REAL TIME CREDIT CARD FRAUD DETECTION

Faisal Tadvi[1] , Sejal Shinde[2] , Devendra Patil[3] , Prof. Sejal Dmello[4]

[1]Faisal Tadvi, Dept of Information Technology Engineering, Atharva College of Engineering
[2]Sejal Shinde, Dept of Information Technology Engineering, Atharva College of Engineering
[3]Devendra Patil, Dept of Information Technology Engineering, Atharva College of Engineering
[4]Prof. Sejal Dmello, Dept of Information Technology Engineering, Atharva College of Engineering,
Maharashtra, India

---***---

Abstract - - *Credit card fraud remains the most common type of identity theft all across the globe resulting in a 28.65 billion dollar loss worldwide in the year 2019. Most of the credit card frauds are card not present (CNP) type which usually involves online transactions or over the phone transaction. As we are aware that machine learning is one of the most effective ways to recognize fraud. This paper mainly focuses on real-world credit card frauds and identifies each fraudulent transaction using Random Forest machine learning algorithm. Another key feature in this project is the real-time fraud detection using Kafka for this we use output of the implemented machine learning algorithms to determine whether transaction is fraudulent or not [2].All the predictions will be finally saved in Cassandra.*

Key Words: Credit Card Fraud, Random Forest Algorithm, Card Not Present, Kafka, Cassandra,

1.INTRODUCTION

Fraud in credit cards is the unauthorized and unwanted usage of an account by someone who is not the authorized owner of that account. Nowadays the world is getting digital and there is arise in online buying and selling platforms which also allow customers to pay online. Frauds are more likely to happen when consumers use online platforms for payment. There is an exponential growth rate in the credit card industry because of the increasing popularity of credit cards and also the growing trend of purchasing the product first and paying later. According to a survey, there are approximately 52 million credit card users in India. In the latest Nilson report of Nov 21st, 2019 fraud losses around the world reached $27.85 billion in 2018. It is predicted to rise to $35.67 billion by 2022 and in the next 10 years, the losses would reach $40.63 billion.

Credit card fraud is classified into two categories Card present (CP) and Card not present (CNP). In this paper, we are focusing on the second category. In Card not present (CNP) the fraudster does not have physical access to the card but rather has all the details of the account to commit the fraud.

The rest of the paper consists of the following; section 2 consists of the proposed approach. Section 3&4 provides the architecture and methodology including the result followed by conclusion and references in section 5.

2. PROPOSED APPROACH

An authorized consumer who owns a credit card his prior history and spending habits will be learned. The bank or the financial company will keep track of the consumer's history. How much amount does the consumer spends, what the accessories consumer purchases, at what frequency does the consumer purchase, etc. In this project, Random Forest machine learning algorithms will be running simultaneously on this data and after going through an extensive comparison we will start a spark streaming job and consume credit card transaction from Kafka after consuming if an anomaly is detected the transaction will be treated as fraudulent otherwise the transaction is genuine and all the predictions are saved in Cassandra.

3. ARCHITECTURE
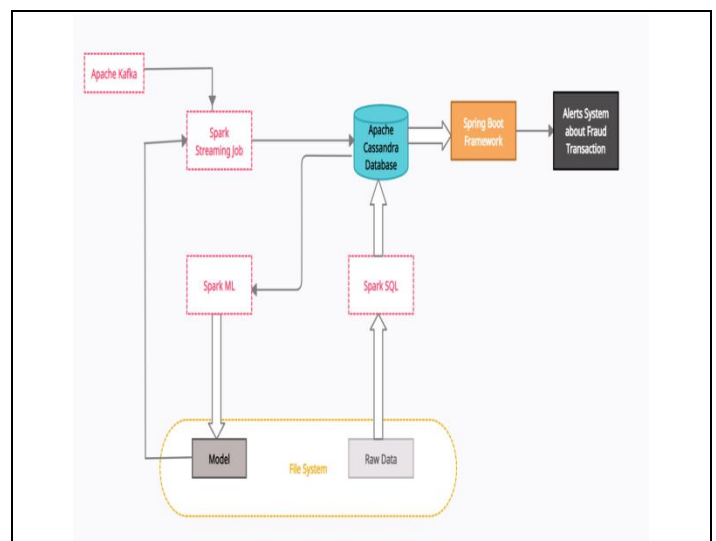
Fig-1: Architecture Of Real Time Credit Card Fraud Detection System

