Data Science

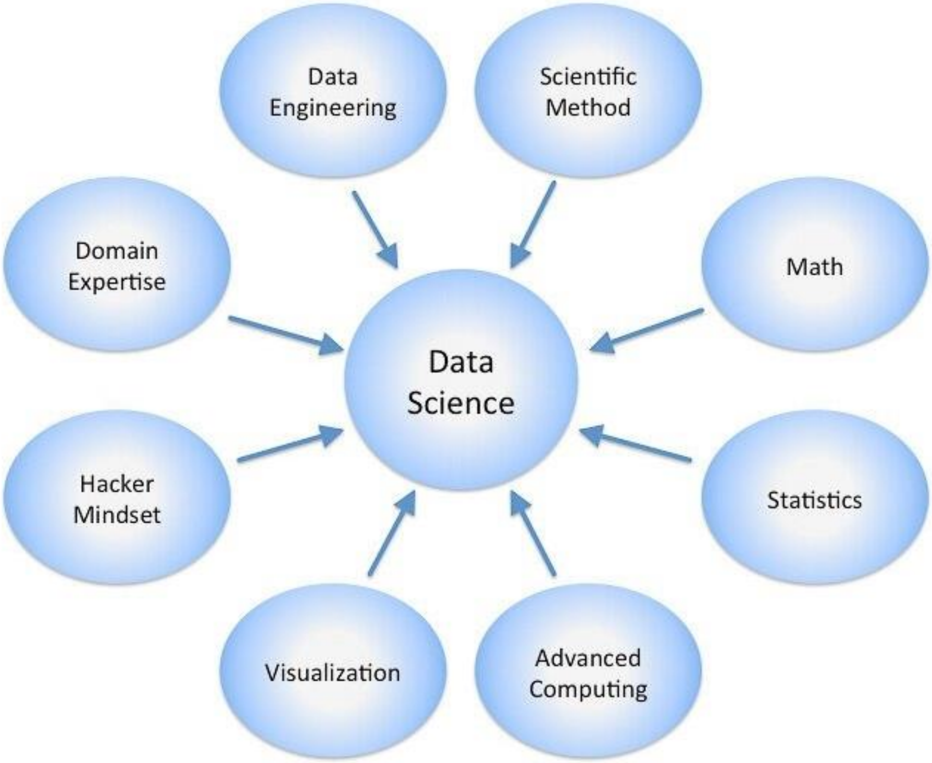
Data science combines math and statistics, specialized programming, advanced analytics, artificial intelligence (AI), and machine learning with specific subject matter expertise to uncover actionable insights hidden in an organization's data. These insights can be used to guide decision making and strategic planning.

Data science is a multidisciplinary field that uses statistical and computational methods to extract insights and knowledge from data. It involves a combination of skills and knowledge from various fields such as statistics, computer science, mathematics, and domain expertise.

