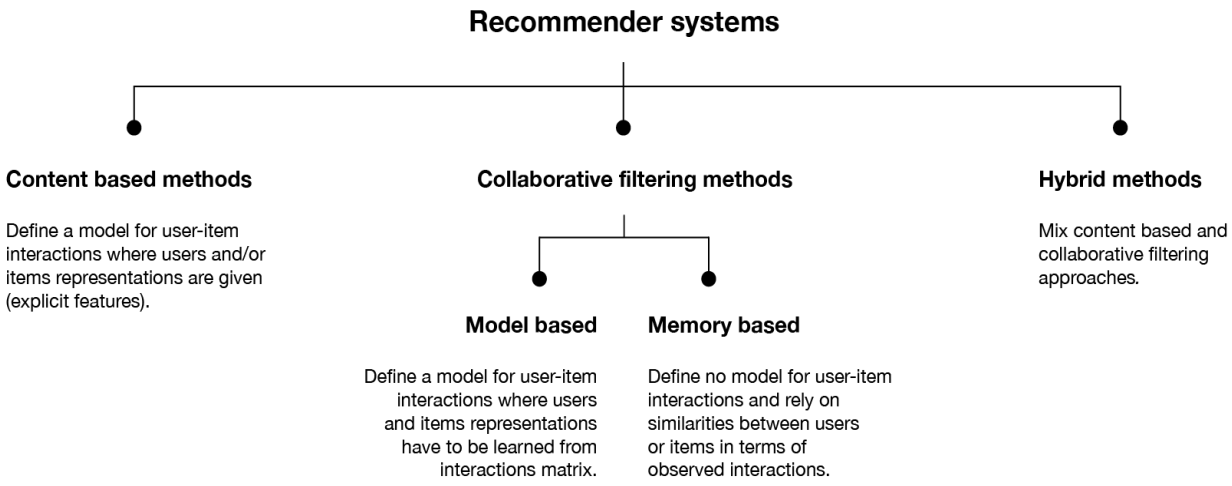
Day 10: Data Science Interview Preparation

Q1. What is a Recommender System?

Answer: A recommender system is today widely deployed in multiple fields like movie recommendations, music preferences, social tags, research articles, search queries and so on.

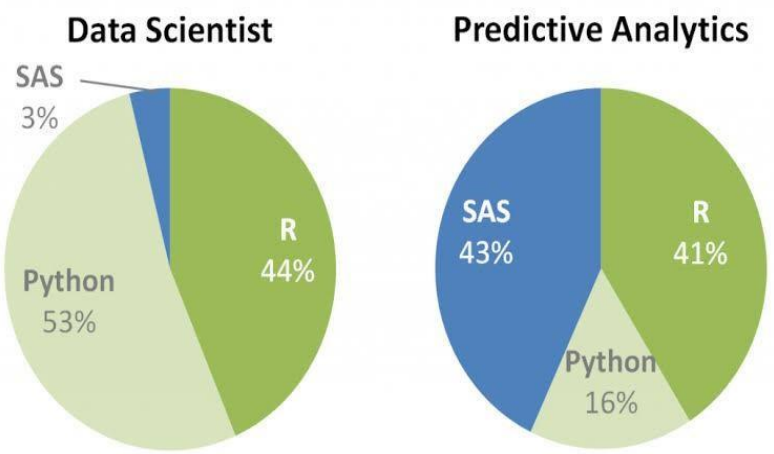


The recommender systems work as per collaborative and content-based filtering or by deploying a personality-based approach.

This type of system works based on a person's past behaviour to build a model for the future. This will predict the future product buying, movie viewing or book reading by people. It also creates a filtering approach using the discrete characteristics of items while recommending additional items.

Q2. Compare SAS, R and Python programming?

Answer: SAS: it is one of the most widely used analytics tools used by some of the biggest companies on earth. It has some of the best statistical functions, graphical user interface, but can come with a price tag and hence it cannot be readily adopted by smaller enterprises.



R: The best part about R is that it is an Open-Source tool and hence used

generously by academia and the research community. It is a robust tool for statistical computation, graphical representation, and reporting. Due to its open-source nature, it is always being updated with the latest features and then readily available to everybody.

