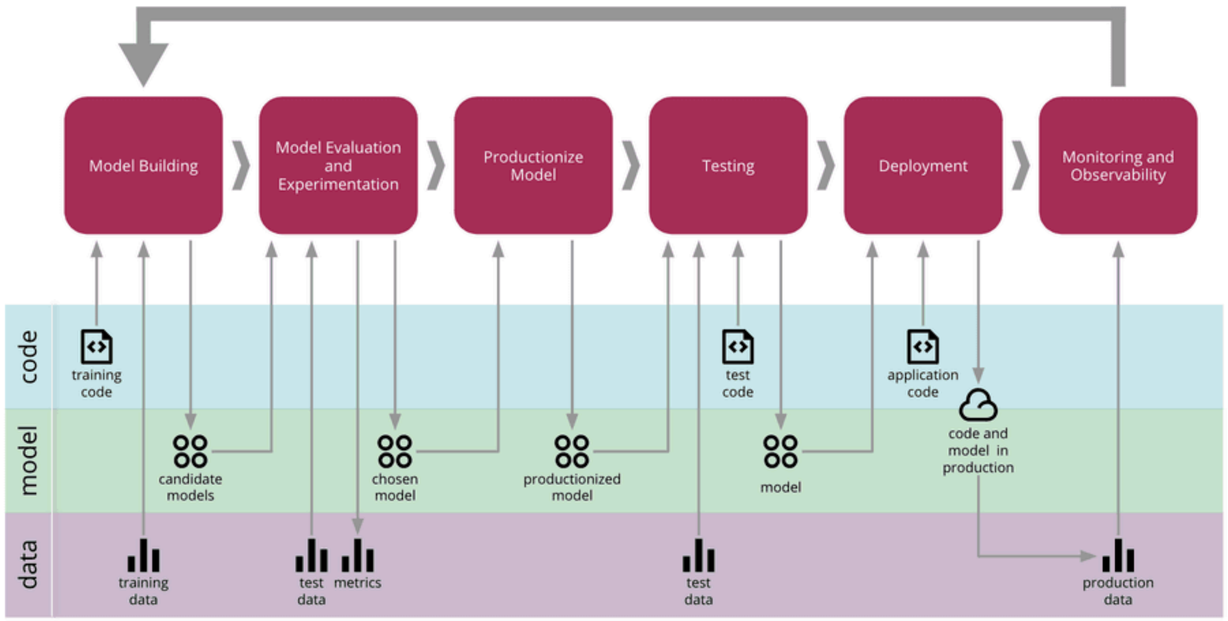
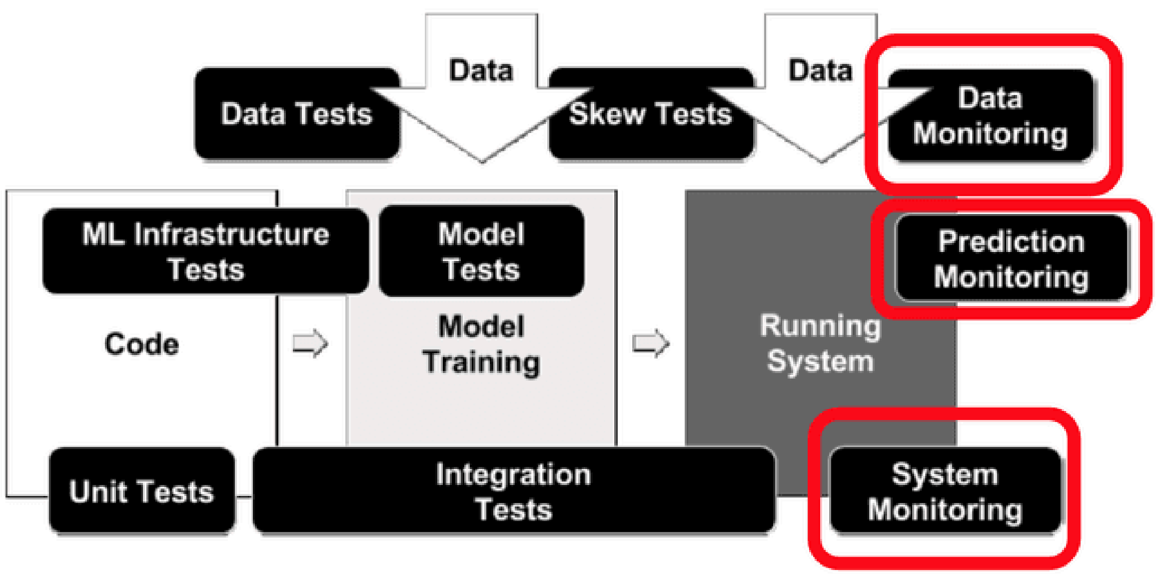
Monitoring Machine Learning Systems