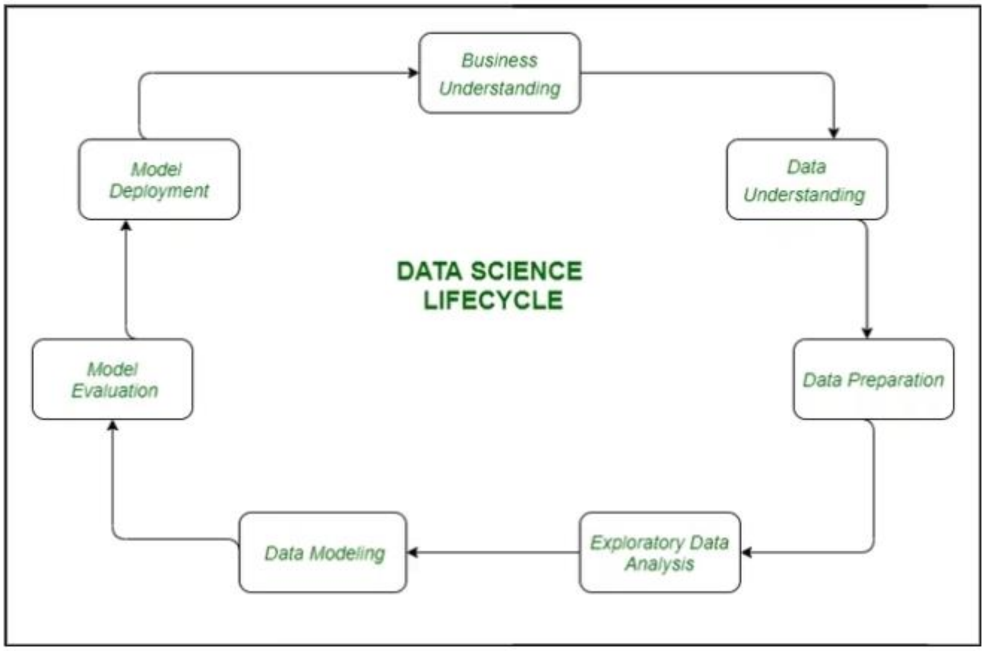
The process of data science involves several steps, including data collection, cleaning, exploration, analysis, and interpretation. These steps are often iterative, and the process may be refined based on the results obtained.

One of the primary goals of data science is to extract insights from data that can be used to inform decision-making. This may involve identifying patterns or trends in data, making predictions about future outcomes, or identifying opportunities for optimization or improvement.

Data Science is not only a synthetic concept to unify statistics, data analysis and their related methods but also comprises its results. It includes three phases, design for data, collection of data, and analysis on data.



Process involved in Data science:



There are some steps that are necessary for any of the tasks that are being done in the field of data science to derive any fruitful results from the data at hand.

Data Collection – After formulating any problem statement the main task is to calculate data that can help us in our analysis and manipulation. Sometimes data is collected by performing some kind of survey and there are times when it is done by performing scrapping.

Data Cleaning – Most of the real-world data is not structured and requires cleaning and conversion into structured data before it can be used for any analysis or modeling.

Exploratory Data Analysis – This is the step in which we try to find the hidden patterns in the data at hand. Also, we try to analyze different factors which affect the target variable and the extent to which it does so. How the independent features are related to each other and what can be done to achieve the desired results all these answers can be extracted from this process as well. This also gives us a direction in which we should work to get started with the modeling process.

Model Building – Different types of machine learning algorithms as well as techniques have been developed which can easily identify complex patterns in the data which will be a very tedious task to be done by a human.

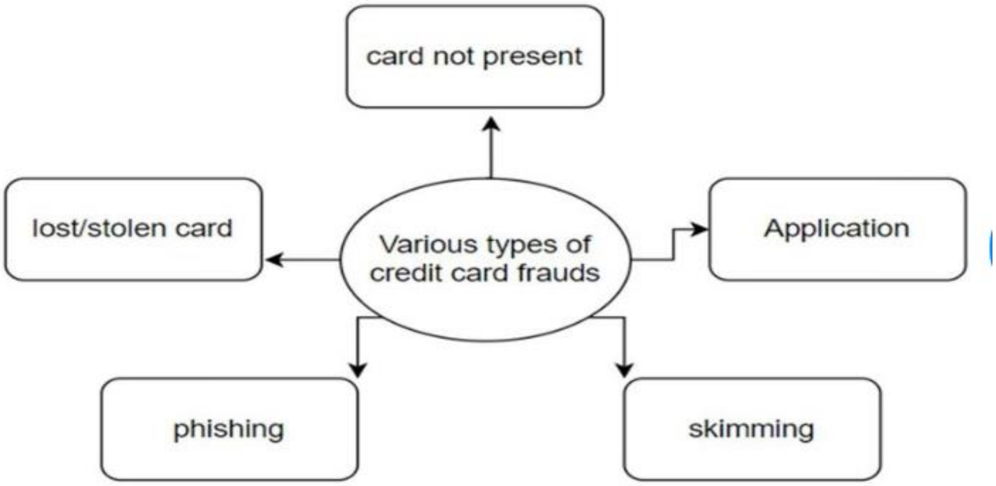
Model Deployment – After a model is developed and gives better results on the holdout or the real-world dataset then we deploy it and monitor its performance. This is the main part where we use our learning from the data to be applied in real-world applications and use cases.

