66 Days of Data (22 February 2022 – 28 April 2022)

Day 1: What is ML (22/2/2022)

Machine Learning (ML) is a field of computer science that uses statistical techniques to give computer systems the ability to "learn" with Data, without being explicitly programmed

Definition: Machine learning is "a set of methods that can automatically **detect patterns** in data, and then use the uncovered patterns to **predict future data**"

Three types of learning

1. Supervised
2. Unsupervised
3. Reinforcement

Supervised Learning

- ➔ Used for **classification** (Fraud Detection, Image classification etc.) or **regression** (Forecasting, predictions etc.) problems
- ➔ Discern results and learn by trying to **find patterns** in a **labelled dataset**
- ➔ Predicts outcome/future
- ➔ Data with clearly defined output given
- ➔ Makes machine learn explicitly

Unsupervised Learning

- ➔ Best for dealing with **unstructured data**
- ➔ Does not predict or find anything specific
- ➔ **Clustering** / **Dimensional Reduction**
- ➔ Machine understand the data (identifies structures/patterns)

Reinforcement Learning

- ➔ Machine learn how to act in a certain environment
- ➔ Learning from + and – reinforcement
- ➔ Maximise rewards
- ➔ Skill acquisition, Game AI etc.

Day 1 Learning Summary

- ➔ Learned about what is machine learning
- ➔ Learned about the different types of machine learning

Day 2: The Data Science Project Cycle (23/2/2022)

Data Science Process

(Source: Introduction to Data Science Chapter 2, PG22)

1. **Setting the research goal**

➔ Define **research goal**

➔ The purpose here is to make sure that all stakeholders understand the **what**, **how** and **why** of the project.

2. **Retrieving data**

➔ We need to have **data available** for analysis

➔ Finding **suitable data** and getting access to the data from the data owner (Raw data)

3. **Data preparation**

➔ Transforming data from a raw form into data that's directly usable in the models.

➔ Data **cleansing** (Missing values, outliers etc.)

➔ Data **transformation** (Aggregating data, creating dummies, reduce number of variables)

➔ **Combining** data (Merging/joining datasets)

4. **Data exploration**

➔ Gain a **deeper understanding** of the data

➔ Looking for **patterns, correlations** based on visual and descriptive techniques

➔ Simple graphs, combined graphs

5. **Data modelling**

➔ Model and variable **selection** / Model **execution** / Model **comparison**

➔ Model building

➔ Make the predictions stated in our project charter / objectives

6. **Presentation and automation**

➔ Presenting the results and **automating** the analysis

In reality, we won't progress in a linear way from step 1 to step 6. Often, we will regress and iterate between the different phases.

Why follow the steps?

➔ Ensures that we have a well-defined research plan, good understanding of the business question and clear deliverables before we look into the data

➔ First steps of the process focus on getting high-quality data as input for our models. The "garbage in equals garbage out" concept.

Day 2 Learning Summary

➔ Learned about the data science project cycle

➔ 6 steps cycle of the data science process which is iterative

Day 3: Bias Variance Trade-off (24/2/2022)

Topic: What's the trade-off between bias and variance?

(Source: Machine Learning Interview Questions, LinkedIn)

Part 1: What is Bias?

- ➔ Bias is error due to erroneous or **overly simplistic** assumptions in the learning algorithm we are using.
- ➔ High error on **training** and **test** data.
- ➔ It can lead to **model underfitting** the data, making it **hard** for it to **have high predictive accuracy.**

Model Underfit

- Happens when our model is **unable to capture** the underlying pattern of the data (Usually high bias, low variance).
- It happens when we have very **less amount of data** to build an accurate model.

Part 2: What is Variance?

- ➔ Variance is error due to **too much complexity** in the learning algorithm we are using.
- ➔ This leads to the algorithm being **highly sensitive** to **high degrees of variation** in the **training data**, which will lead to the model to **overfit** the data.
- ➔ Model with **HIGH variance** pays a lot of attention to **training data** and does not generalise on the data which it hasn't seen before.
- ➔ High variance: **Perform well** on **training data** but has **high error rates** on **test data**.
- ➔ "Carrying" **too much noise** from the **training data** for the model to be very useful for the test data.

Model Overfit

- Happens when our model **captures the noise** along with the underlying pattern in data
- **Complex model** like Decision Trees is prone to overfitting (Low bias, High variance)

Part 3: Bias – Variance Trade-off

Essentially, if we make the model **more complex** and add more variables, we will **lose bias** but **gain some variance**.

Ultimately, we **do not want** either high bias or high variance in our model

Day 3 Learning Summary

- If our model is **too simple**, has very few parameters, it is likely to have **high bias and low variance**.
- If our model has **large number of parameters**, it is likely to have a **high variance and low bias**.
- Hence, it is crucial to find the **right balance** without overfitting and underfitting the data. **Ideal** to have **low bias**, **low variance**.

Additional sources:

https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229

Day 4: Type I & Type II Error (25/2/2022)

Topic: What is the difference between type I and type II error?

(Source: Machine Learning Interview Questions, LinkedIn)

Type I Error (False Positive)

- **Claiming** something **has happened** when **it hasn't**
- Eg: Telling a man he is pregnant
- In statistics, this occurs when a researcher **incorrectly rejects** a true **null hypothesis**
- In other words, the researcher report that his **findings** are **significant** when in fact it **occurred by chance**.
- Setting a lower significance level **decreases a Type I error** risk, but **increases a Type II error** risk.

Type II Error (False Negative)

- Claims that **nothing is happening** when in fact **something is**
- Eg: Telling a pregnant woman she isn't carrying a baby
- Occurs when a researcher **fails to reject** a **null hypothesis** which is really **false**
- Here, a researcher concludes there is **not a significant effect**, when actually **there really is**.
- Type II error can be reduced by **ensuring** that the **sample size is large** enough to detect a practical difference when one truly exists

Type I and Type II Error – The Covid-19 Pandemic Example

Null Hypothesis: A person is healthy and **not infected** with Covid-19 (**Negative**)

Alternative Hypothesis: A person is **infected** with Covid-19 (**Positive**)

Conclusion

- **Type I Error** (False Positive) may lead to patient having some anxiety and need to quarantine

False Positive: the test result says you have coronavirus, but you actually don't.

- **Type II Error** (False Negative) would give the patient an **incorrect assurance** that he or she was **not infected** when **in fact he or she does**.

False Negative: the test result says you don't have coronavirus, but you actually do.

- As a result of Type II Error, the person will not be treated and the virus would be **further spread** in the **community**.

Day 4 Learning Summary

- For **statisticians**, a **Type I Error** is usually **worse**. The rationale is that if we stick to the status quo or default assumption, at least we're **not making things worse**.
- A **Type I error** means mistakenly going against the main statistical assumption of a null hypothesis. This may lead to new policies, practices or treatments that are inadequate or a **waste of resources**. In contrast, a **Type II error** means failing to reject a null hypothesis. It may only result in missed opportunities to innovate.
- In practical terms, however, either type of error could be worse **depending** on your **research context**.

Day 5: Resampling (26/2/2022)

Topic: Why is resampling carried out?

(Source: Data Science Interview Questions and Answers, LinkedIn)

What is Sampling?

- Sampling is an active process of **gathering observations** with the intent of estimating a **population variable**.
- Sampling consists of selecting **some part of the population** to observe so that one may estimate something about the **whole population**.

What is Re-sampling?

- ➔ Resampling is a methodology of economically using a **data sample** to **improve the accuracy** and quantify the uncertainty of a population parameter.
- ➔ The key idea is to **resample** from the **original data** to create **replicate dataset**.
- ➔ A tool consisting in repeatedly **drawing samples** from a dataset and calculating statistics and metrics on each of those samples in order to **obtain further information** about something, in the machine learning setting, this something is the **performance of a model**.

Two Commonly used Re-sampling Methods

- Generally, resampling techniques for estimating model performance operate similarly.
- A subset of samples is used to **fit a model** and the remaining samples are used to **estimate the efficacy** of the model.

1. **Bootstrap**

Multiple samples are drawn from our original sample (resampling) with **replacement** (Allows the same sample to appear more than once in the sample)

We can explore the **different combinations** that could result in the first place, computing standard errors and confidence intervals, which give us a consistent range of values for estimate the true population parameter.

2. **K-Fold Cross Validation (CV)**

A dataset is partitioned into **k groups**, where each group is given the opportunity of being used as a **held-out test set** leaving the remaining groups as the **training set**.

Day 5 Learning Summary

- ➔ Both Bootstrap and Cross Validation are used in different scenarios.
- ➔ **Cross Validation** is widely used when we are assessing a **model performance** through test metrics.
- ➔ **Bootstrapping** is widely used when we want to **make estimates** about **some population value** with the standard errors and confidence intervals.

Additional sources:

https://medium.com/analytics-vidhya/resampling-methods-in-machine-learning-cross-validation-677485fa1b4d

https://machinelearningmastery.com/statistical-sampling-and-resampling/

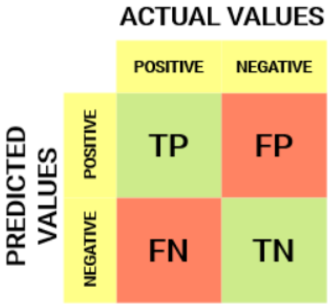
Day 6: Confusion Matrix (27/2/2022)

Topic: What is Confusion Matrix?

(Source: Data Science Interview Questions and Answers, LinkedIn)

Confusion Matrix is an **N x N matrix** used for **evaluating** the **performance** of a classification model, where N is the number of target classes.

For a binary classification problem, we would have a 2 x 2 matrix



True Positive (TP) • Predicted value matches actual value	**False Positive (FP) -> Type I Error** • Predicted value is false • Actual value was negative, • But model predicted a positive value
False Negative (FN) -> Type II Error • Predicted value is false • Actual value was positive, • But model predicted a negative value	**True Negative (TN)** • Predicted value matches the actual value

Accuracy = (TP + TN) / (TP + TN + FP + FN)

- Classification **accuracy** alone can be **misleading** if we have an unequal number of observations in each class (Think **Imbalance dataset**) or if you have more than two classes in your dataset.

Precision = TP / (TP + FP)

- Tells us how many of the currently **predicted cases** actually turned out to be positive
- What proportion of **positive identifications** was actually correct?
- Precision is a useful metric in cases where **False Positive is a higher concern** than False Negatives.
- Precision is important in music or video recommendation systems, e-commerce websites, etc. Wrong results could lead to customer churn and be harmful to the business.

Recall = TP / (TP + FN)

- Tells us how many of the **actual positive cases** we are able to predict correctly with our model (What proportion of actual positives was identified correctly?)
- Recall is a useful metric in cases where **False Negative trumps False Positive**.
- Recall is **important** in **medical cases** where it doesn't matter whether we raise a false alarm but the actual positive cases **should not go undetected**! For example, Recall would be a better metric because we don't want to accidentally discharge an infected person and let them mix with the healthy population thereby spreading the contagious virus.

Day 6 Learning Summary

- A confusion matrix is a summary of prediction results on a **classification problem**.
- Calculating a confusion matrix can give us a better idea of what our classification model is getting right and what types of errors it is making.

Additional sources:

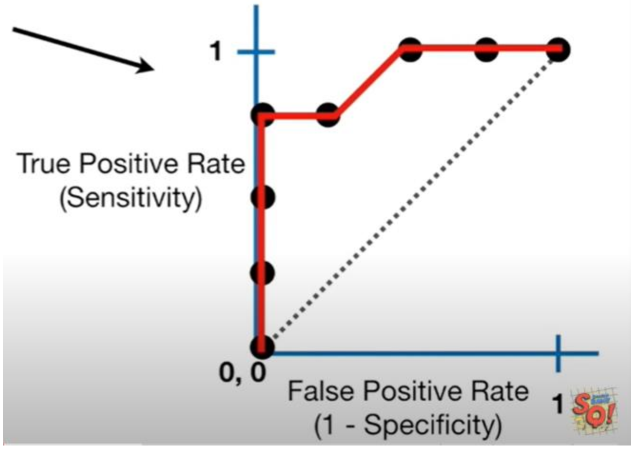
https://www.analyticsvidhya.com/blog/2020/04/confusion-matrix-machine-learning/

Day 7: ROC Curve and AUC (28/2/2022)

Topic: What is ROC curve and AUC?

(Source: Machine Learning Crash Course, Google Developers)

An **ROC curve** (Receiver Operating Characteristics curve) is a graph showing the performance of a classification model at all **classification thresholds**.



The curve plots two parameters:

True Positive Rate (TPR) -> Synonym for **Recall/Sensitivity**.

- ➔ TPR = TP / (TP + FN)

False Positive Rate (FPR)

- ➔ FPR = FP / (FP + TN)
- ➔ Often replaced with **Precision -> TP / (TP +FP)**, specifically useful when met with imbalance sample class

An ROC curve plots TPR vs. FPR at different classification thresholds.

Lowering the classification **threshold** classifies more items as positive, thus **increasing** both **False Positives** and **True Positives**.

To compute the points in an ROC curve, we could evaluate a logistic regression model many times with different classification thresholds, but this would be inefficient.

Area Under the ROC Curve (AUC)

- ➔ AUC measures the entire two-dimensional area **underneath** the entire **ROC curve**.
- ➔ The AUC makes it **easy to compare** one ROC curve to another.

Day 7 Learning Summary

- ➔ The ROC graph **summarizes** all of the **confusion matrices** that **each threshold produced** instead of being overwhelmed with confusion matrices.

Additional sources:

https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc

ROC and AUC, Clearly Explained! - YouTube

Day 8: Univariate, Bivariate & Multivariate Analysis (1/3/2022)

Topic: What are the differences between univariate, bivariate and multivariate analysis?

