UNIT-II (2nd PART)

MULTICOLLINEARITY

When the data set has a large number of independent variables, it is possible that few of these independent variables may be highly correlated. Existence of high correlation between independent variables is called multi-collinearity. Presence of multi-collinearity can destabilize the multiple regression model. Thus, it is necessary to identify the presence of multi-collinearity and take corrective actions. Multi-collinearity can have the following impact on the model:

1. The standard error of estimate of a regression coefficient may be inflated, and may result in re5aininjg of null hypothesis in t-test, resulting in rejection of a statistically significant explanatory variable. The t-statistics value is ($\hat{\beta}/S_e(\hat{\beta})$). If $S_e(\hat{\beta})$ is inflated, then the t-value will be underestimated resulting in high p-value that may result in failing to reject the null hypothesis. Thus, it is possible that a statistically significant explanatory variable may be labelled as statistically insignificant due to the presence of multi-collinearity.
2. The sign of the regression coefficient may be different, that is, instead of negative value for regression coefficient, we may have a positive regression coefficient and vice-versa.
3. Adding/ removing a variable or even an observation may result in large variation in regression coefficients estimates

https://www.investopedia.com/terms/m/multicollinearity.asp#:~:text=Multicollinearity%20is%20a%20statistical%20concept,in%20less%20reliable%20statistical%20inferences.

Variance Inflation Factor (VIF)

VIF measures the magnitude of multi-collinearity. Let us consider a regression model with two explanatory variables defined as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \qquad (2.2)$$

To find whether there is multi-collinearity, we develop a regression model between the two explanatory variables as follows:

$$X_1 = \alpha_0 + \alpha_1 X_2 \qquad (2.3)$$

Let $R^2{}_{12}$ be the R-square value for the regression model in (2.3). Note that when there are only two independent variables, we can use correlation coefficient to calculate R^2 between the two variables. VIF is then given by-

$$VIF = \frac{1}{1 - R_{12}^2} \qquad (2.4)$$

The value $1-R_{12}^2$ is called the **tolerance** $\sqrt{VIF}$ is the value by which the standard error of estimate is inflated in the presence of multi-collinearity or $\sqrt{VIF}$ is the value by which the t-statistic is deflated. So, the actual t-value is given by-

$$t_{actual} = \left(\frac{\hat{\beta}_1}{S_e\left(\hat{\beta}_1\right)} \right) \times \sqrt{VIF} \qquad (2.5)$$

There will be some correlation between explanatory variables in almost all cases, thus the value of VIF is likely to be more than one. The **threshold value** for VIF is 4. VIF value of greater than 4 requires further investigation to assess the impact of multi-collinearity. Before building the multiple regression models, it is advised to check the correlation between different explanatory variables for potential multi-collinearity. VIF value equal to 4 implies that the t-statistics value is deflated by a factor 2 and thus there will be a significant increase in the corresponding p-value. The serious impact of multi-collinearity is that it can change the sign of the regression coefficients (for eg., instead of +ve, the model may have -ve regression coefficient for a predictor and vice-versa).

Remedies for handling Multi-Collinearity:

There are many approaches that can be used to handle the impact of multi-collinearity. One easier approach is to remove one of the variables from the model building. For example, the data scientist may remove a variable that is either difficult or expensive to collect. Another approach suggested by researchers is to use centered variables, that is, use $\left(X_i - \bar{X}_i\right)$ instead of X_i. When there are many variables in the data, the data scientists can use **Principal Component Analysis (PCA)** to avoid multi-collinearity. PCA will create orthogonal components and thus remove potential multi-collinearity. In the recent years, authors use advanced regression models such as **Ridge Regression** and **LASSO regression** to handle multi-collinearity.

https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/

AUTO-CORRELATION

Auto-correlation is the correlation between successive error terms in a time-series data. Consider a time-series model as defined below-

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t \qquad (2.6)$$

In the regression model (2.6), the values of the response variable Y are measured at different time points *t* and X_t, is the value of the independent variable at time *t*. One of the assumptions of regression model is that, there should be no correlation between error terms, ε_t and ε_{t-1} (known as auto-correlation of errors pf lag 1). In general, errors ε_t and ε_{t-k} may be correlated (known as auto-correlation of lag k). If there is an auto-correlation, the standard error estimate of the beta coefficient may be underestimated and that will result in overestimation of the *t*-statistic value, which, in turn, will result in low *p*-value. Thus, a variable that has no statistically significant relationship with the response variable may be accepted in the model due to the presence of auto-correlation. The presence of auto-correlation can be established using Durbin-**Watson test.**

https://www.investopedia.com/terms/a/autocorrelation.asp#:~:text=Autocorrelation%20represents%20the%20degree%20of,value%20and%20its%20past%20values.

How to measure Auto-correlation?

https://online.stat.psu.edu/stat501/lesson/14/14.3

HETEROSCEDASTICITY

https://www.geeksforgeeks.org/heteroscedasticity-in-regression-analysis/

https://www.investopedia.com/terms/h/heteroskedasticity.asp

https://statisticsbyjim.com/regression/heteroscedasticity-regression/

https://corporatefinanceinstitute.com/resources/knowledge/other/heteroskedasticity/

UNIT-III LOGISTIC REGRESSION AND PANEL DATA MODEL

UNIT-III LOGISTIC REGRESSION AND PANEL DATA MODELS

Introduction to Binary Logistic Regression:

Logistic Regression is a statistical model in which the response variable takes a discrete value and the explanatory variables can either be continuous or discrete. Logistic regression is one of the supervised learning algorithms.