Python: Python is a powerful open-source programming language that is easy to learn, works well with most other tools and technologies. The best part about Python is that it has innumerable libraries and community created modules making it very robust. It has functions for statistical operation, model building and more.

Q3. Why is important in data analysis?

Answer: With data coming in from multiple sources it is important to ensure that data is good enough for analysis.

This is where data cleansing becomes extremely vital. Data cleansing extensively deals with the process of detecting and correcting of data records, ensuring that data is complete and accurate and the components of data that are irrelevant are deleted or modified as per the needs.

This process can be deployed in concurrence with data wrangling or batch processing.

Once the data is cleaned it confirms with the rules of the data sets in the system. Data cleansing is an essential part of the data science because the data can be prone to error due to human negligence, corruption during transmission or storage among other things.

Data cleansing takes a huge chunk of time and effort of a Data Scientist because of the multiple sources from which data emanates and the speed at which it comes.

Q4. What are the various aspects of a Machine Learning process?

Answer: Here we will discuss the components involved in solving a problem using machine learning.

Domain knowledge

This is the first step wherein we need to understand how to extract the various features from the data and learn more about the data that we are dealing with.

It has got more to do with the type of domain that we are dealing with and familiarizing the system to learn more about it.

Feature Selection

This step has got more to do with the feature that we are selecting from the set of features that we have. Sometimes it happens that there are a lot of features and we have to make an intelligent decision regarding the type of feature that we want to select to go ahead with our machine learning endeavour.

Algorithm

This is a vital step since the algorithms that we choose will have a very major impact on the entire process of machine learning. You can choose between the linear and nonlinear algorithm. Some of the algorithms used are Support Vector Machines, Decision Trees, Naïve Bayes, K-Means Clustering, etc.

Training

This is the most important part of the machine learning technique and this is where it differs from the traditional programming. The training is done based on the data that we have and providing more real-world experiences. With each consequent training step the machine gets better and smarter and able to take improved decisions.

Evaluation

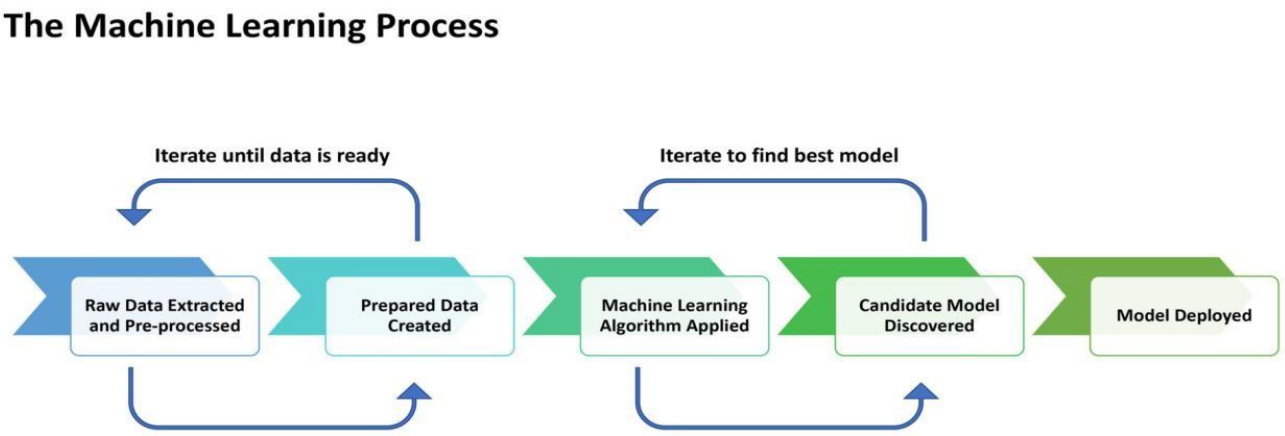
In this step we evaluate the decisions taken by the machine in order to decide whether it is up to the mark or not. There are various metrics that are involved in this process and we have to closed deploy each of these to decide on the efficacy of the whole machine learning endeavour.