(Source: Data Science Interview Questions and Answers, LinkedIn)

Univariate Analysis

- The "uni" refers to one and "variate" refers to variable. As such, there are **only one dependent** variable in univariate analysis.
- The **main objective** of univariate analysis is to **find patterns** that exist within it.
- Patterns can be studied by drawing conclusions using **mean, median, mode,** etc.
- Univariate analysis can be described through **Frequency distribution**, **bar charts** (Discrete data), **histograms** (Continuous data), **pie charts**, **frequency polygons**.

Bivariate Analysis

The "Bi" refers to two and "variate" refers to variable. As such, there are **two variables** involved. Analysis is related to **cause and relationship** between the two variables.

1. **Bivariate analysis of two numerical variables (Numerical – Numerical)**

- **Scatter Plot** (Strength of relationship between two variables, linear correlation)

2. **Bivariate analysis of two categorical variables (Categorical – Categorical)**

- **Chi-Square Test**
- Used for determining the associate between **categorical variables**, probability of 0 indicates complete dependency between two variables while probability of 1 indicates that the two variables are completely independent)

3. **Bivariate analysis of one numerical and one categorical variable (Numerical – Categorical)**

- **z-test and t-test** (z and t-tests are important to calculate if the **difference** between a **sample** and **population is substantial**. If the sample size is large enough, then we use a Z-test, and for a small sample size, we use a T-test.)
- **Analysis of Variance / ANOVA** (Determine whether there is a significant difference among the averages of **more than two groups** that are statistically different)

Multivariate Analysis

- Involves **three or more variables**, it is categorised under **multivariate**. It is similar to a bivariate but contains more than one dependent variable.
- Types of multivariate analysis include **cluster analysis, factor analysis, multiple regression, principle component analysis** (PCA) etc.

Day 8 Learning Summary

➔ The analysis of data to discover relationships between measures in the data and to gain an insight on the relationships among various entities present in the data set with the help of statistics and visualization tools: **Exploratory Data Analysis (EDA)**.

Additional sources:

Statistics made easy ! ! ! Learn about the t-test, the chi square test, the p value and more - YouTube

https://www.analyticsvidhya.com/blog/2021/04/exploratory-analysis-using-univariate-bivariate-and-multivariate-analysis-techniques/

Day 9: Handling Missing Data (2/3/2022)

Topic: How to handle missing values in a dataset?

(Source: Wiem Gargouri LinkedIn post)

Missing data is a huge problem for data analysis because it **distorts findings**.

PART 1: Three types of Missing data

1. **Missing Completely at Random (MCAR)**

Data is completely missing at random across the dataset with no discernible pattern

2. **Missing at Random (MAR)**

Data is not missing randomly, but only within sub-samples of data

3. **Missing Not at Random (MNAR)**

There is a noticeable trend in the way data is missing

PART 2: Dealing with Missing Values

A. Deletion Method

- **Not a robust solution**. Recommended to **only use** this technique when the dataset contains **fewer** missing values.

Listwise Deletion: This method deletes the **entire row** when a column has an empty value.

Pairwise Deletion: This method calculates the correlation between two variables for **every pair of variables** to which data is considered. Less biased for MCAR or MAR data.

B. Data Imputation

- Imputation is that the method of **substituting** missing data with **substituted values**.
- General data is mainly imputed by mean, mode, median, multiple imputations and constants.
- General data is divided into **Continuous** and **Categorical**.
- **Numerical variables** can be **continuous** or **discrete** (variables are finite with specific range and are **countable**)

PART 2.1: Imputing Categorical & Continuous Variables

Categorical variables can always be imputed based on MODE. In other words, replace missing values with the **most frequent value**. Continuous variables can be handled using **mean, mode, median, multiple imputation**, and **linear regression**.

PART 2.2: Multiple Imputation

- Multiple Imputation by Chained Equation (**MICE**)
- A multiple imputation method used to replace missing data values in a data set under certain assumptions about the **data missingness mechanism** (e.g., the data are missing at random, the data are missing completely at random).

Additional sources:

https://www.analyticsvidhya.com/blog/2021/10/guide-to-deal-with-missing-values/#h2_6

https://www.analyticsvidhya.com/blog/2021/04/how-to-handle-missing-values-of-categorical-variables/#h2_2

Day 10: Cross Validation (3/3/2022)

Topic: What is cross validation?

Cross Validation is a technique which involves **reserving** a particular sample of a dataset on which you **do not train** the model. The data scientist will then test the model on this sample.

A resampling procedure used to evaluate machine learning models on a limited data sample.

1. Reserve a sample data set.
2. Train the model using the remaining part of the dataset.
3. Use the reserved sample of the test/validation set to help gauge model performance.

Cross Validation Methods

1. Validation set approach

- Reserving **50%** of the dataset for **validation** and the remaining **50%** for model **training**.
- Huge possibility that we might miss out on some interesting information about the data which will lead to high bias.

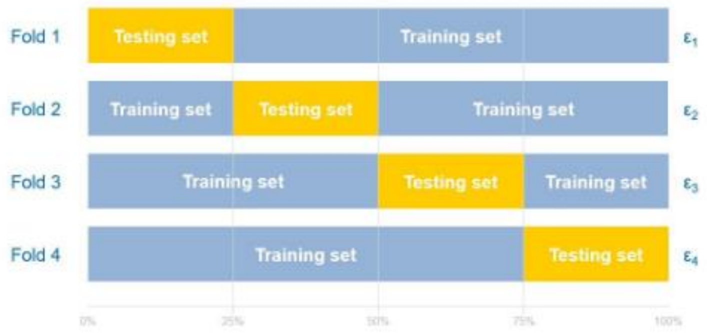
2. Leave one out cross validation (LOOCV)

- Reserve **only one data point** from the available dataset and train the model on the rest of the data. This process iterates for **each** data point.

3. K-fold cross validation

- Randomly split the entire dataset into **f folds**.
- For each k-fold in the dataset, build our model on **k – 1 folds** of the dataset. Then, test the model to check the effectiveness for k**th fold**.
- In practice, it is very common to divide the data into 10 blocks. (Ten-fold cross validation)

4-fold validation (k=4)



4. Stratified k-fold cross validation

- The process of rearranging the data so as to ensure that each fold is a **good representative** of the whole.
- In a binary classification problem where each class comprises of 50% of the data, it is best to arrange the data such that in every fold, **each class** comprises of about half the instances.

Additional sources:

https://www.analyticsvidhya.com/blog/2018/05/improve-model-performance-cross-validation-in-python-r/#h2_4

Day 11: Multicollinearity (4/3/2022)

Topic: What is multicollinearity?

Multicollinearity exists whenever an **independent variable** is **highly correlated** with one or more of the other **independent variables** in a **multiple regression** equation. It occurs when our model includes multiple factors that are correlated to **each other**.

Why Multicollinearity is a problem?

- It **increases** the **standard errors** of the coefficients. It means that coefficients for some independent variables may be found not to be significantly different from 0.
- Multicollinearity makes some variables **statistically insignificant** when they **should be** significant. (Coefficients might be significant if there are no multicollinearity – low standard errors)

How to measure Multicollinearity?

Variance Inflation Factor (**VIF**) is used to **measure multicollinearity**. It identifies the strength of correlation among the predictors.

- Alternatively, correlations can also be used to detect multicollinearity issues.
- If the VIF is **equal to 1**, there is **no multicollinearity** among factors.
- If the VIF is **greater than 1**, the predictors may be **moderately correlated**.
- A VIF between 5 and 10 indicates high correlation. If the VIF goes **above 10**, the regression coefficients are poorly estimated due to **multicollinearity** issues.

How to deal with Multicollinearity in our model?

Option 1 (Do nothing)

- If the model is used for **prediction only** and the variables are not of particular interest to study question & if the correlation is **not extreme**.

Option 2 (Remove highly correlated variables)

- **Remove highly correlated** predictors from the model. If there are more than 2 factors with a high VIF, remove one from the model.
- There is no point in keeping 2 very similar predictors in the model.

Option 3 (Combine correlated variables)

- Eg: Include a ‘seniority’ score **combining** both ‘experience’ and ‘age.

Option 4 (Use PCA or PLS)

- **Stepwise regression** can be used to remove variables. Select the model that has the highest R-Squared value.
- Use Principal Components Analysis (**PCA**) or Partial Least Squares Regression (**PLS**) – Regression methods that **cut** the number of **predictors** to a smaller set of uncorrelated components
- PCA is used when we want to reduce the number of variables in our data but we are not sure which variable to drop.

Additional sources:

https://www.analyticsvidhya.com/blog/2021/02/multicollinearity-problem-detection-and-solution/#h2_4

https://blog.minitab.com/en/understanding-statistics/handling-multicollinearity-in-regression-analysis

Day 12: Introduction to Logistic Regression (5/3/2022)

Topic: What is Logistic Regression?

Logistic Regression is a **Supervised** Machine Learning algorithm that can be used to model the probability of a **certain class** or event.

Logistic Regression is usually used for **binary classification** problems (Predicting output variable that is **discrete** in two classes)

Help data scientist to **predict** the **likelihood** of an event happening.

Types of Logistic Regression

1. **Binary Regression**

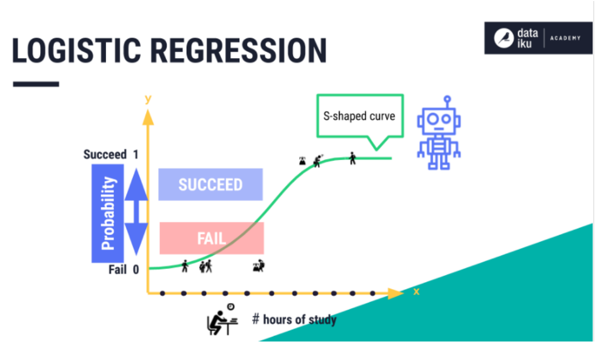
Either A or B

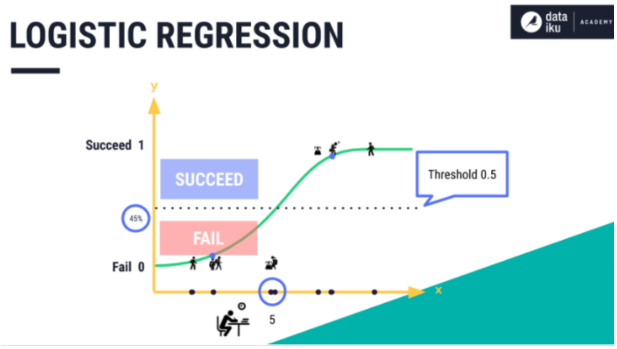
2. **Multinominal Logistic Regression**

The **output variable** is discrete in three or more classes with no natural ordering (A range of finite options A, B, C, D)

Logistic Regression with Example – Student Examinations

In Logistic Regression, data scientist attempts to **predict a class label** (Find out whether the student will succeed or fail on their exam).





The line of best fit is an **S-shaped curve** - **Sigmoid Curve**. We use the S-curve to predict the **likelihood or probability** of a data point belonging to "SUCCEED" category. All data points will ultimately be predicted as either "SUCCEED" or "FAIL".

A pre-defined **probability threshold** is needed for the class prediction.

All data points **above** the probability threshold will be predicted as "SUCCEED".

All data points **below** the probability threshold will be predicted as "FAIL".

Depending on where we set the threshold, the student's outcome could be classified as either "SUCCEED" or "FAIL".

Confusion matrix is used to help determine the optimal threshold.

Additional sources:

https://www.analyticsvidhya.com/blog/2021/07/an-introduction-to-logistic-regression/#h2_4

https://knowledge.dataiku.com/latest/courses/intro-to-ml/classification/classification-summary.html

Day 13: Introduction to Decision Tree (6/3/2022)

Topic: What is Decision Tree?

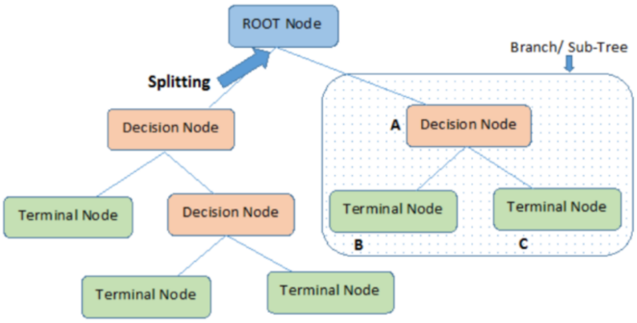
Decision Tree can be used for **classification** and **regression** problems. It uses a flowchart like a tree structure to show the predictions that result from a series of **feature-based splits** (Starts with **root node** and ends with decision made by leaves).

Terminology – Decision Tree (DT)

Root Nodes – It is the node present at the **beginning** of a **decision tree**.

Decision Nodes – The nodes we get after **splitting** the root nodes are called Decision Node.

Leaf Nodes – The nodes where further splitting is **not possible**.



Decision Trees are nothing but a bunch of **if-else statements** in layman terms. It checks if the condition is true and if it is then it goes to the next node attached to that decision.

Hyperparameter Tuning

Many of the real-world datasets have a large number of features, which will result in a large number of splits, it will gives a huge tree and such tree can lead to **overfitting**.

Hyperparameter Tuning can tackle the issue by using **max_depth** and **min_samples_split**.

A. Max_depth

- The more the value of max_depth, the **more complex** our tree will be, leading to **model overfit**. Hence, we will need a value that will not overfit as well as underfit our data (Eg: GridSearchCV).

B. Min_samples_split

- Setting the **minimum number** of samples for **each split**. In other words, we specific the minimum number of samples required to do a split.

Pruning -Improve model by **cutting down** some nodes/branches to **prevent overfitting**.