Assume that the value of Y is either 1 (conventionally known as positive outcome) or 0 (conventionally known as negative outcome). When there are more than two values of Y , then multinomial logistic regression model is used. The binary logistic regression model is given by-

$$
\begin{aligned}
&P\left(Y=1\right)=\frac{e^{z}}{1+e^{z}}\\
&where \qquad\qquad\qquad\qquad\qquad\qquad (3.1)\\
&Z=\beta_0+\beta_1 X_1+\beta_2 X_2+...+\beta_m X_m
\end{aligned}
$$

Here, $X_1, X_2, ...X_m$ are the independent variables. The right-hand side of the above equation is a logistic function. One of the objectives of classification problems is to predict the class probability, that is the probability of an observation belonging to class labelled as 1, that is P(Y=1). Logistic function is a probability function, and has an S-shaped curve.

The logistic regression function defined in the above equation can be transformed as follows-

$$\frac{P\left(Y=1\right)}{1-P\left(Y=1\right)}=e^{z} \qquad (3.2)$$

The above equation can also be written as-

$$\ln\left(\frac{P\left(Y=1\right)}{1-P\left(Y=1\right)}\right)=Z=\beta_0+\beta_1 X_1+\beta_2 X_2+...+\beta_m X_m \qquad (3.3)$$

Now, $\frac{P\left(Y=1\right)}{1-P\left(Y=1\right)}$ is odds. Thus, the left-hand side of (3.3) is log-natural of odds. Equation (3.3) is known as logit (logistic probability unit) function. The left-hand side of the logit function is continuous function and the right-hand side is a linear function.

How is it different from Simple Linear Regression?

Linear Regression is a machine learning algorithm based on **supervised regression algorithm**. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between the dependent and independent variables, they are considering and the number of independent variables being used. **Logistic regression** is basically a **supervised classification algorithm**. In a classification problem, the target variable(or output), y, can take only discrete values for a given set of features(or inputs), X.

Linear Regression	**Logistic Regression**
Linear Regression is a supervised regression model.	Logistic Regression is a supervised classification model.
In Linear Regression, we predict the value by an integer number.	In Logistic Regression, we predict the value by 1 or 0.
Here no activation function is used.	Here activation function is used to convert a linear regression equation to the logistic regression equation
Here no threshold value is needed.	Here a threshold value is added.
Here we calculate Root Mean Square Error(RMSE) to predict the next weight value.	Here we use precision to predict the next weight value.
Here dependent variable should be numeric and the response variable is continuous to value.	Here the dependent variable consists of only two categories. Logistic regression estimates the odds outcome of the dependent variable given a set of quantitative or categorical independent variables.
It is based on the least square estimation.	It is based on maximum likelihood estimation.
Here when we plot the training datasets, a straight line can be drawn that touches maximum plots.	Any change in the coefficient leads to a change in both the direction and the steepness of the logistic function. It means positive slopes result in an S-shaped curve and negative slopes result in a Z-shaped curve.
Linear regression is used to estimate the dependent variable in case of a change in independent variables. For example, predict the price of houses.	Whereas logistic regression is used to calculate the probability of an event. For example, classify if tissue is benign or malignant.
Linear regression assumes the normal or gaussian distribution of the dependent variable.	Logistic regression assumes the binomial distribution of the dependent variable.

Logistic Regression Model Diagnostics:

We have to carry out diagnostics tests before a binary logistic regression model can be accepted for deployment. Statistical significance of logistic regression model is checked using likelihood ratio test (Omnibus test). Wald’s test and Hosmer and Lemeshow test for deployment. We discuss some of these tests below:

1. **Omnibus Test:** Omnibus tests are generic statistical tests used for checking whether the variance explained by the model is more than the unexplained variance. For example, in MLR model, F-test is an omnibus test. F-test in MLR compares the explained variation

with unexplained variation. In the case of logistic regression, this is achieved using likelihood ratio test. Likelihood ratio test usually compares two likelihood functions: one without any independent variable and the other with independent variable. Likelihood ratio test is a chi-square test with degrees of freedom equal to the number of independent variables in the model. Note that a likelihood test can also be used for model comparison; in that case the degrees of freedom will be the difference in the number of independent variables between the models.

2. **Wald's Test:** Wald's test is used for checking whether an individual explanatory variable is statistically significant. Wald's test is chi-square test.
3. **Hosmer-Lemeshow Test:** It is a chi-square goodness of fit test for binary logistic regression.
4. **Pseudo R^2 :** Pseudo R^2 is a measure of the goodness of the model. It is called pseudo R^2 because it does not have the same interpretation of R^2 in the MLR model.

Sensitivity, specificity and precision: In logistic regression, the model performance is often measured using concepts such as sensitivity, specificity and precision. The ability of the model to correctly classify positives and negatives are called sensitivity and specificity respectively. The terminologies sensitivity and specificity originated in medical diagnostics. In medical diagnostics, **sensitivity** (also known as true positive rate) measures the ability of a diagnostic test to identify disease if it is present in a patient (test positive).

