APPLIED DATA SCIENCE PHASE 1 PROJECT

INTRODUCTION TO DATA SCIENCE:

Data science is the practice of mining large data sets of raw data, both structured and unstructured, to identify patterns and extract actionable insight from them. This is an interdisciplinary field, and the foundations of data science include statistics, inference, computer science, predictive analytics, machine learning algorithm development, and new technologies to gain insights from big data.

Data science, in simple words, is the field of study that involves collecting, analyzing, and interpreting large sets of data to uncover insights, patterns, and trends that can be used to make informed decisions and solve real-world problems.

To define data science and improve data science project management, start with its life cycle. The first stage in the data science pipeline workflow involves capture: acquiring data, sometimes extracting it, and entering it into the system. The next stage is maintenance, which includes data warehousing, data cleansing, data processing, data staging, and data architecture.

By 2020, there will be around 40 zettabytes of data—that's 40 trillion gigabytes. The amount of data that exists grows exponentially. At any time, about 90 percent of this huge amount of data gets generated in the most recent two years, according to sources like IBM and SINTEF.

In fact, internet users generate about 2.5 quintillion bytes of data every day. By 2020, every person on Earth will be generating about 146,880 GB of data every day, and by 2025, that will be 165 zettabytes every year.

Simple data analysis can interpret data from a single source, or a limited amount of data. However, data science tools are critical to understanding big data and data from multiple sources in a meaningful way. A look at some of the specific data science applications in business illustrate this point and provide a compelling introduction to data science.

APPLICATIONS OF DATA SCIENCE:

Healthcare: Data science can identify and predict disease, and personalize healthcare recommendations.

Transportation: Data science can optimize shipping routes in real-time.

Sports: Data science can accurately evaluate athletes' performance.

Government: Data science can prevent tax evasion and predict incarceration rates.

E-commerce: Data science can automate digital ad placement.

Gaming: Data science can improve online gaming experiences.

Social media: Data science can create algorithms to pinpoint compatible partners.

In conclusion, the role of a Data Scientist is critical for businesses looking to make data-driven decisions.

Use of data science course:

Data Science helps organizations identify and refine target audiences by combining existing data with other data points for developing useful insights. Data Science also helps recruiters by combining data points to identify candidates that best fit their company needs.

PROBLEM DEFINITION: CREDIT CARD FRAUD DETECTION:

CREDIT CARD FRAUD: AN INTRODUCTION:-

Credit card fraud is an inclusive term for fraud committed using a payment card, such as a credit card or debit card.[1] The purpose may be to obtain goods or services or to make payment to another account, which is controlled by a criminal. The Payment Card Industry Data Security Standard (PCI DSS) is the data security standard created to help financial institutions process card payments securely and reduce card fraud.

Credit card fraud can be authorised, where the genuine customer themselves processes payment to another account which is controlled by a criminal, or unauthorised, where the account holder does not provide authorisation for the payment to proceed and the transaction is carried out by a third party. In 2018, unauthorised financial fraud losses across payment cards and remote banking totalled £844.8 million in the United Kingdom.

There are different types of credit card fraudulent activities occurs everywhere, includes,

1. Skimming
2. Dumpster diving
3. Phishing
4. Keystroke capturing
5. SIM Swap
6. Application fraud
7. Hacking. Etc

In conclusion, implementing robust credit card fraud detection systems is essential to safeguard financial transactions and protect both consumers and businesses from potential fraudulent activities.

ROLE OF MACHINE LEARNING IN CREDIT CARD FRAUD DETECTION :-

Touching a little more on the difficulties of credit card fraud detection, even with more advances in learning and technology every day, companies refuse to share their algorithms and techniques to outsiders. Additionally, fraud transactions are only about 0.01–0.05% of daily transactions, making it even more difficult to spot. Machine learning is similar to artificial intelligence where it is a sub field of AI where statistics is a subdivision of mathematics. With regards to machine learning, the goal is to find a model that yields that highest level without overfitting at the same time. Overfitting means that the computer system memorized the data and if a new transaction differs in the training set in any way, it will most likely be misclassified, leading to an irritated cardholder or a victim of fraud that was not detected. The most popular programming used in machine learning are Python, R, and MatLab. At the same time, SAS is becoming an increasing competitor as well. Through these programs, the easiest method used in this industry is the Support Vector Machine. R has a package with the SVM

function already programmed into it. When Support Vector Machines are employed, it is an efficient way to extract data. SVM is considered active research and successfully solves classification issues as well. Playing a major role in machine learning, it has “excellent generalization performance in a wide range of learning problems, such as handwritten digit recognition, classification of web pages and face detection.” SVM is also a successful method because it lowers the possibility of overfitting and dimensionality.