Pre-Pruning – Stop growing the tree earlier, prune/remove a node if it has **low importance** while growing the tree.

Post-Pruning – Once our 'tree' is **built to its depth**, we can stop pruning the nodes based on their significance.

Additional sources:

https://www.analyticsvidhya.com/blog/2021/08/decision-tree-algorithm/#h2_9

https://knowledge.dataiku.com/latest/courses/intro-to-ml/classification/classification-summary.html

Day 14: Ensemble Learning – Part 1 (7/3/2022)

Topic: What is Ensemble Learning?
Ensemble methods combine several trees base algorithms to **construct better predictive** performance than a single tree base algorithm.

The main principle behind the ensemble model is that a group of **weak learners** come together to form a **strong learner**, thereby increasing the **model accuracy**.

Ensemble helps in reducing the noise, variance and bias.

1. **Sequential Ensemble Learning (Boosting)**

➔ A machine learning ensemble **meta-algorithm** for principally **reducing bias**
➔ Base learners are generated **sequentially**
➔ Eg: Adaboost, Stochastic Gradient Boosting

2. **Parallel Ensemble Learning (Bagging)**

➔ A machine learning ensemble meta-algorithm intended to improve the **strength** and **accuracy** of machine learning algorithms used in classification and regression purpose.
➔ Base learners are generated in **parallel**.
➔ Eg: Random Forest, Bagged Decision Trees

3. **Stacking & Blending**

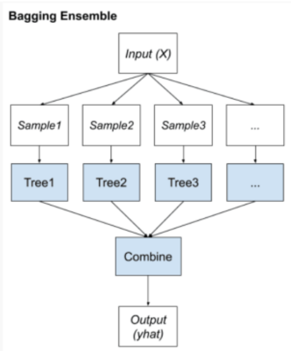
➔ A way of **combining multiple models**, introducing the concept of **meta learner**.
➔ Meta Learner is the science of systematically observing how different machine learning approaches perform on a wide range of learning tasks, and then **learning from this experience**, or meta-data, to learn new tasks much faster than otherwise possible.
➔ Blending is a technique where we can do weighted averaging of final result.
➔ Eg: Voting Classifier

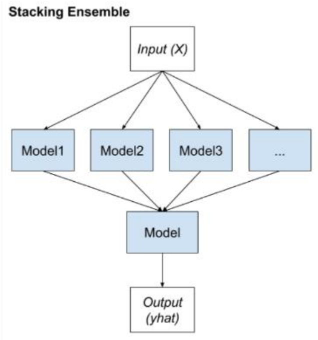
Step 1: Split training set into two disjoint sets

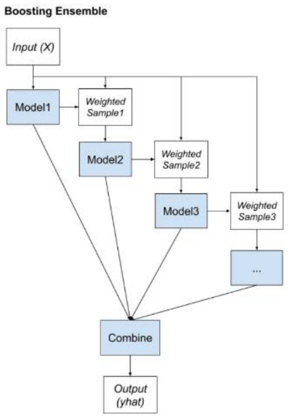
Step 2: Train several base learners on the first part

Step 3: Test the base learners on the second part

Step 4: Using the predictions from Step 3 as the inputs, and the correct responses as the outputs, train a higher-level learner







Additional sources:

https://machinelearningmastery.com/tour-of-ensemble-learning-algorithms/

https://medium.com/ml-research-lab/ensemble-learning-the-heart-of-machine-learning-b4f59a5f9777

Day 15: Outliers in Machine Learning (8/3/2022)

Topic: How to deal with outliers in machine learning?

Part A: What is an outlier?

- Part of the steps in **data pre-processing.**
- An observation in a given dataset that lies **far from the rest** of the observations.
- Outlier is vastly larger or smaller than the remaining values in the set.

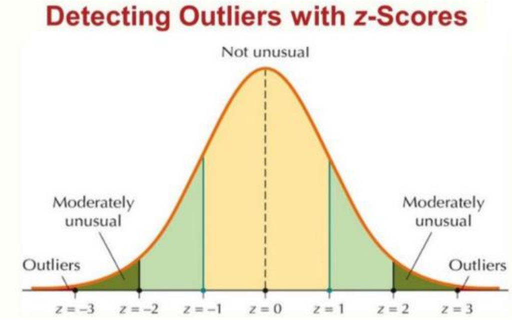
Part B: How to detect an outlier?

1. **Boxplot**

- A Method for graphically depicting groups of numerical data through their quartiles.
- May have lines extending vertically from the boxes (**whiskers**) indicating variability outside the upper and lower quartiles.

2. **Z-Score**

- A numerical measurement that describes a value's relationship to the **mean of a group of values**.
- Z-Score is measured in terms of **standard deviations** from a mean.
- A measure of a point's relationship to the average of all points in the dataset.



Part C: How to handle outliers?

1. **Remove the outliers**

- **Remove outlier values** from the dataset to stop them from skewing our analysis.
- Trim at both ends to remove outliers.
- May not be a good idea if we have a small dataset.

2. **Imputation**

- Alike with imputation of missing values, we can also **impute outliers**.
- Mean, median method can be used but **median** would be **more appropriate** as it is not affected by outliers.

3. **Transforming values**

- Transforming variables can also **eliminate outliers** and **reduces the variation** caused by extreme values.
- Scaling, Log Transformation, Cube Root Normalization, Box Transformation

Additional sources:

https://www.analyticsvidhya.com/blog/2021/05/detecting-and-treating-outliers-treating-the-odd-one-out/#h2_6
https://medium.com/analytics-vidhya/how-to-remove-outliers-for-machine-learning-24620c4657e8

Day 16: Different Types of Missing Data (9/3/2022)

Topic: How many types of Missing Data are there?

There are three distinct types of missing data

Missing Completely at Random (MCAR)

- There's no relationship between whether a data point is missing and any values in the data set.
- The missing data are just a random subset of the data.
- The missingness has **nothing to do** with any **other variable**.
- Assumption is made on whether or not the person (or a data point) has missing data is **completely unrelated** to the other information in the data.

Missing at Random (MAR)

- The missing data here is **affected only by the complete** (observed) **variables** and not by the characteristics of the missing data itself.
- Systematically **related to** the observed by **not the unobserved data**.
- **Missing data** in MAR can somehow **be predicted** from some of the other variables in the dataset
- If data is missing at random, it means that we need to either use an advanced **imputation method** (Multiple imputation) or an analysis method specifically designed for MAR data.

Missing Not at Random (MNAR)

- Data is missing is **related** to the **unobserved data** (The data that we don't have, the missingness is related to factors that we didn't account for).
- **Non-ignorable** missing data.
- **Could not** use any of the **standard methods** for dealing with missing data (Eg: Imputation) as any standard calculations will give the wrong answer.
- Eg: Could be the case where people with very low incomes and very high incomes tend to refuse to answer.

Additional sources:

https://medium.com/@raoufkeskes/missing-data-its-types-and-statistical-methods-to-deal-with-it-5cf8b71a443f
https://www.displayr.com/different-types-of-missing-data/
https://medium.com/analytics-vidhya/different-types-of-missing-data-59c87c046bf7

Day 17: Covariance (10/3/2022)

Topic: What is Covariance?

Covariance is a statistical tool that helps to quantify the total variance of random variables from their **expected value** (**Mean**).

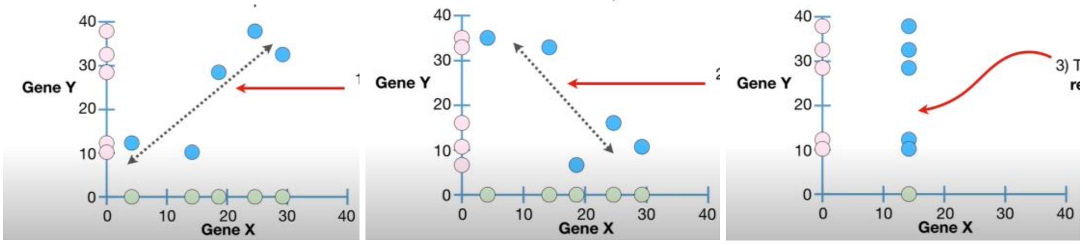
Covariance Formula:

Covariance Formula For Population

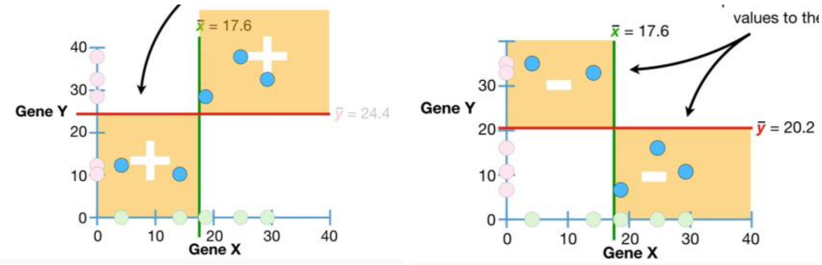
$$\mathrm{Cov}(X,Y) = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{n}$$

The main idea behind Covariance is that it can classify **three types of relationships**:

1) Relationships with positive trends
2) Relationships with negative trends
3) No relationship because there is no trend



Data in the 'NEGATIVE' quandrants contribute negative values to the covariance while data in the 'POSITIVE' quandrants contribute positive values to the total covariance.



- Covariance is a computational stepping-stone to something that is interesting, like correlation.
- When the covariance value is positive, we classify the trend as positive, vice versa.
- The covariance value **does not tell us** if the **slope of the line** representing the relationship is **steep or not**, it just tells us that the slope is positive.
- **Doesn't tell us** if the points are relatively **close** to or relatively **far** from the **dotted line**.
- **Covariance values** are **sensitive** to the **scale of the data**, and this makes them difficult to interpret.
- **Correlation** describes **relationships** and is **not sensitive** to the scale of the data.

Additional sources:

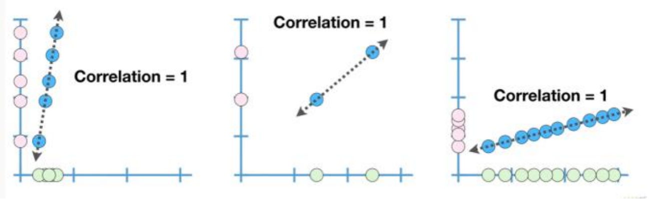
https://www.youtube.com/watch?v=qtaqvPAeEJY

Day 18: Correlation (11/3/2022)

Topic: What is Correlation?

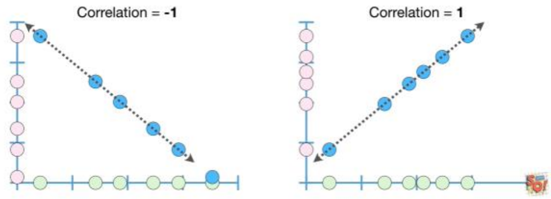
- We can quantify the strength of relationship with correlation.
- Data with a **relatively weak** relationship have a **small** correlation value.
- Data with a **moderate relationship** have a **moderate** correlation value.
- Data with a **strong relationship** have a **strong** correlation value.

The **maximum value** for correlation is 1



Correlation & P-value

- For correlation, a p-value tells us **the probability** that randomly drawn dots will result in a **similar** strong relationship, or stronger.
- The **smaller the p-value**, the **more confidence** we have in the **predictions** we make with the line.
- Correlation = 1 when a straight line with a positive slope can go through the center of every data point.
- Correlation does not depend on the **scale** of the data.



- As long as a straight line goes through all of the data and the slope of the line is negative, correlation = -1 when the slop is large and when the slope is small.
- If a straight line cannot go through all of the data, then we will get correlation values closer to 0, the worse the fit, the closer the correlation is to 0.

$$\text{Correlation} = \frac{\text{Covariance(Gene X, Gene Y)}}{\sqrt{\text{Variance(Gene X)}}\,\sqrt{\text{Variance(Gene Y)}}}$$

Day 18 Learning Summary

- Correlation quantifies the **strength** of **relationships**.
- Correlation values ranges from -1, which is the strongest linear relationship with a negative slope, to 1, which is the strongest linear relationship with a positive slope.
- Our confidence in our inferences depends on the amount of data we have collected the p-value. The **more data** we have, the **smaller the p-value** and vice versa.

Additional sources:

https://www.youtube.com/watch?v=xZ_z8KWkhXE

Day 19: R-Squared (12/3/2022)

Topic: What is R-Squared?

R-Squared (R^2) is a statistical measure that represents the proportion of the **variance** for a **dependent variable** that's **explained** by an **independent variable** or variables in a regression model.

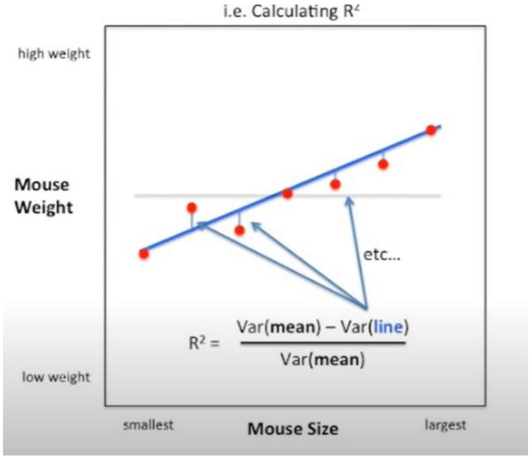
Correlation (The regular "R")

- ✓ Explains the strength of the relationship between independent and dependent variable.
- ✓ Correlation values close to 1 or -1 are good and tell us the two quantitative variables are strongly related.

R-Squared

- ❖ Explains to what extent the **variance** of one variable explains the variance of the second variable.
- ❖ The **percentage** of variation explained by the relationship between two variables.