That is-

$$\text{Sensitivity} = \text{P}(\text{diagnostic test is positive} \mid \text{patient has disease })$$

In general case,

$$\text{Sensitivity} = (\text{model classifies } \text{Y}_\text{i} \text{ as positive} \mid \text{Y}_\text{i} \text{ is positive})$$

Sensitivity is calculated using the following equation:

$$\text{Sensitivity} = \frac{\text{True Positive (TP)}}{\text{True Positive (TP)+False Negative(FN)}}$$

where True Positive (TP) is the number of positives correctly classified as positives by the model and False Negative (FN) is positives misclassified as negative by the model. Sensitivity is also called as **recall**.

Specificity is the ability of the diagnostic test to correctly classify the test as negative when the disease is not present. That is:

$$\text{Specificity} = \text{P}(\text{diagnostic test is negative} \mid \text{patient has no disease })$$

In general case,

$$\text{Specificity} = (\text{model classifies } \text{Y}_\text{i} \text{ as negative} \mid \text{Y}_\text{i} \text{ is negative})$$

Sensitivity is calculated using the following equation:

$$\text{Specificity} = \frac{\text{True Negative(TN)}}{\text{True Negative (TN)+False Positive(FP)}}$$

Where True Negative (TN) is number of the negatives correctly classified as negatives by the model and False Positive (FP) is number of negatives misclassified as positives by the model.

Precision measures the accuracy of positive classified by the model

$$\text{Precision} = \text{P(patient has disease}|\text{diagnostic test is positive)}$$

$$\text{Precision} = \frac{\text{True Positive(TP)}}{\text{True Positive (TP)+False Positive(FP)}}$$

https://www.graphpad.com/guides/prism/latest/curve-fitting/reg_logistic_roc_curves.htm#:~:text=ROC%20curves%20in%20logistic%20regression,logistic%20regression%20is%20shown%20below.

https://towardsdatascience.com/understanding-the-roc-curve-in-three-visual-steps-795b1399481c

https://www.aptech.com/blog/introduction-to-the-fundamentals-of-panel-data/

https://vitalflux.com/fixed-vs-random-vs-mixed-effects-models-examples/

UNIT-IV FORECASTING

https://www.tableau.com/learn/articles/time-series-forecasting#:~:text=Time%20series%20forecasting%20occurs%20when,drive%20future%20strategic%20decision%2Dmaking.

Time-series data is a data on a response variable, Y_t, such as demand for a spare parts of a capital equipment or a product or a service or market share of a brand observed at different time points t Whenever we have a forecasting problem, we will be using a time-series data. The variable Y_t is a random variable. The data points or measurements are usually collected at regular intervals and are arranged in chronological order. If the time-series data contains observations of just a single variable (such as demand of a product at time t), then it is termed as Univariate Time-Series. If the data consists of more than one variable for e.g. Demand for a product at time t, price at time t, amount of money spent by the company on promotion at time t, competitor's price at time t, etc. then it is called Multi-variate Time series data. From a forecasting perspective, a time-series data can be broken into the following components:

1. **Trend (Secular Trend or Long-Term Variation)** It is a longer-term change. Here we take into account the number of observations available and make a subjective assessment of what is long term. It represents a relatively smooth, steady, and gradual movement of a time series in the same direction. To understand the meaning of the long term, consider the climate variables. These variables sometimes exhibit cyclic variation over a very long time period such as 50 years. If one just had 20 years of data, this long-term oscillation would appear to be a trend, but if several hundreds of years of data are available, then long term oscillations would be visible. These movements are systematic in nature where the movements are broad, steady, showing a slow rise or fall in the same direction. The trend may be linear or non-linear (curvilinear). Some examples of secular trends are:

    - Increase in prices,
    - Increase in pollution,
    - an increase in the need for wheat,
    - an increase in literacy rate,
    - the decrease in deaths due to advances in science. Taking averages over a certain period is a simple way of detecting a trend in seasonal data. Change in averages with time is evidence of a trend in the given series. There are more formal tests for detecting a trend in time series.

2. **Seasonal effect (Seasonal Variation or Seasonal Fluctuations)** Many of the time series data exhibits a seasonal variation which is the annual period, e.g., sales and temperature readings. Seasonal variations are easy to understand and can be easily measured or removed from the data to give deseasonalized data. Seasonal Fluctuations describes any regular variation with a period of less than one year. For example, the cost of various types of fruits and vegetables, clothes, unemployment figures, average daily rainfall, increase in the sale of tea in winter, increase in the sale of ice cream

in summer, etc., all show seasonal variations. The changes which repeat themselves within a fixed period, are also called seasonal variations. For example, traffic on roads in the morning and evening hours, Sales at festivals like EID, etc., increase in the number of passengers at the weekend, etc. Seasonal variations are caused by climate, social customs, religious activities, etc. The main causes of seasonal variations are seasons, religious festivals, and social customs.

3. **Cyclical Variation or Cyclic Fluctuations**
   Time series exhibits Cyclical Variations at a fixed period due to some other physical cause, such as daily variation in temperature. Cyclical variation is a non-seasonal component that varies in a recognizable cycle. These variations are considered a more dangerous effect on business and economic activity. Sometimes series exhibits oscillation which does not have a fixed period but is predictable to some extent. For example, economic data affected by business cycles with a period varying between about 5 and 7 years. In weekly or monthly data, the cyclical component may describe any regular variation (fluctuations) in time series data. The cyclical variation is periodic in nature and repeats itself like a business cycle, which has four phases (i) Peak/Prosperity (ii) Recession (iii) Trough/Depression (iv) Expansion.

