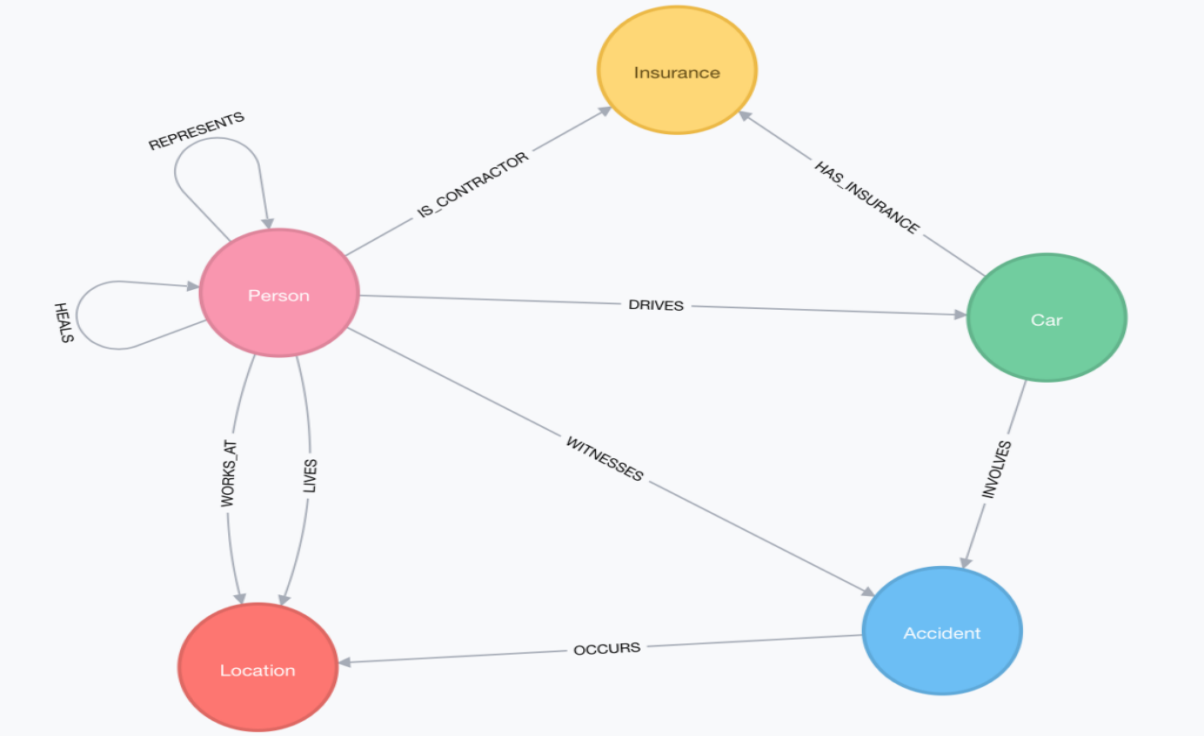
Neo4j stores interconnected data that is neither purely linear nor purely hierarchical and creates graphs and gives us the relationship in our data by making different nodes instead of rows and columns , making it easier to detect links of fraudulent activity regardless of the size and shape of the data.

Neo4j's versatile property makes it easier for organizations to build fraud detection data models and rules, and apply it in the real world applications.



➢ ***Fraud insurance scenarios***

Let's consider a simple model where we want to analyze two common scenarios: various people that are involved in different accidents but with different role for each accident and relationship between first aid location, where injured people live and also checks if there is any link of doctors , lawyers and witnesses with the person that met with the accident.



All of these subjects seem to act normal but what if two or more of them are involved in more than one accident cases with different roles. What if a doctor lives near a driver who gets first aid in a healthcare facility where that doctor works and also the healthcare facility is far from driver's home.

These scenarios are difficult to find through old models but are easy to understand through graph database. Hence, relationships become a key factor in identifying fraud rings or fraudsters.

➢ ***Conclusion:***

Machine learning and Neo4j helped a lot to build this model in which queries are clear, concise and fast enabling real-time detection preventing fraud before they happen.

Modern fraud detection tools can improved by looking at the connections between different points and linking them together to obtain a output rather than looking at the individual point. The use of **native graph database** like Neo4j is the best solution to achieve this.

The completion of this project went quite well, I have learned many new things like about the Neo4j, graph database, Machine Learning, Python language and how hard work helps to build a good project. This project can be used to build a major project out of it which can help the society to tackle the fraud easily