Calculating R^2



- ❖ Quantifying the **difference** between the (**blue line**) and the **mean**.
- ❖ R^2 **does not** indicate the **direction** of the correlation because squared numbers are **never negative**.
- ❖ In Linear Regression (Or other linear models), R^2 and the related p-value are calculated using the **residuals**. We **square the residuals** and then **add them up**.
- ❖ Also called as **SS(fit)** or sum of squares for the residuals around the **best fitting line**.
- ❖ We then compare it to the sum of squared residuals around the **worst fitting line**, the **mean** of the y-axis value – the **SS(mean).**

Additional sources:

https://www.youtube.com/watch?v=2AQKmw14mHM

Day 20: Measure of Variability (13/3/2022)

Topic: What is Range, Variance and Standard Deviation?

Population: Entire group that we are taking for analysis.
Sample: Subset of the population. The size of the sample is always less than the total size of the population.

A population gives a true mean, and a sample statistic is an approximation population parameter which means a population mean is already known.

Measure of Variability/Dispersion

Range

- The difference between the largest and smallest values in a dataset.
- Range can be misleading when there are extremely high or low values.

Variance

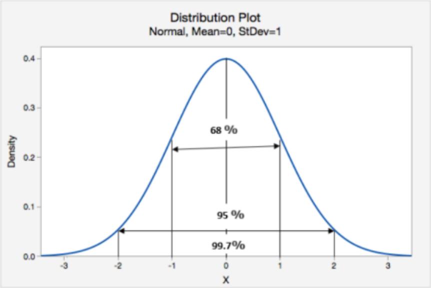
- Variance is a simple measure of dispersion.
- Variance measures how far each number in the dataset from the mean.
- It shows how the data points varied from the measure of central tendency.
- Population variance & sample variance.

Standard Deviation

- The standard deviation quantifies the variation within a set of measurements.
- The Standard Error quantifies the variation in the means from multiple sets of measurements. In short, Standard Error is the standard deviation of the means.
- Numerator is squared to avoid the opposite signs (+ve and -ve) values cancel each other.

Standard deviation is a squared root of the variance to get original values.

- ➔ Low standard deviation indicates data points close to mean



$$\text{Estimated Population Variance} = \frac{\sum(x - \bar{x})^2}{n - 1}$$

$$\text{Estimated Population Standard Deviation} = \sqrt{\text{Estimated Population Variance}} = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

68% of values lie within 1 standard deviation.
95% of values lie within 2 standard deviation.
99.7% of values lie within 3 standard deviation.

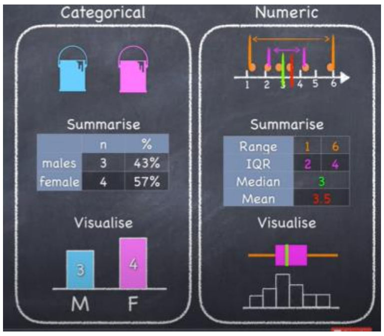
Additional sources:

https://medium.com/analytics-vidhya/statistics-range-variance-and-standard-deviation-f26bbfa0bbaa
https://www.analyticsvidhya.com/blog/2021/04/dispersion-of-data-range-iqr-variance-standard-deviation/#h2_2

Day 21: Statistical test (14/3/2022)

Topic: What is t-test, ANOVA, Chi-squared, Correlation?

Categorical Data: Refers to a data type that can be stored and identified based on the names or labels given to them. Also known as **qualitative data**.



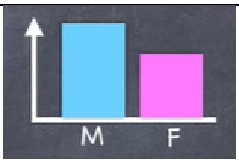
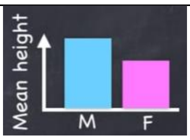
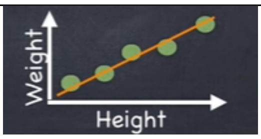
	n	%
males	3	43%
female	4	57%

Range	1	6
IQR	2	4
Median	3	
Mean	3.5	

Numerical Data: Refers to the data that is in the form of numbers, and not in any language or descriptive form. Also known as **quantitative data**.

Before getting into data analysis:

1. Defining Question and **Hypothesis**
2. **Null Hypothesis** and Identify Alpha Value (p-value < 0.05)
3. Analyse the Data

	What we observed in our sample data	**Is it real/Significantly difference?**
One categorical		1 sample proportion test
Two Categorical	Child Adult Elderly	Chi Squared
One Numeric		t-test
One numeric and one categorical		t-test or ANOVA (If categorical variable has more than 2 levels)
Two Numeric		Correlation test

Additional sources:

https://www.youtube.com/watch?v=I10q6fjPxJ0&t=85s

shorturl.at/dxHKQ

Day 22: One Hot Encoding (15/3/2022)

Topic: What is one hot encoding?

One Hot Encoding is the process of converting **categorical data** variables so they can be provided to machine learning algorithms to improve predictions.

Plays a crucial part of **feature engineering** for machine learning.

Categorical data: Variables that are made up of **label values**.

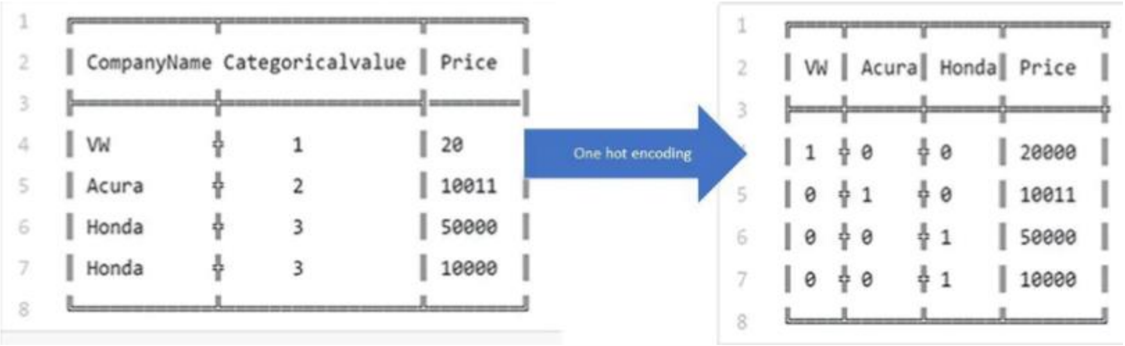
Some machine learning algorithms such as **Decision Tree** are able to work directly with **categorical data**.

However, most machine learning algorithms still require any inputs or outputs to be a number (numeric) in value.

As such, any categorical data must be mapped to integers.

How is One-Hot Encoding carried out?

- Each categorical value is converted into a new categorical column and assigned a **binary value** (Dummy variables) of 1 or 0 to those columns.
- Each integer value is represented as a binary vector. All the values are zero, and the index is marked with a 1.



CompanyName	Categoricalvalue	Price
VW	1	20
Acura	2	10011
Honda	3	50000
Honda	3	10000

VW	Acura	Honda	Price
1	0	0	20000
0	1	0	10011
0	0	1	50000
0	0	1	10000

Why is On-Hot Encoding important?

Useful for data that has **no relationship** to each other.

Machine Learning algorithms treat the **order of numbers** as an **attribute of significance**.

It means they will read a **higher number** as better or **more important** than a lower number. While this may be helpful for some **ordinal situations**, some input data **does not** have any **ranking** for categorical values (Could lead to **poor performance**).

Additional sources:

https://hackernoon.com/what-is-one-hot-encoding-why-and-when-do-you-have-to-use-it-e3c6186d008f

https://www.educative.io/blog/one-hot-encoding

https://www.kaggle.com/dansbecker/using-categorical-data-with-one-hot-encoding#Introduction

Day 23: Label Encoding (16/3/2022)

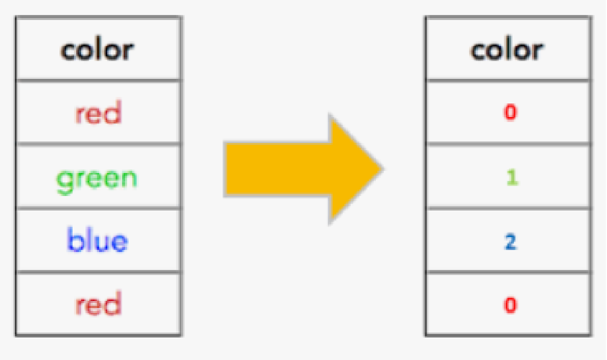
Topic: What is label encoding?

Label Encoding refers to converting the labels into a **numeric form** so as to convert them into the machine-readable form.

In short, it is the process of **converting** categorical values into **numerical values**.

It is an important **pre-processing** step for the structured dataset in supervised learning.

Example of label encoding:



color
red
green
blue
red

color
0
1
2
0

Things to note when doing label encoding:

- Label Encoding induces a new problem since it uses number sequencing.
- They introduce **relation/comparison** between them.
- Apparently, there is **no relation** between the color type, but when looking at the number, one might think that "blue" has **higher precedence** over "red".
- The **algorithm** may **misunderstand** that data has some kind of order 0 < 1 < 2 … and might give 2X **more weight** to "blue" in calculation then than "red".

In short:

Label Encoding is **ordinal** while One-Hot Encoding is not ordinal (No order preference).

Label Encoding has **lesser risk of collinearity** as it creates label of values based on alphabetical order.

Additional sources:

https://medium.com/wicds/label-and-one-hot-encoding-61525a32b99c

https://www.geeksforgeeks.org/ml-label-encoding-of-datasets-in-python/

https://towardsdatascience.com/categorical-encoding-using-label-encoding-and-one-hot-encoder-911ef77fb5bd

Day 24: Introduction to Clustering: K-Means (17/3/2022)

Topic: What is clustering: K-Means Clustering?
Clustering (Unsupervised Learning) is the task of dividing the population or data points into a number of groups.

- The aim is to segregate groups with similar traits and assign them into clusters.
- Eg: Grouping customers into clusters based on their payment history, useful for sales strategies etc.

Types of Clustering algorithms

Every method follows a different set of rules for defining the **"similarity"** among data points.

Connectivity models:

- ✓ Data points closer in data space exhibit more similarity to each other than the data points lying farther away.
- ✓ **First Approach**: Start with classifying all data points into separate clusters and aggregate them as the distance decreases. **Second Approach**: Start with all data points classified as a single cluster and then partitioned as the distance increases.

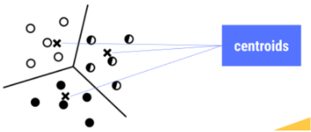
Centroid Models:

- ✓ **Iterative** clustering algorithms in which the notion of similarity is derived by the closeness of a data point to the centroid of the clusters (Eg: K-Means clustering).
- ✓ These models **run iteratively** to find the **local optima**.

K-Means Clustering

- ✓ An iterative clustering algorithm that aims to **find local maxima** in each iteration. In other words, K-Means separates data points into clusters, characterised by their midpoints (centroids).

Steps	Detail
1	Specify the desired **number of clusters K**: Let us choose k=2 for these 5 data points in 2-D space.
2	Randomly assign each data point to a **cluster**: Let's assign three points in cluster 1 shown using red color and two points in cluster 2 shown using grey color.
3	Compute **cluster centroids**: The centroid of data points in the red cluster is shown using red cross and those in grey cluster using grey cross.
4	Re-assign each point to the **closest** cluster centroid: Note that only the data point at the bottom is assigned to the red cluster even though its **closer** to the centroid **of grey cluster**. Thus, we assign that data point into grey cluster.
5	Re-compute cluster centroids: Now, re-computing the centroids for both the clusters.
6	**Repeat** steps 4 and 5 **until no improvements** are possible: Similarly, we'll repeat the 4th and 5th steps until we'll reach **global optima**.



Application of Clustering

- Recommendation Engines / Market Segmentation / Anomaly Detection / Pattern Recognition / Discover distinct groups in customer base

Additional sources:

https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/#h2_11

Day 25: Exploratory Data Analysis (18/3/2022)

Topic: What is Exploratory Data Analysis (EDA)?

EDA is used by Data Scientist to **analyse and investigate** data sets and summarise their main characteristics with the help of data visualisation methods.

Why is EDA important?

- ✓ Helps us to look at the data before making any **assumptions**.
- ✓ Help identify obvious **errors** and understand **patterns** within the data.
- ✓ Detect **outliers** or **anomalous** events.
- ✓ Find interesting **relations** among the variables.
- ✓ Identify **missing values** in the dataset.
- ✓ In short, EDA refers to the critical process of performing initial investigations on data so as to discover patterns, spot anomalies etc. with the help of **summary statistics** and **graphical representations**.

Summary statistics: Standard Deviation, Mean, Count, Quantiles of the data, Minimum and Maximum values

It is also often a good practice to **understand the data** first and gather as many insights as possible.

What are the EDA tools available?

Univariate visualisation of each field in the dataset with summary statistic (Univariate Analysis).

Bivariate visualisation and summary statistics that allow the data scientist to assess the relationship between each variable in the dataset and the target variable (Bivariate Analysis).

Multivariate visualisations, for mapping and understanding interactions between different fields in the data.

What are the example of the types of EDA tools available?

Histograms – A bar plot in which each bar represents the frequency (count) of cases for a range of values.

Box Plots – Graphically depict the five-number summary of minimum, first quartile, median, third quartile, and maximum value.

Scatter Plot – Plot data points on horizontal and vertical axis to show how much one variable is affected by another.

Additional sources:

https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15#:~:text=Exploratory%20Data%20Analysis%20refers%20to,summary%20statistics%20and%20graphical%20representations.

https://www.ibm.com/my-en/cloud/learn/exploratory-data-analysis#toc-why-is-exp-fLvQvfjK

https://www.analyticsvidhya.com/blog/2021/02/introduction-to-exploratory-data-analysis-eda/#h2_3

Day 26: Feature Engineering (19/3/2022)

Topic: What is Feature Engineering?

Part 1: What is Feature?