4. **Other Irregular Variation (Irregular Fluctuations)**
   When trend and cyclical variations are removed from a set of time series data, the residual left, which may or may not be random. Various techniques for analyzing series of this type examine to see "if irregular variation may be explained in terms of probability models such as moving average or autoregressive models, i.e. we can see if any cyclical variation is still left in the residuals. These variations occur due to sudden causes are called residual variation (also called accidental or erratic fluctuations) and are unpredictable. For example, a rise in prices of steel due to strike in the factory, accident due to failure of the break, flood, earth quick, and war, etc.

https://itfeature.com/time-series-analysis-and-forecasting/components-of-time-series

https://www.bmc.com/blogs/machine-learning-vs-predictive-analytics/

https://www.techtarget.com/searchenterpriseai/definition/machine-learning-ML#:~:text=Machine%20learning%20(ML)%20is%20a,to%20predict%20new%20output%20values.

https://data-flair.training/blogs/artificial-neural-networks-for-machine-learning/

https://www.datasciencecentral.com/artificial-neural-network-ann-in-machine-learning/

https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/

UNIT-V DATA MINING AND SIMULATION

Data Mining:

The concept of data mining has been with us since long before the digital age. The idea of applying data to knowledge discovery has been around for centuries, starting with manual formulas for statistical modeling and regression analysis. In the 1930s, Alan Turing introduced the idea of a universal computing machine that could perform complex computations. This marked the rise of the electromechanical computer — and with it, the ever-expanding explosion of digital information that continues to this very day.

We’ve come a long way since then. Data has become a part of every facet of business and life. Companies today can harness data mining applications and machine learning for everything from improving their sales processes to interpreting financials for investment purposes. As a result, data scientists have become vital to organizations all over the world as companies seek to achieve bigger goals than ever before.

Data mining is the process of analyzing massive volumes of data to discover business intelligence that can help companies solve problems, mitigate risks, and seize new opportunities. This branch of data science derives its name from the similarities between the process of searching through large datasets for valuable information and the process of mining a mountain for precious metals, stones, and ore. Both processes require sifting through tremendous amounts of raw material to find hidden value.

Data mining can answer business questions that were traditionally impossible to answer because they were too time-consuming to resolve manually. Using powerful computers and algorithms to execute a range of statistical techniques that analyze data in different ways, users can identify patterns, trends, and relationships they might otherwise miss. They can then apply these findings to predict what is likely to happen in the future and take action to influence business outcomes.

Data mining is used in many areas of business and research, including sales and marketing, product development, healthcare, and education. When used correctly, data mining can give you an advantage over competitors by making it possible to learn more about customers, develop effective marketing strategies, increase revenue, and decrease costs.

How data mining works

Any data mining project must start by establishing the business question you are trying to answer. Without a clear focus on a meaningful business outcome, you could find yourself poring over the same set of data over and over without turning up any useful information at all. Once you have clarity on the problem you are trying to solve, it’s time to collect the right data to answer it — usually by ingesting data from multiple sources into a central data lake or data warehouse — and preparing that data for analysis.

Success in the later phases is dependent on what occurs in the earlier phases. Poor data quality will lead to poor results, which is why data miners must ensure the quality of the data they use as input for analysis.

For a successful data mining process that delivers timely, reliable results, you should follow a structured, repeatable approach. Ideally, that process will include the following six steps:

1. **Business understanding.** Develop a thorough understanding of the project parameters, including the current business situation, the primary business objective of the project, and the criteria for success.
2. **Data understanding**. Determine the data that will be needed to solve the problem and gather it from all available sources.
3. **Data preparation.** Get the data ready for analysis. This includes ensuring that the data is in the appropriate format to answer the business question, and fixing any data quality problems such as missing or duplicate data.
4. **Modeling.** Use algorithms to identify patterns within the data and apply those patterns to a predictive model.
5. **Evaluation**. Determine whether and how well the results delivered by a given model will help achieve the business goal. There is often an iterative phase in which the algorithm is fine-tuned in order to achieve the best result.
6. **Deployment**. Run the analysis and make the results of the project available to decision makers.

Throughout this process, close collaboration between domain experts and data miners is essential to understand the significance of data mining results to the business question being explored.

Advantages of data mining

Data is pouring into your businesses every day from a dazzling array of sources, in a multitude of formats, and at unprecedented speed and volumes. Deciding whether or not to be a data-driven business is no longer an option; your business' success depends on how quickly you can discover insights from big data and incorporate them into business decisions and processes to drive better actions across your enterprise. However, with so much data to manage, this can seem like an insurmountable task.

Data mining gives businesses an opportunity to optimize operations for the most likely future by understanding the past and present, and making accurate predictions about what is likely to happen next.

For example, sales and marketing teams can use data mining to predict which prospects are likely to become profitable customers. Based on past customer demographics, they can establish a profile of the type of prospect who would be most likely to respond to a specific offer. With this knowledge, they can increase return on investment (ROI) by targeting only those prospects likely to respond and become valuable customers.

You can use data mining to solve almost any business problem that involves data, including:

- Increasing revenue
- Understanding customer segments and preferences
- Acquiring new customers
- Improving cross-selling and up-selling
- Retaining customers and increasing loyalty
- Increasing ROI from marketing campaigns
- Detecting and preventing fraud
- Identifying credit risks
- Monitoring operational performance

Through the application of data mining techniques, decisions can be based on real business intelligence — rather than instinct or gut reactions — and deliver consistent results that keep businesses ahead of the competition.