Optimization

This process involves improving the performance of the machine learning process using various optimization techniques. Optimization of machine learning is one of the most vital components wherein the performance of the algorithm is vastly improved. The best part of optimization techniques is that machine learning is not just a consumer of optimization techniques, but it also provides new ideas for optimization too.

Testing

Here various tests are carried out and some these are unseen set of test cases. The data is partitioned into test and training set. There are various testing techniques like cross-validation in order to deal with multiple situations.

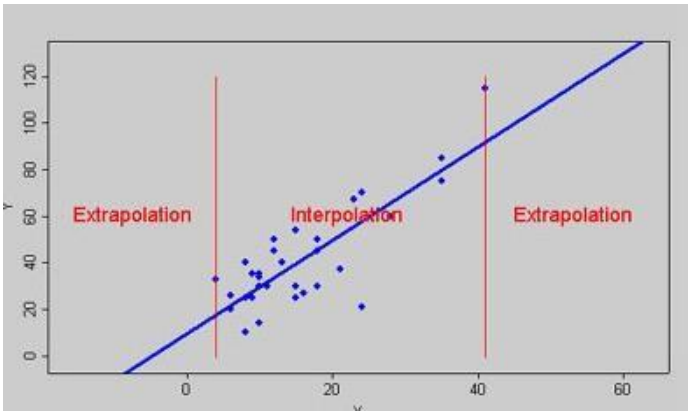


Q4. What is Interpolation and Extrapolation?

Answer:

The terms of interpolation and extrapolation are extremely important in any statistical analysis. Extrapolation is the determination or estimation using a known set of values or facts by extending it and taking it to an area or region that is unknown. It is the technique of inferring something using data that is available.

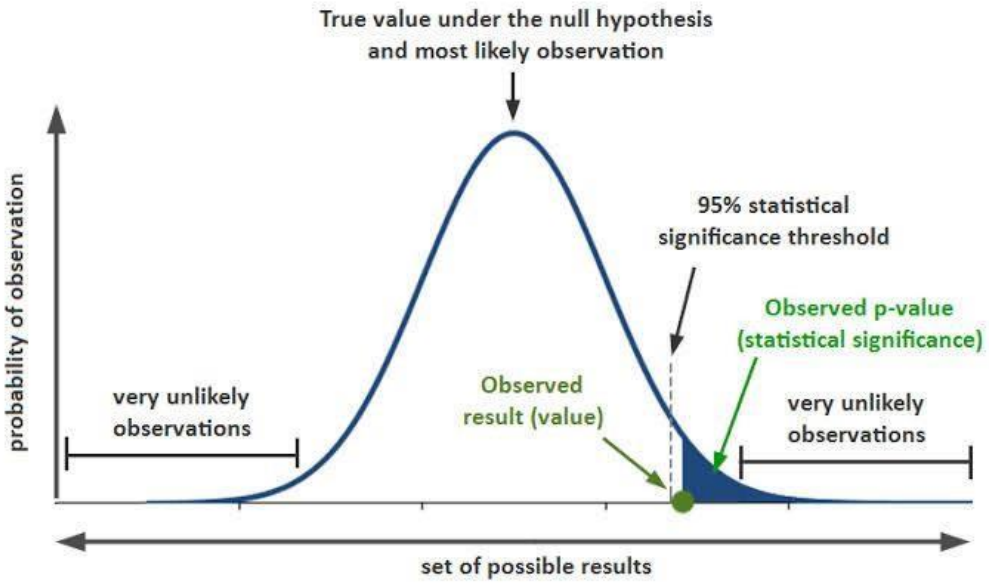
Interpolation on the other hand is the method of determining a certain value which falls between a certain set of values or the sequence of values. This is especially useful when you have data at the two extremities of a certain region, but you don't have enough data points at the specific point. This is when you deploy interpolation to determine the value that you need.



Q5. What does P-value signify about the statistical data?

Answer: P-value is used to determine the significance of results after a hypothesis test in statistics. P-value helps the readers to draw conclusions and is always between 0 and 1.