- All machine learning algorithms take **input data** to **generate** the **output**. The input data or **attributes** are often known as **features**.
- A feature is an attribute that **impacts** a problem.

Part 2: What is Feature Engineering?

- Feature Engineering is the **pre-processing** step of machine learning, it is used to transform raw data into features that can be used for creating a predictive model using machine learning.
- The main aim for Feature Engineering is to **improve performance** of models.

Part 3: What are the processes of Feature Engineering?

Feature Selection

- ✓ While developing the machine learning model, only a few **variables** in the dataset are **useful** for building the model, and the rest features are either **redundant or irrelevant**.
- ✓ If we input the dataset with all the redundant features, it may negatively impact and **reduce** the **overall performance** and **accuracy** of the model.
- ✓ In short, Feature Selection is about selecting the **most important independent** features which have more relation with the target variable. (EX. Correlation Matrix)

Handling Missing Values

- ✓ Missing values within the dataset **highly affect** the **performance** of the algorithm.
- ✓ **Imputation techniques** were used to handle missing values.

Handling outliers

- ✓ Outliers are the deviated values or data points that are **observed too away** from **other data points** in such a way that they **badly affect** the performance of the model.
- ✓ Replace outliers with **mean / quantile values** or **totally drop** the outliers.

Handling Imbalanced data

- ✓ Reduce **overfitting** and **underfitting** problem.
- ✓ **Under-sampling** majority class / **Over-sampling** minority class by duplication.

Label and One-Hot Encoding

- ✓ **Label** Encoder and **One-Hot** Encoder are used to convert object datatype to **integer datatype**.
- ✓ It is a technique that **converts** the categorical data in a form so that they can **be easily understood** by machine learning algorithms

Additional sources:

https://www.javatpoint.com/feature-engineering-for-machine-learning

https://www.analyticsvidhya.com/blog/2021/03/step-by-step-process-of-feature-engineering-for-machine-learning-algorithms-in-data-science/#h2_10

Day 27: Feature Engineering: Feature Scaling (20/3/2022)

Topic: What is Feature Scaling?

Feature Scaling is a method used to **normalise** the **range** of **independent** variables or features of data.

In data processing, it is also known as **data normalisation** and is generally performed during the data **pre-processing** step.

Applied to **reduce the variance** effect and to overcome the **fitting** problem.

Example:

Multiple variables with range as Age(18-100 Years), Salary(25,000-75,000 Euros), Height(1-2 Meters). Feature Scaling will help them all to be in the **same range** (in the range of 0 to 1.

Methods for Scaling

Normalization

Also known as **min-max normalization**. It is the simplest method and consists of rescaling the range of features to scale the range in [**0,1**].

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Makes the training process less sensitive by the scale of the features. This results in getting better coefficients after training.

Standardization

Makes the values of feature in the data have **zero mean** and **unit variance**.

The general method of calculation is to determine the **distribution mean** and **standard deviation** for each feature and calculate the new data point.

$$Z = \frac{X - \mu}{\sigma}$$

Additional sources:

https://www.analyticsvidhya.com/blog/2021/03/step-by-step-process-of-feature-engineering-for-machine-learning-algorithms-in-data-science/#h2_10

https://www.atoti.io/articles/when-to-perform-a-feature-scaling/#:~:text=Feature%20scaling%20is%20a%20method,during%20the%20data%20preprocessing%20step.

Day 28: Feature Selection: Boruta Algorithm (21/3/2022)

Topic: What is Feature Selection – Boruta Algorithm?

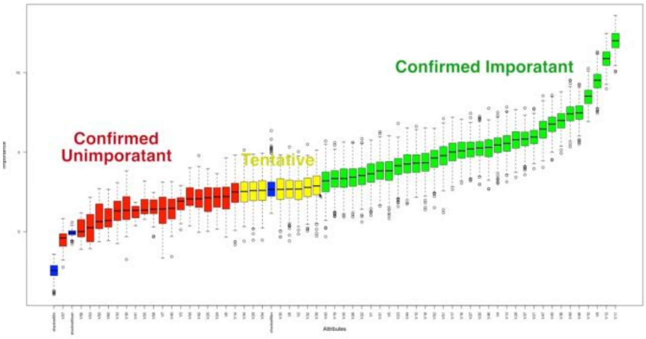
Feature Selection is an important aspect because it helps in building predictive models free from correlated variables, biases and unwanted noise.

Boruta Algorithm

- ✓ Feature Selection Algorithm.
- ✓ Works as a **wrapper algorithm** around **Random Forest**.
- ✓ Adds **randomness** to the given data set by **creating shuffled copies** of all features which are called **Shadow Features**.

Steps for Boruta Algorithm

1. Trains a **Random Forest classifier** on the extended data set and applies a feature importance measure (**Mean Decrease Accuracy**) to evaluate the importance of each feature where **higher** means more **important**.
   - Mean Decrease Accuracy plot **expresses how much accuracy the model losses by excluding each variable.**
2. At every iteration, it checks whether a **real feature** has a **higher importance** than the best of its **shadow features** (Higher Z score than the Z score of its shadow features)
3. Constantly **removes features** which are deemed **highly unimportant**.
4. Finally, the algorithm stops either when all features gets **confirmed or rejected** or it reaches a specific limit of random forest runs.



Boruta Algorithm in R – Practical (Parameters Involved)

maxRuns: Maximum number of **random forest** runs. Could increase this parameter if tentative attributes are left. **Default** is **100**.

doTrace: It refers to **verbosity level**. 0 means no tracing. 1 means reporting attribute decision as soon as it is cleared. 2 means all of 1 plus **additional reporting each iteration**. **Default** is **0**.

holdHistory: The full history of importance runs is stored if set to **TRUE (Default)**. Gives a plot of classifier run vs. Importance when plotImpHisoty function is called.

Additional sources:

https://www.analyticsvidhya.com/blog/2016/03/select-important-variables-boruta-package/#h2_4

https://www.andreaperlato.com/mlpost/feature-selection-using-boruta-algorithm/

Day 29: Model Explainability (22/3/2022)

Topic: What is Model Explainability?

Model Explainability refers to the concept of being able to understand the machine learning model.

Essential for model developers to explain the model once it is deployed in the real world.

Why is Model Explainability Required?

- The ability to **interpret a model** increases trust in a machine learning model.
- Once we understand a model, we can detect if there is any **bias** present in the model.
  - Ex: If a healthcare model has been trained on the American population, it might not be suitable for Asian people.
- Important while **debugging a model** during development phase.
- Critical for getting models to be vet by regulatory authorities like Food and Drug Administration (FDA) etc.
- Helps to determine if the models are **suitable to be deployed** in real life.

How do we develop Model Understanding?

First Method: Build models that are inherently **interpretable** – Glass Box Models

Ex: Linear Regression Model -> Y = b0 + b1*x (When x increases by 1%, Y will increase by b1% keeping other factors constant)

Second Method: Post-hoc explanation of **pre-built models** – Black Box Models

Ex: Deep learning model -> Model developers are not aware of how the input variables have combined to produce a particular output.

Ways to interpret a Model

Global Interpretation	**Local Interpretation**
Helps in understanding how a model makes decisions for the **overall structure**	This helps in understanding how the model makes decisions for a **single instance**
Can explain the **complete behaviour** of the model	Can explain the **individual predictions**
Helps in understanding the **suitability** of the model for **deployment**	Helps in understanding the **behaviour** of the model in the **local neighbourhood**
Ex: Predicting the **risk of disease** in patients	Ex: **Understanding why** a specific person has a high risk of a disease
Methods: PDP & ICE	Methods: LIME & SHAP

Additional sources:

https://www.analyticsvidhya.com/blog/2021/11/model-explainability/#h2_16

https://www.analyticsvidhya.com/blog/2021/06/beginners-guide-to-machine-learning-explainability/

Day 30: Stepwise Regression (23/3/2022)

Topic: What is Stepwise Regression?

Stepwise regression is the **step-by-step iterative** construction of a regression model that involves the selection of independent variables to be used in a final model.

It involves **adding or removing** potential explanatory variables in succession and testing for statistical significance after each iteration.

The idea is often to **find patterns** that existed in the past that might also recur in the future.

Goal of Stepwise Regression

Through a series of tests (Ex: t-test) to find a set of independent variables that **significantly** influence the dependent variable.

Types of Stepwise Regression

1. Forward Selection

Begins with **no variables** in the model, tests each variable as it is added to the model, then **keeps** those that are deemed **most statistically significant** – Repeating the process until the results are optimal.

2. Backward Selection

Starts with a **set of independent variables**, deleting one at a time, then testing to see if the **removed variable** is statistically significant.

3. Bidirectional elimination

A **combination** of the first two methods that test which variables should be **included or excluded**.

Limitations

Market conditions often change and relationships that have held in the past do not necessarily hold true in the present or future.

Additional sources:

https://www.investopedia.com/terms/s/stepwise-regression.asp

Day 31: Discriminant Analysis (24/3/2022)

Topic: What is Discriminant Analysis?

Discriminant Analysis is a statistical technique used to **classify observations** into non-overlapping groups – based on one or more quantitative predictor variables.

Ex: Doctor analysing (classify) a patient's risk of stroke (High / Low) using discriminant analysis based on personal attributes (Cholesterol level, body mass etc.).

Objective is to develop **discriminant functions** that are nothing but the **linear combination** of **independent variables** that will discriminate between the **categories** of the dependent variable.

Ways to Conduct Discriminant Analysis

a. Two-Group Discriminant Analysis

Used when the dependent variable has only **two categories** and is expressed as a dummy variable (Ex: 0 or 1).

Regression equation is called the "**discriminant function**".

Efficacy of the discriminant function is measured by the proportion of **correct assignments**.

Biggest **difference** between **discriminant analysis** and standard **regression analysis** is the use of a **categorical variable** as a **dependent variable**.

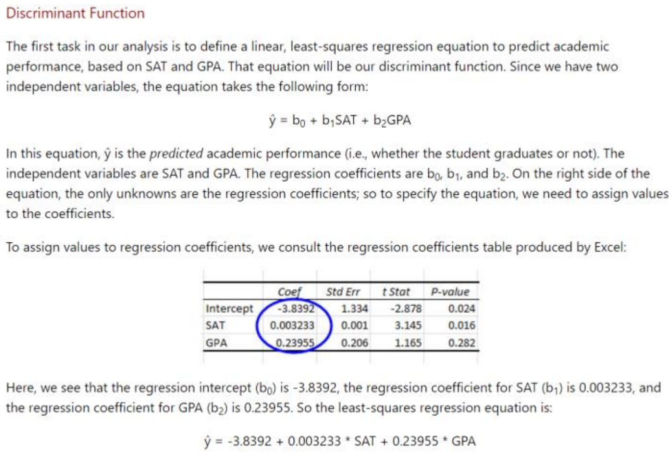
b. Multiple Discriminant Analysis

Used when the dependent variable has **more than two** classification groups.

Ex: Classifying voters into one of three political groups (BN, PH or Independent)

in the case of multiple discriminant analysis, **more than one** discriminant function can be computed.

Discriminant Function Formula



Discriminant Function

The first task in our analysis is to define a linear, least-squares regression equation to predict academic performance, based on SAT and GPA. That equation will be our discriminant function. Since we have two independent variables, the equation takes the following form:

$$\hat{y} = b_0 + b_1 SAT + b_2 GPA$$

In this equation, $\hat{y}$ is the *predicted* academic performance (i.e., whether the student graduates or not). The independent variables are SAT and GPA. The regression coefficients are b_0, b_1, and b_2. On the right side of the equation, the only unknowns are the regression coefficients; so to specify the equation, we need to assign values to the coefficients.

To assign values to regression coefficients, we consult the regression coefficients table produced by Excel:

	Coef	*Std Err*	*t Stat*	*P-value*
Intercept	-3.8392	1.334	-2.878	0.024
SAT	0.003233	0.001	3.145	0.016
GPA	0.23955	0.206	1.165	0.282

Here, we see that the regression intercept (b_0) is -3.8392, the regression coefficient for SAT (b_1) is 0.003233, and the regression coefficient for GPA (b_2) is 0.23955. So the least-squares regression equation is:

$$\hat{y} = -3.8392 + 0.003233 * SAT + 0.23955 * GPA$$

Additional sources:

https://www.statisticssolutions.com/discriminant-analysis/

https://stattrek.com/multiple-regression/discriminant-analysis.aspx

Day 32: Introduction: Recursive Feature Elimination (25/3/2022)

Topic: What is Recursive Feature Elimination (RFE)?

Recursive Feature Elimination (RFE) is a popular **feature selection** algorithm.

It fits a model and **removes the weakest features** until the specified number of features is reached.

It is a **greedy optimization** algorithm which aims to find the best performing feature subset.

Popular as it is **easy to configure** and effective at selecting features in a training dataset that are most relevant in predicting target variable.

Steps to carry out RFE

Performs greedy search to find the **best performing feature** subset.

It iteratively creates models and determines the best or the worst performing feature at **each iteration**.

Constructs the subsequent models with the feature left until all the features are **exhausted**.

It then **ranks the features** based on the order of their elimination.

In a worst-case scenario, if a dataset contains N number of features, RFE will do a greedy search for **2nd combinations** of features.

How RFE ranks feature importance?

Features are ranked by the model's `coef_` or `feature_importances_` attributes, and by recursively **eliminating a small number** of features per loop, RFE attempts to **eliminate dependencies** and **collinearity** that may exist in the model.