As large-scale data processing technologies such as machine learning and artificial intelligence become more readily accessible, companies are now able to automate these processes to dig through terabytes of data in minutes or hours, rather than days or weeks, helping them innovate and grow faster.

Data mining use cases and examples

Organizations across industries are achieving transformative results from data mining:

- **Groupon aligns marketing activities** — One of Groupon's key challenges is processing the massive volume of data it uses to provide its shopping service. Every day, the company processes more than a terabyte of raw data in real time and stores this information in various database systems. Data mining allows Groupon to align marketing activities more closely with customer preferences, analyzing that 1 terabyte of customer data in real time and helping the company identify trends as they emerge.
- **Air France KLM caters to customer travel preferences** — The airline uses data mining techniques to create a 360-degree customer view by integrating data from trip searches, bookings, and flight operations with web, social media, call center, and airport lounge interactions. They use this deep customer insight to create personalized travel experiences.
- **Domino's helps customers build the perfect pizza** — The largest pizza company in the world collects 85,000 structured and unstructured data sources, including point of sales systems and 26 supply chain centers, and through all its channels, including text messages, social media, and Amazon Echo. This level of insight has improved business performance while enabling one-to-one buying experiences across touchpoints.

These are just a few examples of how data mining capabilities can help data-driven organizations increase efficiency, streamline operations, reduce costs, and improve profitability.

Key data mining concepts

Achieving the best results from data mining requires an array of tools and techniques. Some are probably already familiar, but others might be new to you. Here are a few of the most common terms and concepts in the field of data mining.

Data processes

The first batch of concepts relate to the data itself, and how it is moved and managed.

- **Data cleansing and preparation**. Raw data flows in from any number of sources in a wild mix of formats and quality. Before it can be used in any meaningful way, that data must be transformed from its raw state into a format that's more suitable for analysis and processing. This includes processes such as identifying and removing errors, calling out missing data, and flagging outliers.
- **Data warehousing**. Unless you are working with only a small subset of data, you will probably need to collect data from a range of sources combine it into a single data repository before you can use data to make decisions. This repository is generally known as a data warehouse. It is the foundational component of most large-scale data mining efforts.
- **Data analytics**. Once your data has been cleaned and collected, you can start examining it for past trends that could be applied to future decision-making. The process of evaluating historical digital information to provide useful business intelligence is known as data analytics.
- **Predictive analytics**. Where data analytics looks to the past to identify trends, predictive analytics uses that data to anticipate future outcomes. Predictive analytics relies on data modeling, machine learning, and artificial intelligence to uncover patterns in big data.

Computer science concepts

Next, you should be familiar with some common computer science terms that describe how various programs and algorithms interact with the data to deliver meaningful insights.

- **Artificial intelligence (AI)**. With modern technology, automated systems can perform analytical activities that used to be possible only by applying human intelligence. These activities can include things like planning, learning, reasoning, and problem solving. When it comes to data mining, this refers to using a computer program to identify meaningful trends in the data.
- **Machine learning (ML)**. The earliest computers needed an explicit program to instruct them through any process, step by step — but that assumes that the programmer is already aware of every possible scenario that may arise. More recently, programmers

use statistical probabilities to write machine learning algorithms that give computers the ability to "learn" and adapt without being explicitly programmed.

- **Natural language processing (NLP).** Many valuable data sources, such as social media, aren't easily broken down into simple fields. Natural language processing is a feature of AI that gives a computer program the ability to "read" and understand casual or unstructured data sources.
- **Neural networks**. Sometimes a single machine learning algorithm isn't powerful enough to do the job alone. A neural network is a collection of algorithms that work together to solve more complex problems, thinking more like a human brain. Just like a simple machine learning algorithm, neural networks have the ability to learn and adapt.

Data mining techniques

There are many techniques used by data mining technology to make sense of your business data. Here are a few of the most common:

- **Association rule learning**. Also known as market basket analysis, association rule learning looks for interesting relationships between variables in a dataset that might not be immediately apparent, such as determining which products are typically purchased together. This can be incredibly valuable for long-term planning.
- **Classification**. This technique sorts items in a dataset into different target categories or classes based on common features. This allows the algorithm to neatly categorize even complex data cases.
- **Clustering**. To help users understand the natural groupings or structure within the data, you can apply the process of partitioning a dataset into a set of meaningful sub-classes called clusters. This process looks at all the objects in the dataset and groups them together based on similarity to each other, rather than on predetermined features.
- **Decision trees**. Another method for categorizing data is the decision tree. This method asks a series of cascading questions to sort items in the dataset into relevant classes.
- **Regression**. This technique is used to predict a range of numeric values, such as sales, temperatures, or stock prices, based on a particular data set.

The future of data mining

We are living in a world of data. The volume of data that we create, copy, use, and store is growing exponentially. We've already crossed the threshold of creating 1.7 megabytes of new information every second for every human being on the planet.

That means that the future is bright for data mining and data science. With so much data to sort through, we are going to need ever more sophisticated methods and models to draw meaningful insights and fuel business decision making.

Just like mining techniques have evolved and improved because of improvements in technology, so too have technologies to extract valuable insights out of data. Once upon a time, only organizations like NASA could use their supercomputers to analyze data — the cost of storing and computing data was just too great. Now, companies are doing all sorts of interesting things with machine learning, artificial intelligence, and deep learning with cloud-based data lakes.