Machine learning represents an essential pillar for fraud detection. Its toolkit provides two approaches:

Supervised methods: k-nearest neighbors, logistic regression, support vector machines, decision tree, random forest, time-series analysis, neural networks, etc.

Unsupervised methods: cluster analysis, link analysis, self-organizing maps, principal component analysis, anomaly recognition, etc.

ROLE OF DATA SCIENCE IN CREDIT CARD FRAUD DETECTION :-

Nowadays, data has become the most valuable asset in this sphere. Data science is a necessary requirement for banks to keep up with their rivals, attract more clients, increase the loyalty of existing clients, make more efficient data-driven decisions, empower their business, enhance operational efficiency, improve existing services/products and introduce new ones, reinforce security, and, as a result of all these actions, obtain more revenue. It is not surprising that the majority of all data science job demand comes from banking.

Data science allows the banking industry to successfully perform numerous tasks, including:

- Investment risk analysis
- Customer lifetime value prediction
- Customer segmentation
- Customer churn rate prediction
- Personalized marketing
- Customer sentiment analysis

Virtual assistants and chatbots

Data science plays a critical role in modern credit card fraud detection, revolutionizing how financial institutions safeguard their customers' assets. By harnessing the power of data analysis, machine learning, and real-time monitoring, data science helps identify and prevent fraudulent activities, providing a more secure and seamless experience for cardholders. In this introduction, we will explore the fundamental aspects of this role, highlighting how data science contributes to the ongoing battle against credit card fraud.

As the digital economy grows, so does the sophistication of fraudulent activities. Credit card fraudsters continually devise new tactics to exploit vulnerabilities, making it imperative for financial institutions to stay ahead of these threats..

*credit card fraud detection work:*
*Credit card fraud detection strategies can vary depending on the credit card issuer. According to Inscribe, some of the most common practices involve using AI, machine learning and data analysis to review spending patterns and account behavior.*

*If account activity falls outside the user's typical behavior, the analysis tool can alert the credit card issuer. The card issuer can then decide whether to approve the transaction or notify the cardholder.*

*What's defined as unusual account activity can vary. But Inscribe says some examples of anomalies include*:

- Multiple transactions in a short period of time
- Sudden increases in spending
- Making a particularly large purchase
- Multiple transactions from the same retailer
- Purchases that take place in a foreign country or unusual location

DESIGN THINKING :

Design thinking is a non-linear, iterative process that teams use to understand users, challenge assumptions, redefine problems and create innovative solutions to prototype and test.

Main challenges involved in credit card fraud detection are:

1. Enormous Data is processed every day and the model build must be fast enough to respond to the scam in time.
2. Data i.e most of the transactions (99.8%) are not fraudulent which makes it really hard for detecting the fraudulent ones
3. Data availability as the data is mostly private.
4. Misclassified Data can be another major issue, as not every fraudulent transaction is caught and reported.
5. Adaptive techniques used against the model by the scammers.

Data science plays a crucial role in credit card fraud detection by utilizing various techniques and technologies to identify and prevent fraudulent transactions. Here's how data science is applied:

Data Source:.

A data source is the location where data that is being used originates from. A data source may be the initial location where data is born or where physical information is first digitized, however even the most refined data may serve as a source, as long as another process accesses and utilizes it.

Data Preprocessing:

. Data Preprocessing is the process of converting raw data into a format that is understandable and usable. It is a crucial step in any Data Science project to carry out an efficient and accurate analysis. It ensures that data quality is consistent before applying any Machine Learning or Data Mining techniques.

Feature Engineering:

Feature engineering involves a set of techniques that enable us to create new features by combining or transforming the existing ones. These techniques help to highlight the most important patterns and relationships in the data, which in turn helps the machine learning model to learn from the data more effectively.

Model Selection:

Model selection is the process of selecting the best model from all the available models for a particular business problem on the basis of different criterions such as robustness and model complexity.

Model Training:

Model training is the phase in the data science development lifecycle where practitioners try to fit the best combination of weights and bias to a machine learning algorithm to minimize a loss function over the prediction range.

Evaluation:

Model evaluation is the process of using different evaluation metrics to understand a machine learning model's performance, as well as its strengths and weaknesses. Model evaluation is important to assess the efficacy of a model during initial research phases, and it also plays a role in model monitoring.

In conclusion, the introduction sets the stage for our exploration into the critical realm of credit card fraud detection using data science techniques. As the financial landscape becomes increasingly digital, the need to safeguard sensitive financial transactions grows more urgent. Through this project, we aim to harness the power of data science to mitigate the risks associated with credit card fraud.