Additional sources:

https://www.scikit-yb.org/en/latest/api/model_selection/rfecv.html#:~:text=Recursive%20feature%20elimination%20(RFE)%20is,number%20of%20features%20is%20reached.

https://machinelearningmastery.com/rfe-feature-selection-in-python/

==========

https://www.tutorialspoint.com/data_mining/dm_cluster_analysis.htm#:~:text=What%20is%20Clustering%3F,into%20classes%20of%20similar%20objects.&text=A%20cluster%20of%20data%20objects,the%20labels%20to%20the%20groups.

https://knowledge.dataiku.com/latest/courses/intro-to-ml/clustering/clustering-summary.html

Day 33: Data Science Interview Question: ML (26/3/2022)

Topic: 101 Data Science Interview Questions for Machine Learning

Q1) **Explain Logistic Regression and its assumptions.**

There are two types of logistic regression: **Binary** and **Multinomial**.

Binary Logistic Regression deals with **two categories** whereas multinomial deals with **three or more** categories.

Q2) **Explain Linear Regression and its assumptions.**

Linear Regression is useful for finding the **relationship** between two **continuous variables**. One is the predictor or **independent variable** and the other is the response or **dependent variable**.

Q3) **How do you split your data between training and validation?**

First, ensure the validation set is large enough to yield **statistically meaningful** results.

Second, the validation set should be **representative** of the data set as a whole.

Q4) **Describe Binary Classification.**

Binary classification is the process of **predicting the class** of a given set of data points. These classes are also known as **targets/labels**.

Q5) **Explain the working of decision trees.**

A decision tree classification algorithm uses a training dataset to **stratify or segment** the predictor space into **multiple regions**. Each such region has only a subset of the training dataset.

Additional sources:

https://online.datasciencedojo.com/blogs/data-science-interview-questions?utm_content=202172219&utm_medium=social&utm_source=linkedin&hss_channel=lcp-3740012

Day 34: Data Science Interview Question: STATS (27/3/2022)

Topic: 101 Data Science Interview Questions for Statistics & Mathematics

Q1) **How would you select a representative sample of search queries from 5 million queries?**

Some key features need to be kept in mind while selecting a representative sample

1. Diverse – A sample must be as diverse as the 5 million search queries. It should be **sensitive** to all the **local differences** between the search queries.
2. Consistency – Important to ensure that any change we see in our sample data is also **reflected in the true population** which is the 5 million queries.
3. Transparency – It is extremely important to decide the **appropriate sample size** and structure so that it is a **true representative**.

Q2) **Discuss how to randomly select a sample from a product user population.**

The **sampling techniques** to select a sample from a product user population can be divided into two categories:

1. Probability Sampling Methods – Simple Random Sampling, Stratified Sampling, Clustered Sampling, Systematic Sampling
2. Non-Probability Sampling Methods – Convenience Sampling, Quota Sampling, Judgement Sampling, Snowball Sampling

Q3) **What does P-Value mean?**

P-Value is used to determine the **statistical significance** in the **Null Hypothesis**.

It stands for probability value and indicates how likely it is that a result **occurred by chance** alone. If P-Value is **small**, it indicates that the result was unlikely to have occurred by chance alone. Hence, it is known as being **statistically significant**.

A **large P-Value** indicates that the result is within change or normal sampling error which means that **nothing has happened** (not significant). It also indicates **weak evidence** against null hypothesis. Hence, we fail to reject the null hypothesis.

Additional sources:

https://online.datasciencedojo.com/blogs/data-science-interview-questions?utm_content=202172219&utm_medium=social&utm_source=linkedin&hss_channel=lcp-3740012

Day 35: Data Science Interview Question: DA (28/3/2022)

Topic: 101 Data Science Interview Questions for Data Analysis

Q1) **What are the core steps of the data analysis process?**

Data Analysis is a process in which we can change or analyse the data to **draw a conclusion** which will help to **achieve** a certain **goal**.

In short, it means the **process** involving inspecting, cleansing, transforming and modelling data to discover useful information.

A data analysis process consists of the following phases that are **iterative in nature**:

Setting Goals -> Data Gathering -> Data Processing -> Data Cleaning -> Data Analysis -> Result Interpretation -> Communication of Result

Q2) **How do you detect if a new observation is an outlier?**

A. Use **Boxplot / Whiskers plot** to visualise outlier:

Any value that will be more than the **upper limit** or lesser than the **lower limit** of the plot will be the outliers.

B. Standard Deviation

Find the points which lie more than **3 times** the standard deviation of the data.

Q3) **How do you solve for multicollinearity?**

Multicollinear occurs when **independent variables** in a regression model are **correlated**. This correlation is a problem because independent variables should be independent.

To solve this issue, we have to **remove highly correlated predictors** from the model. Remove one from the model if we have two or more correlated variables since they **supply redundant information**.

Regularization can be used to omit the problem of correlation because it stabilizes the regression coefficients. **Principle Component Analysis** can also be used to cut the number of correlated predictors.

Additional sources:

https://online.datasciencedojo.com/blogs/data-science-interview-questions?utm_content=202172219&utm_medium=social&utm_source=linkedin&hss_channel=lcp-3740012

Day 36: Data Science Interview Question: ML – P2 (29/3/2022)

Topic: 101 Data Science Interview Questions for Machine Learning – Part 2

Q1) **Why is regularization used in machine learning models? What are the differences between L1 and L2 regularization?**

Regularization is a technique used to **reduce the error** by fitting a function appropriately on the given training set thereby **avoiding overfitting**.

Key difference between L1 and L2 regularization is the **penalty term**.

A. Least Absolute Shrinkage and Selection Operator (LASSO Regression) – L1

Adds "**absolute value** of magnitude" **coefficient** as the penalty term to the loss function.

B. Ridge Regression – L2

Adds "**squared magnitude**" of **coefficient** as the penalty term to the loss function

Ridge sets weights of some features to **small values** whereas **Lasso shrinks** the less important features coefficient **to zero** thus, removing some features altogether. As such, it works well for **feature selection / dimensionality reduction** in case there is a huge number of features.

Q2) **Why is overfitting a problem in ML? What steps can we take to avoid it?**

- ✓ Overfitting is a phenomenon when a model **fits too closely** on the **training data**.
- ✓ It is said to have "**memorized**" the data and **performs very poorly** on **unseen data** because it is **not generalised**. The noise / outliers in the training data are picked up and learned as concepts by the model.
- ✓ As such, it **reduces** the predictive ability.

Overfitting can be reduced by applying:

- ❖ **Resampling techniques** such as k-fold cross validation that creates multiple train-test splits.
- ❖ Using **Ensemble techniques** that combines predictions from separate models and **reduce variance**.
- ❖ Increase **generalizability** using Regularization techniques that **add a penalty** to the cost function and **makes model** more **flexible**.

Additional sources:

https://online.datasciencedojo.com/blogs/data-science-interview-questions?utm_content=202172219&utm_medium=social&utm_source=linkedin&hss_channel=lcp-3740012

Day 37: Data Science Interview Question: STATS - P2 (30/3/2022)

Topic: 101 Data Science Interview Questions for Statistics & Mathematics - Part 2

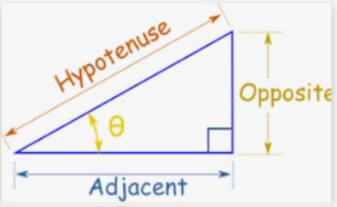
Q1) **Define Variance**

Variance of a distribution is a measure of the **variability of data**. It measures how far a set of (random) numbers are **spread out** from their **average value**.

It can be formulated as the average of the **squared differences** from the **mean**.

Q2) **Explain Euclidean Distance**

Euclidean Distance is used to **calculate distance** between **2 points** P and Q. It steams out from the **Pythagoras theorem** where the distance from point P to Q (2-dimensional space) is calculated by considering the line P to Q as **hypotenuse of a triangle**.



In terms of the Euclidean Distance in machine learning, it could be used to measure the "**similarity**" between two vectors. It is used by several classification and clustering algorithms.

Q3) **Define Central Limit Theorem and its Application**

Central Limit Theorem (CLT) states that the sampling distribution of a sample mean approached normal distribution as the **sample size gets larger**.

To make **statistical inferences** about the data.

CLT gives us the ability to quantify the probability that the random sample will **deviate from the population** without having to take any new sample to compare it with.

With CLT, we don't need the characteristics about the whole population to understand the likelihood of our **sample** being **representative** of it.

Confidence intervals, hypothesis testing and p-value analysis is based on the CLT. As such, CLT can **make inferences** from a **sample about a population**.

Additional sources:

https://online.datasciencedojo.com/blogs/data-science-interview-questions?utm_content=202172219&utm_medium=social&utm_source=linkedin&hss_channel=lcp-3740012

Day 38: Data Science Interview Question: DA – P2 (31/3/2022)

Topic: 101 Data Science Interview Questions for Data Analysis – Part 2

Q1) **If you are working at Facebook and you want to detect bogus/fake accounts. How will you go about that?**

The company can use the stored data to identify inauthentic profiles by **looking for patterns**.

Patterns such as **repeatedly posting** the same thing over and over OR a **sudden spike** in messaging activity.

Q2) **How do you inspect missing data?**

The following techniques can be used to handle missing data:

- ✓ **Imputation** of missing values depending on whether the data is numerical or categorical.
- ✓ Replacing values with **mean, median, mode**.
- ✓ Using the average value of **K nearest neighbours** as an **imputation estimate**.
- ✓ Using **linear regression** to predict values.

Q3) **What are the core steps for data pre-processing before applying machine learning algorithms?**

Data pre-processing is the process of giving structure to the data for **better understanding** and decision making related to the data.

- ○ Data Acquisition: Gather the data from all the sources and try to **understand the data**.
- ○ Cleaning: **Imputing null values** and treating **outliers/anomalies** in the data to make the data ready for further analysis.
- ○ Exploratory Data Analysis (EDA): **Find patterns** in the dataset and **extract new features** from given data.

Q3) **Facebook wants to analyse why the "likes per user and minutes spent on a platform are increasing, but total number of users are decreasing". How can they do that?**

There can be **multiple approaches** to answer this question. One way is to gather the **context information** for this problem. The following factors can be analysed from the data in order to reach a sound conclusion:

- ❖ **Timeline** – Is the drop in users a **one-time event** or has it happened **progressively**?
- ❖ **Region** – Is the decline in the number of users happening from a **specific region**? If it is, it may involve the country's regulations or a competitive product there.
- ❖ **Platforms** – Is the decline happening on **specific platforms**? (Ex: iOS, Android etc.) If it is, compare the number of users who are leaving on each platform.

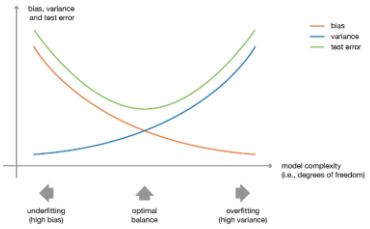
Additional sources:

https://online.datasciencedojo.com/blogs/data-science-interview-questions?utm_content=202172219&utm_medium=social&utm_source=linkedin&hss_channel=lcp-3740012

Day 39: Ensemble Learning – Part 2 (1/4/2022)

Topic: What is Ensemble Learning – Part 2?

- ➔ Ensemble learning is a machine learning paradigm where multiple models or "**weak learners**" are **trained** to solve the same problem and **combined** to get better results.
- ➔ The idea is that when **weak models** are correctly **combined**, we can obtain **more accurate** models.
- ➔ **Low bias** and **low variance** are the most fundamental feature expected for a model.



The idea of ensemble methods is also to **reduce bias** and **variance** of such weak learners by combining several of them together to create a strong learner with better performances.

Bagging: Often considers homogeneous weak learners, learns them **independently from each other** in **parallel** and combines them following some kind of deterministic averaging process.

Boosting: Often considers homogeneous weak learners, learns them **sequentially** in a very **adaptative way** (base model depends on the **previous ones**) and combines them following some kind of deterministic strategy.

- ✓ **Bagging** will mainly focus at getting an ensemble model with **less variance** than its components whereas **boosting** will mainly try to produce strong models **less biased** than their components (even if variance can also be reduced).

BAGGING

- Stands for "**Bootstrap Aggregating**" that aims at producing an ensemble model that is **more robust** than the individual models composing it.
- In parallel methods, we fit different considered learners **independently** from each other's and so, it is possible to **train them concurrently**.

BOOTSTRAPPING

- Consists of generating samples of size B (bootstrap samples) from an initial dataset of size N by randomly drawing **with replacement** B observations.

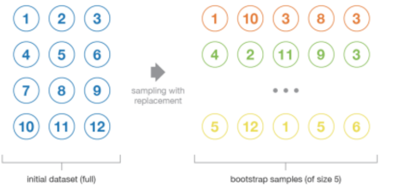


Illustration of the bootstrapping process.

Additional Source:

https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205

Day 40: Ensemble Learning – Part 3 (2/4/2022)

Topic: What is Ensemble Learning – Part 3?

BAGGING

When training a model, no matter if we are dealing with a classification or a regression problem, we obtain a function that **takes an input** and **returns an output**.

The idea of bagging is:

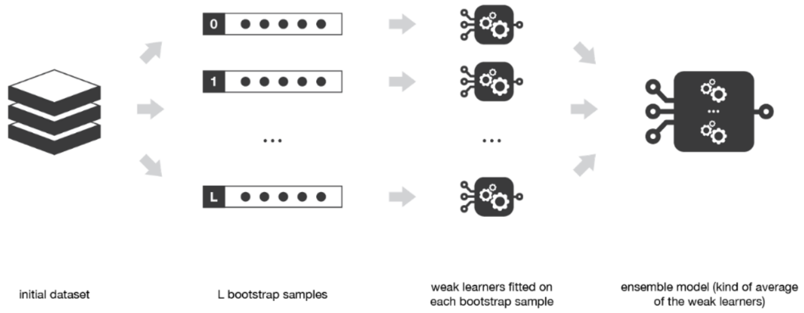
We wanted to fit several independent models and "**average**" their predictions in order to obtain a model with a **lower variance**.