We have data in the file system this could be the last one month's data or last 6 months data. First, we will import this existing data from the file system to the Cassandra database

and to import this data we will run a spark job. This job will read this data it will do some ETL transformation and it will finally save this data to the Cassandra database. Next, we will run one more spark job this job is the spark ML job. This job will read the data from Cassandra, train on this data and will create a model. Finally, this model will be saved to the file system. After the model is saved to the file system we will start a spark streaming job this job will load the model and also it will start consuming credit card transactions from Kafka. After consuming these transactions it will predict whether these transactions are fraud or not and finally it will save these predictions to Cassandra. All this is done in real-time, now That we have our predictions we will use the springboot framework to display fraud transactions on the dashboard. These fraud transactions will be alerted on the dashboard in real-time.

4. METHODOLOGY

Machine Learning-

Machine learning is a related application of artificial intelligence that has systems that learn mechanically and can improve from experience without being specifically programmed. In this project, we train the machine using a machine learning algorithm to detect fraudulent transactions. Machine learning is one of the most exciting technologies one would find.

ALGORITHMS-

Random Forest-

We apply a random forest algorithm to classify the credit card record. Random Forest is a classification and regression algorithm. In short, it is a collection of decision tree classifiers. The random forest has the advantage over the decision tree that it corrects the habit of over fitting. A subset of the training set is sampled at random to train each individual tree, and then a decision tree is created, dividing each node into a selected feature of a random subset of the total functionality. Training data sets with many features and data instances is extremely fast in the random forest and since each tree is trained independently of the others. The random forest algorithm was found to provide a good estimate of the generalization error and to be resistant to over fitting.

One of the main reasons for using the random forest is that the training time is comparatively very small compared to other algorithms. While the accuracy of predicting credit card fraud is extremely important, the random forest predicts the output with great precision, even for the large data set it is running.

STEPS-
1. First, start by selecting random samples from a given dataset.

2. Next, this algorithm creates a decision tree for each sample. Then you get the result of the prediction for each decision tree.
3. In this voting will be performed for every predicted result.
4. Finally, select the prediction result with the most votes as the final prediction result.

BIG DATA OVERVIEW-

As we all know that there are numerous kinds of credit card fraud it's now become one in all the foremost common drawback for financial corporations and banks. Per annum these companies face a huge financial loss. There are several solutions to the present problem however none of it's enough. Big Data is an adequate and well structured technology that has real time process capabilities and is proving to be extremely helpful for credit card fraud detection.

Big data is a group of huge datasets that can't be processed by using regular computing techniques. It's not one technique or a tool, rather it's become an entire subject, that involves varied tools, technqiues and frameworks. In this project we are using big data technologies because it provides a detailed analysis which leads the firm to take more concrete decision which results in better operational efficiencies and additionally reduces the future risk of the firm.

Kafka-

Apache Kafka is a well-known streaming platform and is able to deal with of billions and trillions of activities in a day. Kafka is a distributed persistent message queue. Though it is a message queue , it is not implemented on a single system , instead it is implemented over a cluster of system. Messages to Kafka are stored on the disk, Messages are also stored on memory for faster access and also for durability purpose. Hence it is persistent message queue.

Since Kafka is a message queue , there must be a bunch of applications that are sending messages to Kafka. And there must be a bunch of applications that must be continuously consuming messages from Kafka. For example, credit card transaction application, this application will send all the transactions made by a customer using credit card in the form of messages to Kafka. We simulated this application in our model.

We will be using Kafka to consume the transactions from the transactions dataset and publish it in real-time. So, these real-time transactions will be consumed by spark streaming job.