For example, the Internet of Things (IoT) and wearable technology have turned people and devices into data-generating machines that can yield unlimited insights about people and organizations — if companies can collect, store, and analyze the data fast enough.

By 2020, there were already more than 20 billion connected devices on the Internet of Things. The data generated by this activity will be available on the cloud, creating an urgent need for flexible, scalable analytics tools that can handle masses of information from disparate datasets.

With data pouring in from sales, marketing, the web, production and inventory systems, and more, cloud-based analytics solutions are making it more practical and cost-effective for organizations to access massive data and computing resources. Cloud computing helps companies accelerate data collection, compile, and prepare that data, then analyze it and act on it to improve outcomes.

Open source data mining tools also afford users new levels of power and agility, meeting analytical demands in ways many traditional solutions cannot and offering extensive analyst and developer communities where users can share and collaborate on projects. In addition, advanced technologies such as machine learning and AI are now within reach for just about any organization with the right people, data, and tools.

Data mining software and tools

There is no doubt that data mining has the power to transform enterprises; however, implementing a solution that meets the needs of all stakeholders can frequently stall platform selection. The wide range of options available to analysts, including open source languages such as R and Python and familiar tools like Excel, combined with the diversity and complexity of tools and algorithms, can further complicate the process.

Businesses that gain the most value from data mining typically select a platform that meets the following criteria:

- It incorporates best practices for their industry or type of project — for example, healthcare organizations have different needs than e-commerce companies.
- It manages the entire data mining lifecycle, from data exploration to production.
- It aligns with all enterprise applications, including BI systems, CRM, ERP, financial systems, and other enterprise software.
- It integrates with leading open source languages, providing developers and data scientists with the flexibility and collaboration tools to create innovative applications.

- It meets the needs of IT, data scientists, and analysts, while also serving the reporting and visualization needs of business users.

The Talend Big Data Platform provides a complete suite of data management and data integration capabilities to help data mining teams respond more quickly to the needs of their business.

Based on an open, scalable architecture and with tools for relational databases, flat files, cloud apps, and platforms, this solution complements your data mining platform by putting more data to work in less time — which translates into faster time to insight for a competitive advantage.

https://www.talend.com/resources/data-mining-techniques/

https://www.geeksforgeeks.org/ml-classification-vs-clustering/

https://wisdomplexus.com/blogs/data-mining-clustering-vs-classification/

https://leanbi.ch/en/blog/modelling-of-causes-and-effects/

https://www.juran.com/blog/the-ultimate-guide-to-cause-and-effect-diagrams/

A simulation imitates the operation of real-world processes or systems with the use of models. The model represents the key behaviors and characteristics of the selected process or system while the simulation represents how the model evolves under different conditions over time.

Simulations are usually computer-based, using a software-generated model to provide support for the decisions of managers and engineers as well as for training purposes. Simulation techniques aid understanding and experimentation, as the models are both visual and interactive.

Simulation systems include discrete event simulation, process simulation and dynamic simulation. Businesses may use all of these systems across different levels of the organisation.

What Does it Mean?

A simulation is a model that mimics the operation of an existing or proposed system, providing evidence for decision-making by being able to test different scenarios or process changes. This can be coupled with virtual reality technologies for a more immersive experience.

Simulations can be used to tune up performance, optimise a process, improve safety, testing theories, training staff and even for entertainment in video games! Scientifically modelling systems allows a user to gain an insight into the effects of different conditions and courses of action.

Simulation can also be used when the real system is inaccessible or too dangerous to assess or when a system is still in the design or theory stages.

Key to any simulation is the information that is used to build the simulation model and protocols for the verification and validation of models are still being researched and refined, particularly with regard to computer simulation.

How Simulation Works

Simulation works through the use of intuitive simulation software to create a visual mock-up of a process. This visual simulation should include details of timings, rules, resources and constraints, to accurately reflect the real-world process.

This can be applied to a range of scenarios, for example, you can model a supermarket and the likely behaviours of customers as they move around the shop as it becomes busier. This can inform decisions including staffing requirements, shop floor layout, and supply chain needs.

Another example would be a manufacturing environment where different parts of the line can be simulated to assess how their processes interact with those of others. This can provide an overview of how the entire system will perform in order to devise innovative methods to improve performance.

Advantages

There are a range of advantages to be gained through the use of simulation, including:

1. Less Financial Risk

Simulation is less expensive than real life experimentation. The potential costs of testing theories of real world systems can include those associated with changing to an untested process, hiring staff or even buying new equipment. Simulation allows you to test theories and avoid costly mistakes in real life.

2. Exact Repeated Testing

A simulation allows you to test different theories and innovations time after time against the exact same circumstances. This means you can thoroughly test and compare different ideas without deviation.

3. Examine Long-Term Impacts

A simulation can be created to let you see into the future by accurately modelling the impact of years of use in just a few seconds. This lets you see both short and long-term impacts so you can confidently make informed investment decisions now that can provide benefits years into the future.

4. Gain Insights for Process Improvement

The benefits of simulation are not only realised at the end of a project. Improvements can be integrated throughout an entire process by testing different theories.

5. Assess Random Events

A simulation can also be used to assess random events such as an unexpected staff absence or supply chain issues.