Credit Card Fraud detection:

The challenge is to recognize fraudulent credit card transactions so that the customers of credit card companies are not charged for items that they did not purchase.

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. Imbalanced 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.

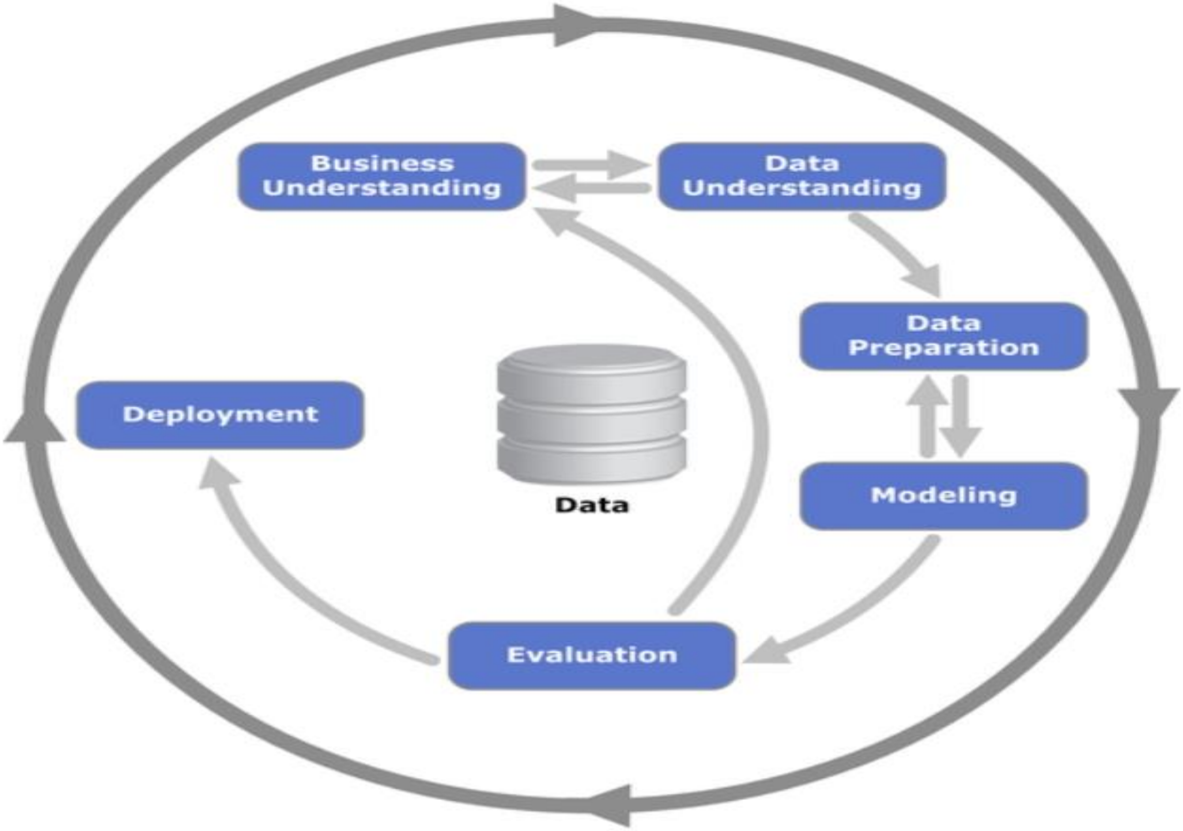


Credit fraud detection: Approach

Data Understanding :

Good data allows organizations to establish baselines, benchmarks, and goals to keep moving forward. Because data allows you to measure, you will be able to establish baselines, find benchmarks and set performance goals. A baseline is what a certain area looks like before a particular solution is implemented.