Cassandra-

Cassandra is largely an allotted NoSQL columnar database. Most of the projects have used Cassandra as a storage layer in data-intensive projects. Cassandra uses a hoop architecture. It can be used by many applications as its storage layer. For example, Web applications developed using web frameworks can use Cassandra as its storage layer. Cassandra is a column-oriented database, so data is stored in a columnar manner. For each primary key , there could be a different number of columns , it's not mandatory to have a fixed number of columns as we have in RDBMS. There is no single point failure in Cassandra as it is not a master-slave model. Every node is equal in Cassandra , so in case if one node goes down there is always another node for a client to access data. As our data increases, we can scale the cluster without compromising on read and write throughput, which makes Cassandra linearly scalable. Although so many benefits , Cassandra does have some limitations like, it does not have built-in aggregation. Any aggregation must be precomputed and stored.
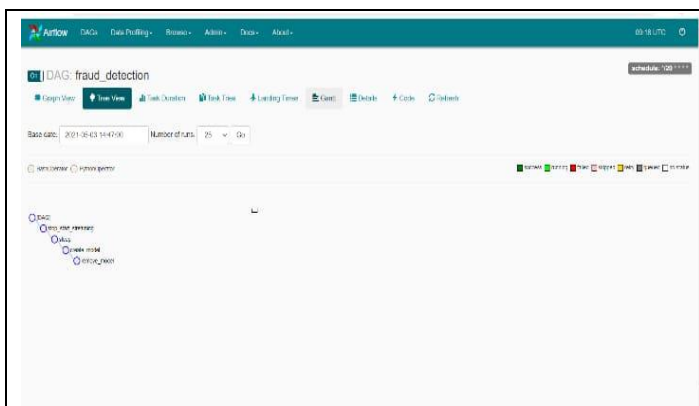
RESULTS-

Fig-2: AIRFLOW WEB UI

Fig-3: FRAUD ALERT MONITORING DASHBOARD

5. CONCLUSIONS

The number of credit card frauds has grown exponentially over the past few years, resulting in huge losses for financial companies, banks and consumers. It has become a necessity for financial companies to develop a system that can accurately identify these fraudulent transactions. To solve this problem, we developed a system that uses big data technology and machine learning. This system learns the consumer's buying pattern through studies to predict historical data and fraud. Our goal is to offer a robust architecture that can predict fraudulent transactions in real time.

Apache Kafka is a messaging system and well known streaming platform that can handle billions and trillions of activities every day. We use Kafka to consume the transactions from the transaction record and publish them in real time So, this real-time transactions will be consumed by spark streaming job. Predictions are stored in Cassandra. In this project, we are using Cassandra as the storage layer. Machine learning, on the other hand, is used to train the system in such a way that it continuously learns from all transactions and uses algorithms to differentiate between real and fraudulent transactions. This proposed system offers the user maximum security and precision.

REFERENCES

[1]   AnuruddhaThennakoon, Chee Bhagyani, Sasitha Premadasa, ShalithaMihiranga, Nuwan Kuruwitaarachchi, "Real-time Credit Card Fraud Detection Using Machine Learning," 9th International Machine Learning Conference, pp. 978-1-5386-5933, 2019.

[2]   "Credit Card Fraud Losses Reached $ 27.85 Billion ", Nilson Report, Number 1164, November 2019.

[3]   SoumayaOunacer, Soufiane Ardchir, Zahira Rachad, Mohamed Azouazi Real-Time Credit Card Fraud", International Magazine of Engineering & Technology, 7 (4.32) 93-97, 2018.

[4]   S.P.Maniraj, Aditya Saini, Shadab Ahmed, Swarna Deep Sarkar, "Credit Card Fraud Detection using Machine Learning and Data Science," IJRET, ISSN. 2278-0181, Vol. 8 issue 09,September-2019.

[5]   P.Kumar, Fahad Iqbal, "Credit Card Fraud Identification Using Machine Learning Approaches," 1st International Conference on Innovations in Information and Communication Technology, 2019

[6]   Lakshmi SVS, Selvani Deepthi Kavila, "Machine Learning For Credit Card Fraud Detection System,"

IJAERV, ISSN 0937-4562, Vol. 13, pp. 16819-16824, 2018.

[7]　Devi Meenakshi, Janani, Gayathri, Mrs. Indira, "Credit Card Fraud Detection Using Random Forest," IRJET, ISSN 2395-0056, Vol. 06, March 2019.