APPLIED DATA SCIENCE PHASE-1 PROJECT

DATA SCIENCE:

1. Data science is an interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data.
2. Data Science is about data gathering, analysis and decision-making.
3. Data Science is about finding patterns in data, through analysis, and make future predictions.

By using Data Science, companies are able to make:

1. Better decisions (should we choose A or B)
2. Predictive analysis (what will happen next?)
3. Pattern discoveries (find pattern, or maybe hidden information in the data)

A Data Scientist requires expertise in several backgrounds:

1. Machine Learning
2. Statistics
3. Programming (Python or R)
4. Mathematics
5. Databases

A Data Scientist must find patterns within the data. Before he/she can find the patterns, he/she must organize the data in a standard format.

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:

Data science has a wide range of applications across various industries. Here are some key application areas:

1. Business Analytics: Data science helps businesses analyze customer behavior, optimize operations, and make data-driven decisions for improved profitability.
2. Healthcare: It aids in patient diagnosis, drug discovery, and predicting disease outbreaks through analysis of medical data.
3. Finance: Data science is used for fraud detection, algorithmic trading, credit risk assessment, and portfolio optimization in the financial sector.
4. E-commerce: Recommender systems use data science to suggest products to customers, improving user experience and increasing sales.

5. Marketing: Marketers use data science for customer segmentation, A/B testing, and personalized marketing campaigns.

6. Manufacturing: Predictive maintenance uses data science to anticipate machinery failures, reducing downtime and costs.

7. Transportation and Logistics: It optimizes route planning, inventory management, and demand forecasting in supply chains.

8. Energy Management: Data science helps in optimizing energy consumption, grid management, and renewable energy integration.

9. Social Media: Platforms use data science for content recommendation, sentiment analysis, and ad targeting.

10. Education: Personalized learning platforms leverage data science to adapt content to individual student needs.

11. Government: Data science is used for policy analysis, crime prediction, and optimizing public services.

12. Environmental Science: It aids in climate modeling, natural disaster prediction, and ecological conservation.

13. Sports Analytics: Teams use data science for player performance analysis, injury prevention, and game strategy.

14. Telecommunications: It optimizes network performance, predicts equipment failures, and improves customer service.

15. Human Resources: Data science is applied to talent acquisition, employee retention, and workforce planning.

In conclusion, the role of a Data Scientist is critical for businesses looking to make data-driven decisions. Data Scientists are responsible for collecting, organizing, analyzing, and interpreting data to identify trends and correlations.

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

n 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 :

In simple words, you can explain machine learning as a type of artificial intelligence (AI) or a subset of AI which allows any software applications or apps to be more precise and accurate for finding and predicting outcomes. Machine learning algorithms use historical data to predict new outcomes or output values. There are different use cases for machine learning like fraud detection, malware threat detection, recommendation engines, spam filtering, healthcare, and many others.

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:

1. Investment risk analysis
2. Customer lifetime value prediction
3. Customer segmentation
4. Customer churn rate prediction
5. Personalized marketing
6. Customer sentiment analysis
7. Virtual assistants and chatbot

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..

DESIGN THINKING :

The goal of this project is to use machine learning techniques to improve the accuracy of detecting credit card fraudulent activities. The model will be trained on a dataset of credit card transactions and the goal is to improve the measure called Area Under the Precision-Recall Curve (AUPRC).

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.
Here are some of the challenges that complicate the fraud detection process :

1. Changing fraud patterns over time : This one is the toughest to address since the fraudsters are always in the lookout to find new and innovative ways to get around the systems to commit the act. Thus it becomes all-important for the deep learning models to be updated with the evolved patterns to detect. This results in a decrease in the model's performance and efficiency. Thus the machine learning models need to keep updating or fail their objectives.
2. Class Imbalance : Practically only a small percentage of customers have fraudulent intentions. Consequently, there's an imbalance in the classification of fraud detection models (that usually classify transactions as either fraudulent or non-fraudulent) which makes it harder to build them. The fallout of this challenge is a poor user experience for genuine customers, since catching the fraudsters usually involves declining some legitimate transactions.
3. Model Interpretations: This limitation is associated with the concept of explainability since models typically give a score indicating whether a transaction is likely to be fraudulent or not — without explaining why.
4. Feature generation can be time-consuming :Subject matter experts can require long periods of time to generate a comprehensive feature set which slows down the fraud detection process.

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.

Some of the important data science algorithms include regression, classification, clustering techniques, decision trees, random forests, and machine learning techniques like supervised, unsupervised, and reinforcement learning.

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.