However, we can't fit fully independent models because it would require **too much data** in practice.

Instead, we rely on the good "approximate properties" of **bootstrap samples** to fit models.

- ➔ First, we create **multiple bootstrap** samples so that each new bootstrap sample will **act** as another **independent dataset** drawn from true distribution.
- ➔ Then, we can **fit a weak learner** for each of these samples and finally **aggregate them** such that we kind of "**average**" their outputs. Thus, obtaining an ensemble model with less variance.

One of the advantages of Bagging is that it can be **parallelised**.


0
1
...
L
initial dataset
L bootstrap samples
weak learners fitted on each bootstrap sample
ensemble model (kind of average of the weak learners)

Random Forest

Learning tress are very popular **base models** for ensemble methods.

Strong learners composed of **multiple trees** can be called "forests". Tress that composes a forest can be chosen to be either **shallow** (Few depths) or **deep** (Lots of depths).

Shallow trees -> **Less variance** but higher bias (better off going for **sequential methods**)

Deep trees -> Low bias but **high variance** (relevant for **bagging method** that is focused at reducing variance)

Random Forest approach is a **bagging method** where **deep trees**, fitted on bootstrap samples are combined to product an output with **lower variance**.

Additional Source:

https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205

Day 41: Ensemble Learning – Part 4 (3/4/2022)

Topic: What is Ensemble Learning – Part 4?

BOOSTING

In **sequential methods**, the different combined weak models are no longer fitted independently from each other's.

The idea is to fit models **iteratively** such that the training of model at a given step depends on the **models fitted** at the **previous steps**.

"Boosting" produces an ensemble model that is in general **less biased** than the weak learners that compose it.

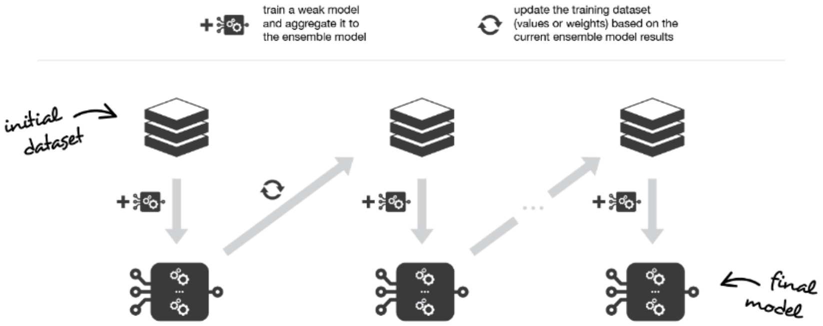
How does "Boosting" works?

In "Boosting", we build a family of models that are aggregated to obtain a strong learner that performs better. It **fits sequentially** multiple weak learners in a very **adaptive way**.

Each model in the sequence is fitted giving **more importance** to **observations** in the dataset that were **badly handled** by the previous models in the sequence.

Each new model focus its efforts on the **most difficult observations** to fit up to now to end up with a strong learner with lower bias.

The base models are often considered for boosting are models with **low variance but high bias**. In the case of using trees as base models, most of the time **shallow decision trees** with only a **few depths** will be chosen.


train a weak model and aggregate it to the ensemble model
update the training dataset (values or weights) based on the current ensemble model results
initial dataset
final model

Boosting consists in, iteratively, fitting a weak learner, aggregate it to the ensemble model and "update" the training dataset to better take into account the strengths and weakness of the current ensemble model when fitting the next base model.

Adaptive Boosting (Adaboost)

Updates the **weights attached** to each of the training dataset observations.

Gradient Boosting

Updates the **value** of these observations.

Additional Source:

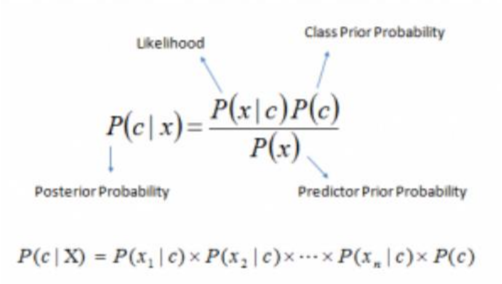
https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205

Day 42: Introduction to Naïve Bayes (4/4/2022)

Topic: What is Naïve Bayes classifier?

Naïve Bayes is a classification technique based on **Bayes Theorem** with an assumption of **independence** among predictors.

In other words, a Naïve Bayes classifier assumes that the presence of a particular feature in a class is **unrelated** to the presence of **any other feature**.



$$P(c \mid x) = \frac{P(x \mid c)P(c)}{P(x)}$$

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

P (C | X) -> The posterior probability of class (C, Target) given predictor (X, attributes).

P (C) -> The prior probability of class.

P (X | C) -> The likelihood which is the probability of predictor given class.

P (X) -> The prior probability of predictor.

What are the pros of Naïve Bayes?

- ➔ It is easy and fast to predict class of test data set. It also performs well in **multi class prediction**.
- ➔ **Lesser** training data is **needed**.
- ➔ **Perform well** in case of **categorical input variables** compared to numerical variables.

What are the cons of Naïve Bayes?

If categorical variable has a category, which was **not observed** in training data set, the model will assign a (**zero**) probability and will be unable to make a prediction. This is often known as "**Zero Frequency**".

Smoothing technique (**Laplace estimation**) is used to solve the issue.

4 Main Applications of Navie Bayes Algorithms

Real Time Prediction: Ability to make predictions in real time as it is fast.

Multi Class Prediction: Multiple classes of target variable.

Text Classification / Sentiment Analysis: Higher success rate as compared to the others.

Recommendation System: Data mining techniques applied together (Naïve Bayes & Collaborative Filtering) to filter unsees information.

Additional Source:

https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/#h2_6

Day 43: Linear Regression Analysis (5/4/2022)

Topic: What is Linear Regression Analysis?

Regression Analysis is a form of predictive modelling technique which investigates the **relationship** between a **dependent** (Target) and **independent variable (s)** (Predictor).

Generally used for finding the **causal effect relationship** between the variables.

Benefits of using Regression Analysis

Indicates the **significant relationships** between dependent variable and independent variable.

Indicates the **strength of impact** of multiple independent variables on a dependent variable.

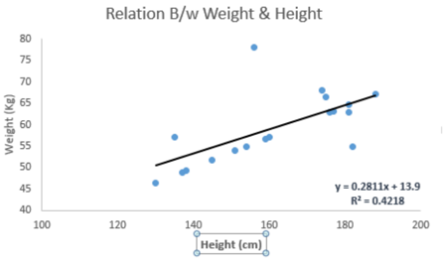
Linear Regression

Dependent variable is **continuous**, independent variable(s) can be **continuous or discrete**.

Linear Regression establishes a **relationship** between **dependent variable (Y)** and one or more **independent variables (X)** using a best fit straight line (Regression line)

Represented by an equation **Y = a + b*X + e** (a is intercept, b is the slope of the line, e is error term)

The equation can be used to predict the value of target variable based on given predictor variable(s).



Fitting the best fit line with Least Square Method

Most common method for fitting a regression line. It calculates the best-fit line for observed data by **minimizing the sum of squares** of the **vertical deviations** from each data point to the line.

As the deviations are first **squared**, there is **no cancelling out** between positive and negative values when added.

Model can be further evaluated with the **metric R-Squared**.

Additional Source:

https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/#h2_4

Day 44: Market Basket Analysis (6/4/2022)

Topic: What is Market Basket Analysis

Market Basket Analysis is also known as Association Rule Discovery or Affinity analysis.

In the simplest situation, the data consists of two variables: a transaction and an item.

To perform a Market Basket Analysis and identify potential rules, a data mining algorithm called the 'Apriori algorithm' is commonly used.

The purpose of carrying out market basket analysis is to determine what products customers purchase together. It takes its name from the idea of customers throwing all their purchases into a shopping cart (a "market basket") during shopping.

What could Market Basket Analysis be used for?

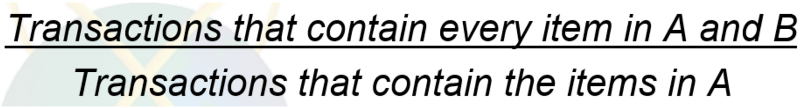
- Design store layout
- Marketing / Promotional Messages
- Maintain Inventory
- Recommendation Engines

Support: The percentage of transactions that contain all of the items in an item set (e.g., A & B). The higher the support the more frequently the item set occurs. Rules with a high support are preferred since they are likely to be applicable to a large number of future transactions.

Within a dataset, i.e. a list of transactions, how many transactions contain **item A**, so it is just the probability of **item A** occurring.

Confidence: The probability of a transaction that contains the items on the left-hand side of the rule (A) also contains the item on the right-hand side (B). The higher the confidence, the greater the likelihood that the item on the right-hand side will be purchased.

The confidence of an association rule A=>B is the conditional probability



$$\frac{\textit{Transactions that contain every item in A and B}}{\textit{Transactions that contain the items in A}}$$

Lift: the probability of all of the items in a rule occurring together.

For example, if pencil, paper and rubber occurred together in 2.5% of all transactions, pencil and paper in 10% of transactions and rubber in 8% of transactions, then the lift would be: 0.025/(0.1*0.08) = 3.125.

A lift of more than 1 suggests that the presence of pencil and paper increases the probability that a rubber will also occur in the transaction. Overall, lift summarizes the strength of association between the products on the left- and right-hand side of the rule; the larger the lift the greater the link between the two products.

Additional Source:

https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/#h2_4

https://www.thedataschool.co.uk/liu-zhang/understanding-lift-for-market-basket-analysis

Day 45: Dimensionality Reduction (7/4/2022)

Topic: What is Dimensionality Reduction?

Dimensionality reduction is the process of reducing the number of random variables under consideration, and can be divided into feature selection and feature extraction.

Feature Selection: **Selecting** a subset of original feature set.

Feature Extraction: Building a **new set of features** from the original feature set.

Why dimension reduction is crucial?

- It helps in data compressing and **reducing storage space** required.
- Less dimensions leads to **less computing**, also less dimensions can allow usage of algorithms unfit for a large number of dimensions.
- Removes **redundant features**.

What are the methods of dimension reduction?

Missing Values: First step of encountering missing values should be to **identify the reason** before carrying out imputation of missing values or drop the variables using appropriate methods.

Low Variance: In case of high number of dimensions, we should drop variables having low variance compared to others because these variables will **not explain the variation** in **target variables**.

Decision Trees: Can be used as an ultimate solution to tackle multiple challenges like **missing values, outliers** and identifying **significant variables**.

Random Forest: Similar with Decision Trees. However, have to be careful as RF have a **tendency to bias** towards variables that have **a greater number** of distinct values (favouring numeric variables over binary or categorical values).

High Correlation: Dimensions that exhibit **high correlation** can lower down the performance of model. Not good to have multiple variables of similar information known as "**multicollinearity**". Variables with Variance Inflation Factor (**VIF**) **more than 5** can be dropped.

Principle Component Analysis (PCA): Variables are transformed into a **new set of variables**, which are **linear combination** of original variables. These new set of variables are known as principal components.

Additional Source:

ABAV SASEM 6 PPT slides

Day 46: Artificial Neural Networks – Part 1(8/4/2022)

Topic: What is Artificial Neural Networks – Part 1?

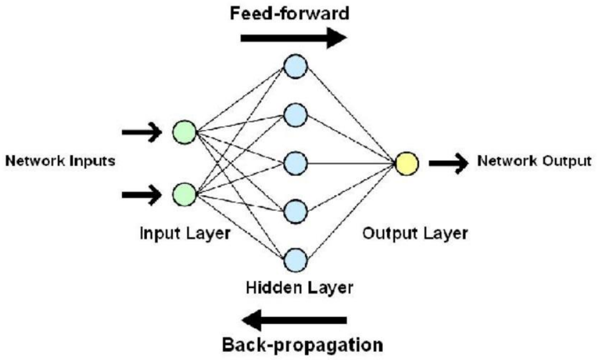
Artificial Neural Networks (ANN) are algorithms based on brain function and are used to model complicated patterns and forecast issues.

ANN is a **deep learning** method. (Concept of human brain biological neural networks)

Other deep learning algorithms include Convolutional Neural Networks (**CNN**) and Recursive Neural Networks (**RNN**): Used to accept unstructured and non-numeric data forms such as image, text and speech.

Artificial Neural Networks Architecture

There are three layers in the network architecture: **Input layer**, **Hidden layer** (More than 1) and the **Output layer**.



Hidden layer extracts some of the **most relevant patterns** from the **inputs** and sends them on to the **next layer** for further analysis.

Accelerates and improves the **efficiency** of the network by recognising **just** the **most important information** from the inputs and discarding the redundant information.

Activation function is crucial:

Captures the presence of **non-linear relationships** between the inputs.

Contributes to the **conversion of the input** into a more usable **output**.

Finding the optimal values of **W (Weights)** that **minimise prediction error**. Critical to building a successful model.

Backpropagation algorithm does this by converting ANN into a learning algorithm by **learning from mistakes**.

Gradient Descent technique is used to quantify **prediction errors**.

To find the **optimum value** for **W**, small adjustments in W are tried, and the impact on prediction errors is examined. Finally, those W values are chosen as ideal since further W changes do not reduce mistakes.

Additional Source:

https://www.analyticsvidhya.com/blog/2021/09/introduction-to-artificial-neural-networks/

Day 47: Recurrent Neural Networks – Part 2 (9/4/2022)

Topic: What is Recurrent Neural Networks?

Recurrent Neural Network (RNN)

Neural networks such as a **feed-forward networks** move data in one direction.