Even though we've trained and thoroughly evaluated our model, the real work begins once we deploy to production. "Training and deploying ML models is relatively fast and cheap, but maintaining, monitoring and governing them over time is difficult and expensive.”



Different testing in ML



Drift

Drift is the change in an entity with respect to a baseline. Machine learning models are trained with historical data, but once they are used in the real world, they may become outdated and lose their accuracy over time due to a phenomenon called drift. Drift is when Real life data distribution changes from the one that it was trained on. This can cause the model to become less accurate or perform differently than it was designed to.

It refers to quantifying the changes in the observed data with respect to the training data.

Why does it drift

There are several reasons why machine learning models can drift over time.

- Outdated data: One common reason is simply that the data that the model was trained on becomes outdated or no longer represents the current conditions.

For example, consider a machine learning model trained to predict the stock price of a company based on historical data. If we train the model with data from a stable market, it might do well at first. However, if the market becomes more volatile over time, the model might not be able to accurately predict the stock price anymore because the statistical properties of the data have changed.

- Inefficient Model: Another reason for model drift is that the model was not designed to handle changes in the data. Some machine learning models can handle changes in the data better than others, but no model can avoid drift completely.

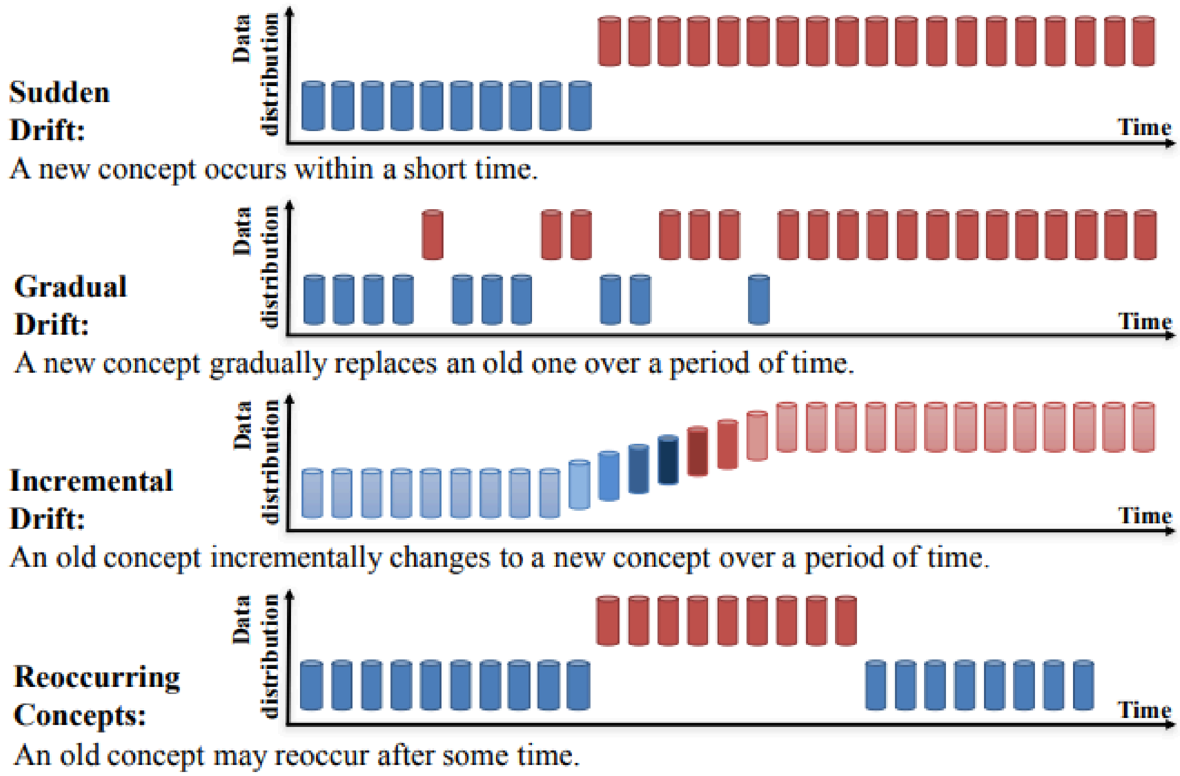
Types of drift

Two major drifts that often occur after a model has been deployed are concept drift and data drift.

Concept drift

Concept drift, also known as model drift, occurs when the task that the model was designed to perform changes over time.

It refers to a change in relationships between input data, ‘x’ and output data ‘y’ over a period of time. Eg- Consider a housing price prediction model. In an anomalous situation such as the pandemic, there is a sudden change in real estate prices and hence, the model may not make accurate predictions anymore. A home that costs 50 lakh INR in a normal scenario might now cost 75 lakh INR for the same set and quantity of features like bedrooms, ACs, area in square feet, etc.



Data drift

Data drift refers to a change in the input features. Mathematically, it is a change in the distribution of variables that causes their meaning to change.

The change in input data or independent variable leads to poor performance of the model. Microsoft has stated data drift to be one of the top reasons model accuracy degrades over time.

Data drift is generally a consequence of changes in consumer preferences over time.

For instance,

- educational data collected before Covid shows a lesser preference for online learning than post-covid.
- Similarly, the demand for lipsticks has reduced considerably after Covid while face masks became a norm. As a result,
- Models trained on previous data will be useless. Since the input data has changed, the distribution of the variables becomes different and confuses the model.