- P- Value > 0.05 denotes weak evidence against the null hypothesis which means the null hypothesis cannot be rejected.
- P-value <= 0.05 denotes strong evidence against the null hypothesis which means the null hypothesis can be rejected.
- P-value=0.05 is the marginal value indicating it is possible to go either way.



Q6. During analysis, how do you treat missing values?

Answer:

The extent of the missing values is identified after identifying the variables with missing values. If any patterns are identified the analyst must concentrate on them as it could lead to interesting and meaningful business insights. If there are no patterns identified, then the missing values can be substituted with mean or median values (imputation) or they can simply be ignored.

There are various factors to be considered when answering this question-

Understand the problem statement, understand the data, and then give the answer. Assigning a default value which can be mean, minimum, or maximum value. Getting into the data is important.

If it is a categorical variable, the default value is assigned. The missing value is assigned a default value.

If you have a distribution of data coming, for normal distribution give the mean value.

Should we even treat missing values is another important point to consider? If 80% of the values for a variable are missing, then you can answer that you would be dropping the variable instead of treating the missing values.

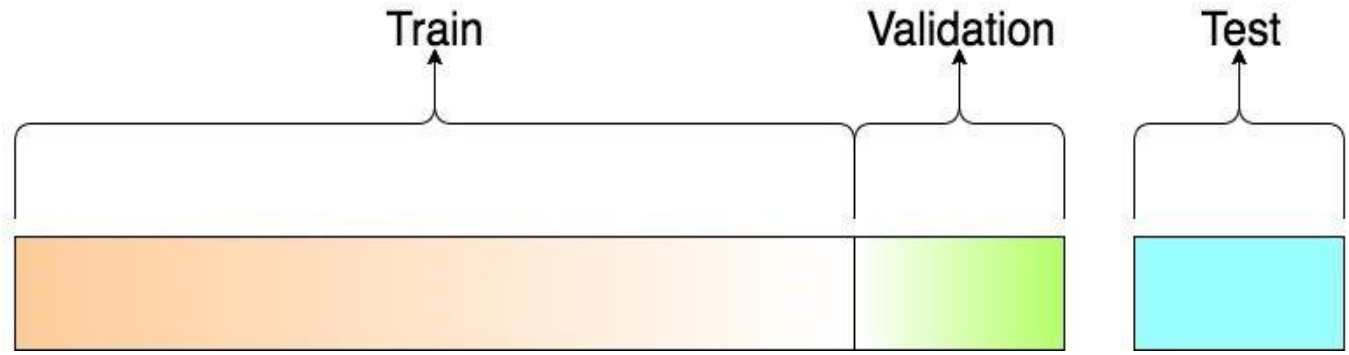
Q7. Explain the difference between a Test Set and a Validation Set?

Answer:

Validation set can be considered as a part of the training set as it is used for parameter selection and to avoid Overfitting of the model being built. On the other hand, test set is used for testing or evaluating the performance of a trained machine leaning model. In simple terms, the differences can be summarized as- Training Set is to fit the parameters i.e., weights.

Test Set is to assess the performance of the model i.e., evaluating the predictive power and generalization.

Validation set is to tune the parameters.



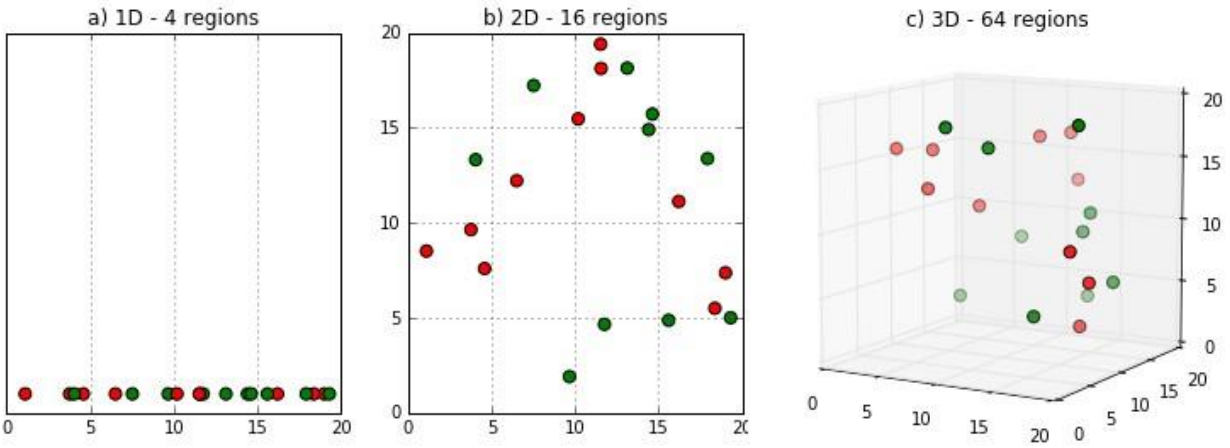
Q8. What is the curse of dimensionality? Can you list some ways to deal with it? Answer:

The curse of dimensionality is when the training data has a high feature count, but the dataset does not have enough samples for a model to learn correctly from so many features. For example, a training dataset of 100 samples with 100 features will be very hard to learn from because the model will find random relations between the features and the target. However, if we had a dataset of 100k samples with 100 features, the model could probably learn the correct relationships between the features and the target.

There are different options to fight the curse of dimensionality:

- **Feature selection.** Instead of using all the features, we can train on a smaller subset of features.
- **Dimensionality reduction.** There are many techniques that allow to reduce the dimensionality of the features. Principal component analysis (PCA) and using autoencoders are examples of dimensionality reduction techniques.
- **L1 regularization.** Because it produces sparse parameters, L1 helps to deal with highdimensionality input.

- **Feature engineering.** It's possible to create new features that sum up multiple existing features. For example, we can get statistics such as the mean or median.



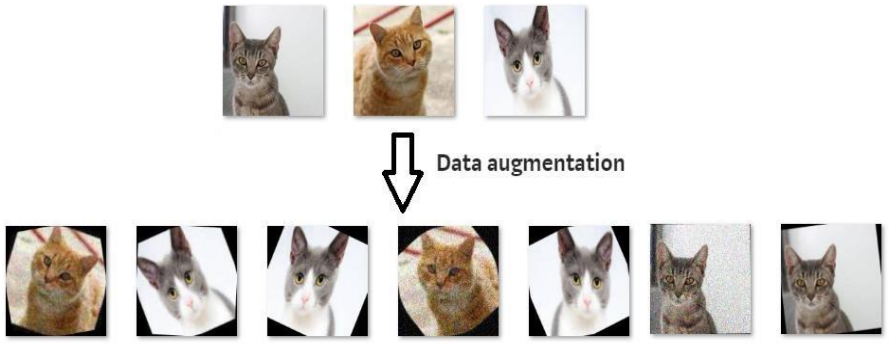
Q9. What is data augmentation? Can you give some examples?

Answer: Data augmentation is a technique for synthesizing new data by modifying existing data in such a way that the target is not changed, or it is changed in a known way.

Computer vision is one of fields where data augmentation is especially useful. There are many modifications that we can do to images:

- Resize
- Horizontal or vertical flip
- Rotate
- Add noise.
- Deform and Modify colours.

Each problem needs a customized data augmentation pipeline. For example, on OCR, doing flips will change the text and will not be beneficial; however, resizes and small rotations may help.



Q10. What is stratified cross-validation and when should we use it?

Answer: Cross-validation is a technique for dividing data between training and validation sets. On typical cross validation this split is done randomly. But in *stratified* cross-validation, the split preserves the ratio of the categories on both the training and validation datasets.

For example, if we have a dataset with 10% of category A and 90% of category B, and we use stratified cross-validation, we will have the same proportions in training and validation. In contrast, if we use simple cross-validation, in the worst case we may find that there are no samples of category A in the validation set.

Stratified cross-validation may be applied in the following scenarios:

- **On a dataset with multiple categories.** The smaller the dataset and the more imbalanced the categories, the more important it will be to use stratified cross-validation.
- **On a dataset with data of different distributions.** For example, in a dataset for autonomous driving, we may have images taken during the day and at night. If we do not ensure that both types are present in training and validation, we will have generalization problems.