6. Test Non-Standard Distributions

A simulation can take account of changing and non-standard distributions, rather than having to repeat only set parameters. For example, when simulating a supermarket you can input different types of customer who will move through the shop at different speeds. A young businesswoman who is picking up a sandwich will move through the shop differently from an old couple or a mother doing a weekly shop with two children in tow. By taking such changing parameters into account, a simulation can more accurately mimic the real world.

7. Encourages In-Depth Thinking

Even the process of designing a simulation and determining the different parameters can offer solutions. By thinking in-depth about a process or procedure it is possible to come up with solutions or innovations without even using the final simulation.

8. Improve Stakeholder Buy-In

A visual simulation can also help improve buy-in from partners, associates and stakeholders. You can visually demonstrate the results of any process changes and how they were achieved, improving engagement with interested parties or even enabling a simulation based sales pitch.

Limitations

While there are a great many advantages to using simulation, there are still some limitations when compared to other similar techniques and technologies, such as digital twin.

A digital twin expands on simulation to incorporate real time feedback and a flow of information between the virtual simulation and a real life asset or assets. The difference being that while a simulation is theoretical, a digital twin is actual.

Due to this, simulations have limitations when it comes to assessing actual real-world situations as they occur.

Why is Simulation Used?

Simulation is used to evaluate the effect of process changes, new procedures and capital investment in equipment. Engineers can use simulation to assess the performance of an existing system or predict the performance of a planned system, comparing alternative solutions and designs.

Simulation is used as an alternative to testing theories and changes in the real world, which can be costly. Simulation can measure factors including system cycle times, throughput under different loads, resource utilisation, bottlenecks and choke points, storage needs, staffing requirements, effectiveness of scheduling and control systems.

What can be Simulated?

Any system or process that has a flow of events can be simulated. As a general rule, if you can draw a flowchart of the process, you can simulate it. However, simulation is most effective when applied to processes or equipment that change over time, have variable factors or random inputs. For example, our supermarket from earlier has variable and random factors due to customer use times, requirements and stocks.

Using simulation to model complex and changeable dynamic systems can offer insights that are difficult to gain using other methods.

While simulation can be used to manage processes, procedures and assets, Swedish philosopher Nick Bostrom took the notion of simulation further in his 2003 paper, ‘Are You Living in a Computer Simulation?’ He argues that by adding artificial consciousness to simulations, you can blur the lines between reality and simulation, making it difficult to tell if you are living in reality or if you are living in a simulation. This simulation hypothesis argues that, should you become aware that your ‘reality’ was not actually ‘real,’ your memories could be edited by the simulation to once again make you blissfully unaware that you are not actually a real person in the real world!

Moving away from the realms of post-human simulation, let’s return to some ‘real world’ types of simulation…

Types of Simulation

Simulation can be broken down into three overarching types, as follows:

1. Discrete Event Simulation

Modelling a system as it progresses through time, for example;

- factory operations (stamping, turning, milling)
- traffic analysis (roads, networks, queues)

2. Dynamic Simulation

Modelling a system as it progresses through space, for example;

- machine kinematics
- human ergonomics
- aerodynamic testing
- virtual prototyping

3. Process Simulation

Modelling physical interactions between two or more systems, for example;

- in-service product modelling
- in-manufacture product modelling
- weather forecasting

Examples

There are many examples of simulation across industry, entertainment, education, and more. Here are a few notable examples:

Automotive

Simulation allows the characteristics of a real vehicle to be replicated in a virtual environment, so that the driver feels as if they are sitting in a real car. Different scenarios can be mimicked so that the driver has a fully immersive experience. These type of simulators can help train both new and experienced drivers, offering a route to teach driving skills that can reduce maintenance and fuel costs and ensure the safety of the drivers themselves.

Biomechanics

Simulation can be applied to biomechanics to create models of human or animal anatomical structures in order to study their function and design medical treatments and devices. Biomechanics simulation can also be used to study sports performance, simulate surgical procedures, and assess joint loads. An additional example is neuromechanical simulation that unites neural network simulation with biomechanics to test hypotheses in a virtual environment.

City and Urban Planning

Simulation can be used to design new cities and urban environments as well as to test how existing urban areas can evolve as a result of policy decisions. This includes city infrastructure and traffic flow among other potential models.

Digital Lifecycle Design

Simulations can assist with product design, allowing digital prototyping and testing to create better performing products with a shorter time-to-market, while also assessing the lifecycle of the finished product.

Disaster Preparation

Simulations can replicate emergency situations, to help with disaster preparedness.This includes training and designing responses to events such as natural disasters, pandemics or terrorist attacks. Responses can be tracked and assessed through the simulation, highlighting potential problems and areas where more training may be required for responders, as well as ensuring any mistakes are made in a safe environment ahead of any real life event.

Economics and Finance

Economics, macroeconomics and finance also benefit from simulations. A mathematical model of the economy can, for example, be tested using historical data as a proxy for the actual economy. This can be used to assess inflation, unemployment, balance of trade and budgets.Elsewhere, simulations can replicate the stock exchange or be used to test financial models. Banks also use simulations to replicate payment and securities settlement systems.

Engineering Systems

Simulation is widely used for engineering systems to imitate operations and functions of equipment, processes and procedures. Engineering simulations can combine mathematical models and computer-assisted simulation for design or improvement of existing processes.

Ergonomics

Simulation can be used to analyse virtual products and working environments incorporating an anthropometric virtual representation of the human, also known as a mannequin or Digital Human Model (DHM). These DHMs can mimic the performance and capabilities of humans in simulated environments. This type of simulation has applications ranging from assembly lines to disaster management and video gaming to waste collection.