Data preparation is the process of preparing raw data so that it is suitable for further processing and analysis. Key steps include collecting, cleaning, and labeling raw data into a form suitable for machine learning (ML) algorithms and then exploring and visualizing the data.



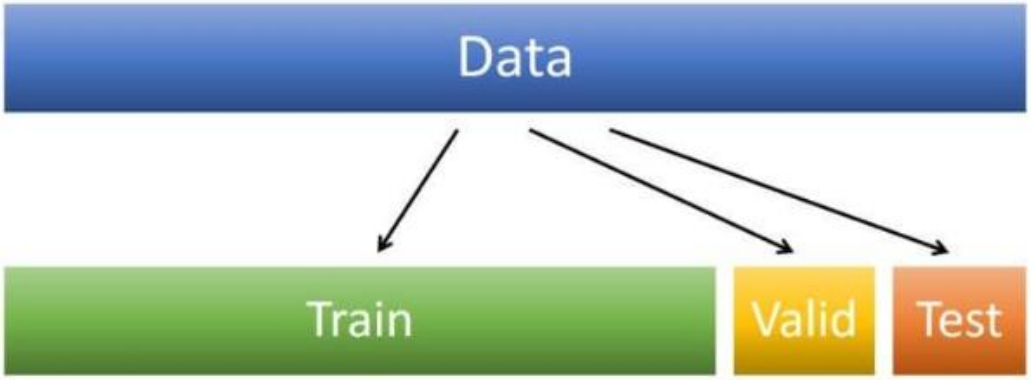
Data Analysis:

Data analysis is a process for obtaining raw data, and subsequently converting it into information useful for decision-making by users. Data is collected and analyzed to answer questions, test hypotheses, or disprove theories. It is the process of collecting, modeling, and analyzing data using various statistical and logical methods and techniques.

Splitting the data:

Data splitting is when data is divided into two or more subsets. Typically, with a two-part split, one part is used to evaluate or test the data and the other to train the model. Data splitting is an important aspect of data science, particularly for creating models based on data.Here are a few common processes for splitting data:Train-Test Split: The dataset is divided right into a training set and a trying out set test validation.

Test Split: The dataset is split into three subsets – a schooling set, a validation set, and a trying out set.



Model building:

Model building is an essential part of data analytics and is used to extract insights and knowledge from the data to make business decisions and strategies. In this phase of the project data science team needs to develop data sets for training, testing, and production purposes. These data sets enable data scientists to develop an analytical method and train it while holding aside some of the data for testing the model. Model building in data analytics is aimed at achieving not only high accuracy on the training data but also the ability to generalize and perform well on new, unseen data. Therefore, the focus is on creating a model that can capture the underlying patterns and relationships in the data, rather than simply memorizing the training data.

Cross validation:

Cross-validation is a statistical technique used by data scientists for training and evaluating a machine learning model, with a focus on ensuring reliable model performance. To understand how cross-validation supports model development, we first need to understand how data is used to train and evaluate models.

Oversampling:

SMOTE (synthetic minority oversampling technique) is one of the most commonly used oversampling methods to solve the imbalance problem. It aims to balance class distribution by randomly increasing minority class examples by replicating them. SMOTE synthesises new minority instances between existing minority instances.

Hyper parameter tuning:

Hyperparameter tuning allows data scientists to tweak model performance for optimal results. This process is an essential part of machine learning, and choosing appropriate hyperparameter values is crucial for success. For example, assume you’re using the learning rate of the model as a hyperparameter.

Conclusion:

This is an overview of approach for credit card fraud detection, if the above steps were put into to process the fraudulent occurs in credit card can be detected.