- Data drift, feature drift, population, or covariate shift. Quite a few names to describe essentially the same thing.

Reasons for data drift

- changes in the data collection process
- changes in the data sources which provide the inputs
- changes in the business needs.
- Change in the distribution of the label/features in the data

Methods Available to detect Data Drift:

There are several methods available for detecting feature drift in machine learning:

- **Visual inspection:** One simple method is to visually inspect the input data over time to see if there are any noticeable changes in the distribution of the features. This can be done by plotting histograms or scatter plots of the data at different times and comparing them.
- **Statistical tests:** Statistical tests such as the Chi-Squared Test can be used to compare the input data distribution at different times and detect significant differences. These tests can provide a quantitative measure of the degree of drift in the data.
- **Model performance monitoring:** Another approach is to monitor the performance of the model over time and look for significant changes in accuracy or other performance metrics. If the model's performance begins to degrade, this could indicate that there is a drift in the data.
- **Data quality checks:** Regularly checking the quality of the input data can also help detect feature drift. For example, if there are sudden changes in the range, mean, median or variance of the features, this could indicate that there is a drift in the data.

The Kolmogorov-Smirnov Test

The KS test is a test of the equality between two univariate probability distributions. It can be used to compare a sample with a reference probability distribution or compare two samples.

When comparing two samples, we are trying to answer the following question:

"What is the probability that these two sets of samples were drawn from the same probability distribution?"

The null hypothesis is that the two samples come from the same distribution. The KS-test is applied to reject or accept it.

It returns the p-value. If the p-value is less than 0.05, you can usually declare that there is strong evidence to reject the null hypothesis and consider the two samples different.

You can also set a different significance level and, for example, react only to p-values less than 0.01. It is good to remember that the p-value is not a "measurement" of drift size but a declaration of the statistical test significance.

Chi-square

Chi-square test in another popular divergence test well-suited for categorical data.

The chi square statistic is a statistical hypothesis testing technique to test how two distributions of categorical variables are related to each other. Specifically, the chi-square statistic is a single number that quantifies the difference between the observed counts versus the counts that are expected if there was no relationship between the variables at all.

The divergence can range from zero to infinity. A value of zero means there is no difference between the data sets.

The null and alternative hypotheses of the Chi-Square Test of Fit Test are:

Null Hypothesis: The null hypothesis is: two groups have no significant difference.