Flight Simulation

Flight simulators have been used for years to train new pilots in a safe environment. This not only allows pilots to be assessed safely, but can also test instrument failures and other problems without risking the pilot, the instructor or the aircraft. You can also easily repeat the exact same scenarios, such as approaching a runway to land, under different conditions, not to mention saving fuel and other costs compared to actual flying time.

Marine Craft Simulation

Much like flight simulation, it is also possible to simulate working in a ship or submarine. Simulators can include those that mimic the bridge, engine rooms, cargo handling bays, communications or remotely operated vehicles. These are used in training institutions, colleges and navies.

Military Applications

Sometimes referred to as ‘war games,’ military simulations can be used to test out military plans in a virtual environment using computer models. These can also incorporate social and political factors and are used by governments and military organisations around the world.

Network Systems

Simulations have been applied to network and distributed systems to test new algorithms and protocols before they are implemented in live systems. These can be applied to applications including content delivery networks, smart cities and the Internet of Things.

Project Management

Simulation can be used for project management analysis and training purposes. Whether training managers or analysing the outcomes of different decisions, simulation is frequently conducted with software tools.

Robotics

Robotics simulations are used to mimic situations that may not be possible to recreate and test in real life due to time, cost or other factors. The results of these tests can then be assessed and transferred to real life robots.

Production Systems

Production systems can be simulated using methods such as discrete event simulation to assess manufacturing processes, assembly times, machine set-up, and more.

Sales

Sales can be simulated to examine the flow of transactions and customer orders as well as costs, labour times and more.

Satellites and Space

The Kennedy Space Centre used simulation to train space shuttle engineers for launch operations. This would see people interact with a simulated shuttle and ground support equipment. Simulation is also used for satellite navigation tests.

Sport

Statistics are widely used as part of sport simulation to predict the outcome of events and the performance of individual sportspeople. Sports simulation can also be used to predict the outcome of games and events as well as for fantasy sports leagues. Biomechanics models can also be used to assist training, assess fatigue levels and their effect on performance and more.

Weather

Weather forecasting uses simulations based on past data to predict extreme weather conditions such as hurricanes or cyclones.

Conclusion

Simulations are used for a range of applications across industry, saving time and expense while being able to test theories and ideas before implementing them in the real world. Although related techniques such as digital twin may provide added benefits due to the two-way flow of information this allows, simulations still have a great many uses.

Whether testing theories, assessing procedural performance or determining the lifecycle of an asset simulation is a useful tool for many businesses and organisations.

What Is a Monte Carlo Simulation?

Monte Carlo simulations are used to model the probability of different outcomes in a process that cannot easily be predicted due to the intervention of random variables. It is a technique used to understand the impact of risk and uncertainty in prediction and forecasting models.

A Monte Carlo simulation can be used to tackle a range of problems in virtually every field such as finance, engineering, supply chain, and science. It is also referred to as a multiple probability simulation.

KEY TAKEAWAYS

- A Monte Carlo simulation is a model used to predict the probability of different outcomes when the intervention of random variables is present.
- Monte Carlo simulations help to explain the impact of risk and uncertainty in prediction and forecasting models.
- A variety of fields utilize Monte Carlo simulations, including finance, engineering, supply chain, and science.
- The basis of a Monte Carlo simulation involves assigning multiple values to an uncertain variable to achieve multiple results and then averaging the results to obtain an estimate.
- Monte Carlo simulations assume perfectly efficient markets.

Understanding Monte Carlo Simulations

When faced with significant uncertainty in the process of making a forecast or estimation, rather than just replacing the uncertain variable with a single average number, the Monte Carlo Simulation might prove to be a better solution by using multiple values.

Since business and finance are plagued by random variables, Monte Carlo simulations have a vast array of potential applications in these fields. They are used to estimate the probability of cost overruns in large projects and the likelihood that an asset price will move in a certain way.

Telecoms use them to assess network performance in different scenarios, helping them to optimize the network. Analysts use them to assess the risk that an entity will default, and to analyze derivatives such as options.

Insurers and oil well drillers also use them. Monte Carlo simulations have countless applications outside of business and finance, such as in meteorology, astronomy, and particle physics.

Monte Carlo Simulation History

Monte Carlo simulations are named after the popular gambling destination in Monaco, since chance and random outcomes are central to the modeling technique, much as they are to games like roulette, dice, and slot machines.

The technique was first developed by Stanislaw Ulam, a mathematician who worked on the Manhattan Project. After the war, while recovering from brain surgery, Ulam entertained himself by playing countless games of solitaire. He became interested in plotting the outcome of each of these games in order to observe their distribution and determine the probability of winning. After he shared his idea with John Von Neumann, the two collaborated to develop the Monte Carlo simulation.

Monte Carlo Simulation Method

The basis of a Monte Carlo simulation is that the probability of varying outcomes cannot be determined because of random variable interference. Therefore, a Monte Carlo simulation focuses on constantly repeating random samples to achieve certain results.

A Monte Carlo simulation takes the variable that has uncertainty and assigns it a random value. The model is then run and a result is provided. This process is repeated again and again while assigning the variable in question with many different values. Once the simulation is complete, the results are averaged together to provide an estimate.

https://corporatefinanceinstitute.com/resources/knowledge/modeling/monte-carlo-simulation/