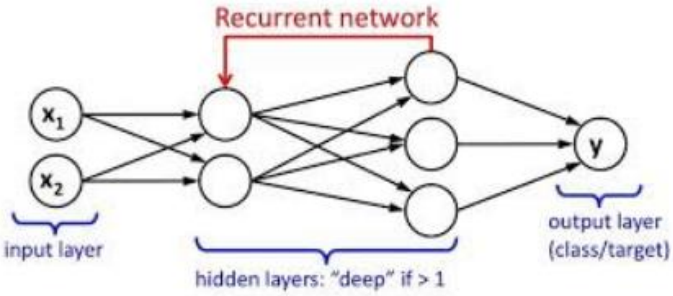
However, one disadvantage of this is that it does not remember the data in past inputs.

RNN is a network good at modelling **sequential data** (Data that follow a particular order in that a thing follows another).

How RNN works?

In RNN, the output of the previous stage goes back in as an input of the current step. Data runs through a loop such that each node remembers data in the previous step.

Essentially, RNN have a memory that helps the network recall what happened earlier in the sequence data. Neurons act as memory cells while carrying out operations.



Where is RNN commonly used in?

RNN are used to solve problems in stock predictions, text data and audio data.

Ex: text to speech conversion and language translation

Additional Source:

https://www.section.io/engineering-education/introduction-to-neural-networks/

Day 48: Convolution Neural Networks – Part 3 (10/4/2022)

Topic: What is Convolution Neural Networks?

Convolution Neural Network (CNN)

CNN are commonly used for **image recognition**. CNNs contain three-dimensional neuron arrangement.

This type of network understands the **image in parts**. It also computes the operations multiple times to complete the processing of the image.

CNN in Stages

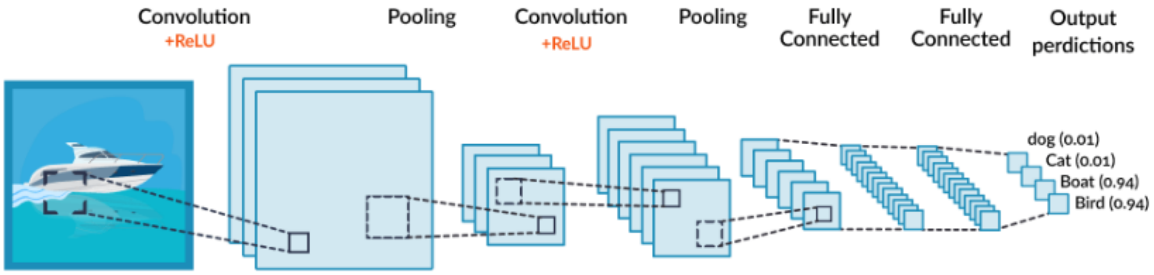
First stage is the **convolutional layer**. Neurons in a convolutional layer only process information from a **small part** of the visual field (image). Input features in convolution are **abstracted in batches**.

Second stage is **pooling**. It **reduces the dimensions** of the **features** and sustain valuable data at the same time.

Third phase starts when the **features** get to the **right granularity level**.

At the final stage, the final **probabilities** are **analysed** and decide which class the image belongs to.

CNN is mainly used in signal and image processing.



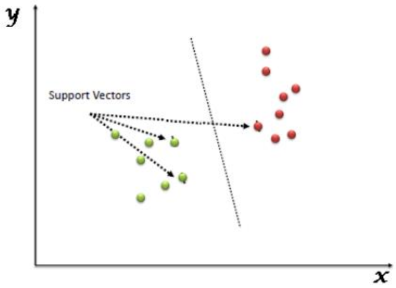
Additional Source:

https://www.section.io/engineering-education/introduction-to-neural-networks/

Day 49: Introduction to Support Vector Machine (11/4/2022)

Topic: What is Support Vector Machine (SVM)?

SVM is a supervised machine learning algorithm that can be used for both classification or regression tasks. (Mostly classification problems)



Each data item is plotted as a point in **n-dimensional space** (n is the number of features we have) with the value of each feature being the value of a particular coordinate.

We then **perform classification** by finding the **hyper-plane** that differentiates the two classes very well.

SVM is a frontier that best segregates the two classes (hyper-place / line).

How SVM works?

Example: Imagine we have 2 tags (**red and blue**), and our data has two features (**x and y**). The task is to have a classifier that, given a pair of **(x,y) coordinates**, **outputs** if its either red or blue.

A SVM takes these data points and outputs the hyperplane that best separates the tags. The line is called **decision boundary** (Anything that falls to one side of it will be classified as blue, anything that falls to the other as red).

For SVM, the best hyperplane is the one that **maximizes the margins** from both tags. (Whose distance to the nearest element of each tag is the largest).

How to tune parameters in SVM?

Tuning the **parameters' values** for machine learning algorithms effectively improves model performance.

Kernel: Various options available with kernel like “Linear”, “RBF”, “poly”

“RBF” and “POLY” are useful for non-linear hyper-plabe

Gamma: Kernel coefficient for “rbf”, “poly” and “sigmoid”. Higher value of gamma, will try to exact fit the training data set (overfit issue)

Additional Source:

https://monkeylearn.com/blog/introduction-to-support-vector-machines-svm/

https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/#h2_7

Day 50: Factor Analysis (12/4/2022)

Topic: What is Factor Analysis?

Factor Analysis is a technique that is used to **reduce** a large number of variables into **fewer** numbers of factors.

One of the unsupervised machine learning algorithms which is used for dimensionality reduction. This technique extracts **maximum common variance** from all variables and puts them into a common score.

Common variance: Variance due to **correlation** among these features (Ex: Grouping lazy and careless features to form a factor “unsuccessful behaviours” as they are correlated).

BARTLETT’S TEST OF SPHERICITY

- Checks whether the **correlation** is present in the given data. It **tests the null hypothesis** that the correlation matrix is an identical matrix. (Aim is to **reject null hypothesis** as factor analysis aims at explaining the common variance).

KAISER-MEYER-OLKIN (KMO) TEST

- Measures the **proportion of variance** that might be a **common variance** among the variables.
- Large proportions are expected as it represents **more correlation** is present **among the variables** thereby giving way for the application of dimensionality reduction techniques.

Determining Number of Factors

Number of factors can be decided on the basis of the amount of common variance the factors explain. In general, factors and their eigenvalues will be plotted. Eigenvalues is the amount of variance the factor explains. Select number of factors whose eigenvalues are greater than 1.

Common Factor Analysis -Extracts common variance and puts them into factors.

Factor Loading: The correlation coefficient for the variable and factor. Factor loading shows the variance explained by the variable on that factor.

Eigenvalues: Eigenvalues is also called characteristic roots. Eigenvalues shows variance explained by that particular factor out of the total variance.

From the commonality column, we can know how much variance is explained by the first factor out of the total variance. For example, if the first factor explains 68% variance out of the total, this means that 32% variance will be explained by the other factor.

Additional Source:

https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/factor-analysis/

https://www.analyticsvidhya.com/blog/2020/10/dimensionality-reduction-using-factor-analysis-in-python/

Day 51: Data Leakage (13/4/2022)

Topic: What is Data Leakage?

Data Leakage occurs when the data used in the training process contains information about what the model is trying to predict.

Also refers to a mistake that is made by the creator of a machine learning model in which they accidentally share the information between the test and training data sets.

How Data Leakage occurs?

When we split data into training and testing subsets, some of the data present in the test set is also copied in the train set, vice versa.

When the model is trained with this type of split, it will give really good results on the train and test set.

But when we deploy the model into production, it will not perform well, because when a new type of data comes in, it won't be able to handle it.

Leakage during Data Pre-Processing

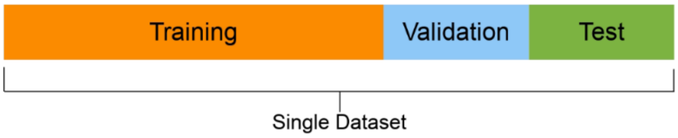
- Evaluating the parameters for normalizing or rescaling features
- Finding the minimum and maximum values of a particular feature
- Normalize the particular feature in our dataset
- Removing the outliers
- Fill or completely remove the missing data in our dataset

The above-described steps should only be done using training set. If entire dataset is used to perform these operations, data leakage may occur.

Ultimately, applying pre-processing techniques to the entire dataset will cause the model to learn not only the training set but also the test set.

How to fix Data Leakage issues?

- Extract appropriate set of features
- Create separate validation set



Purpose of validation set is to mimic the real-life scenario and can be used as a final step.

- Apply data pre-processing separately to both train and test subsets.
- Use cross validation to split data into k folds and iterates over the entire dataset in k number of times and each time we are using k-1 fold for training and 1-fold for testing our model.

Additional Source:

https://www.analyticsvidhya.com/blog/2021/07/data-leakage-and-its-effect-on-the-performance-of-an-ml-model/#h2_4

Day 52: Data Visualisation (14/4/2022)

Topic: What is Data Visualisation?

It is the practice of translating information into a visual context, such as a map or graph to make data easier for human to understand and pull insights from.

The term is often used interchangeably with others, including information graphics, information visualisation and statistical graphics.

The main goal of Data Visualisation is to:

Identify patterns, trends and outliers in large data sets.

When a data scientist is writing advanced predictive analytics or machine learning algorithms, it becomes important to visualise the outputs to monitor results and ensure that models are performing as intended.

Why is data visualisation important?

- Provides a quick and effective way to communicate information in a universal manner using visual information.
- Help businesses identify which factors affect customer behaviour and pinpoint areas that need to be improved.
- Make data more memorable for stakeholders.

Important to note:

The insights provided by big data visualisation will only be as accurate as the information being visualised.

As such, it is essential to have people and processes in place to govern and control the quality of corporate data, metadata and data sources.

Example of data visualisation

Infographics / Bubble clouds / Heat maps / Time series charts / Fever charts

Scatter plots / Treemaps / Area charts / Line charts

Additional Source:

https://www.techtarget.com/searchbusinessanalytics/definition/data-visualization

Day 53: Data Mining Methodology – Part 1 (15/4/2022)

Topic: What is Data Mining Methodology? – P1

Several data mining processes could be applied to modern data science projects.

CRISP-DM

The key difference between CRISP-DM and (SEMMA & KDD) in the structure is that the transitions between stages can be reversed.

Originally developed in IBM for data mining task.

Business Understanding

This stage is aimed toward getting a general understanding of the client’s business.

It is crucial in most cases to understand the application of the product to be developed.

Determine business objectives, assess the situation, determine goals, produce product plan.

Data Understanding

The second stage consists of collecting and exploring the input dataset. The set goal might be unsolvable using the input data.

Collect initial data, describe data, explore data, verify data quality etc.

Data Preparation

This phase, which is often referred to as “data munging”, prepares the final data set(s) for modelling. Bad input inevitably leads to bad output.

Select data, clean data, construct data, integrate data and format data.

Modelling

Build and assess various models based on several different modelling techniques.

Select modelling techniques, build model etc.

Evaluation

This stage is aimed at the evaluation of the obtained results. This stage is also crucial for checking if the business goal was fulfilled.

Evaluate results, review process, determine next steps

Deployment

A model is not particularly useful unless the customer can access its results.

Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise.

Additional Source:

https://www.datascience-pm.com/crisp-dm-2/

Day 54: Data Mining Methodology – Part 2 (16/4/2022)

Topic: What is Data Mining Methodology? – P2

Knowledge Discovery in Databases (KDD)

A method of how specialists can extract patterns and/or required information from data.

Consists of 5 stages

- **Selection**: Creating a target data set or focusing on a subset of variables or data samples that require further exploration.
- **Pre-processing**: Target data pre-processing to obtain consistent data.
- **Transformation:** Data transformation using dimensionality reduction or transformation methods.
- **Data Mining**: Searching for patterns of interest in a particular representational form that depends on the Data Mining goal (e.g. prediction).
- **Interpretation/Evaluation**: Interpretation and evaluation of the mined patterns.

SEMMA

Has similar structure to KDD but it is easier to apply to general Data Science tasks as it does not focus as heavily on data-specific stages.

- **Sample**: A portion of a large data set is taken that is big enough to extract significant information and small enough to manipulate quickly.
- **Explore**: Data exploration can help in gaining understanding and ideas as well as refining the discovery process by searching for trends and anomalies.
- **Modify**: Data modification stage focuses on creating, selecting and transformation of variables to focus model selection process. This stage may also look for outliers and reducing the number of variables.
- **Model**: There are different modelling techniques present and each type of model has its strengths and is appropriate for a specific goal for data mining.
- **Access**: This final stage focuses on the evaluation of the reliability and usefulness of findings and estimates the performance.

Additional Source:

https://www.techtarget.com/searchbusinessanalytics/definition/data-visualization

https://medium.datadriveninvestor.com/data-science-project-management-methodologies-f6913c6b29eb

Day 55: Recommender Systems – Collaborative Filter (17/4/2022)

Topic: What is Recommender System?

A recommender system predicts what a user would rate a specific product based on their preferences.

Recommender Systems (RS) are information filtering systems that help deal with the problem of information overload by filtering and creating fragments out of large amounts of dynamically generated information according to user's preferences, interests about a particular item.

RS has the ability to predict whether a particular user would prefer an item or not based on the user's profile and its historical information.

It can be mainly split into two different areas:

Collaborative Filtering

A method that is solely based on the past interactions that have been recorded between users and items.

Tends to find what similar users would like and the recommendations to be provided -> to classify the users into clusters of similar types and recommend each user according to the preference of its cluster.

Memory Based approach: Directly work with the values of recorded interactions and are essentially core based on nearest neighbours' search.

Example: Finding the closest users from a user of interest and suggest the most popular items among these neighbours.

Recommending an item to user A based on the interests of a similar user B.

Main idea:

Through past user item interactions when processed through the system, it becomes sufficient to detect similar users to make predictions based on these estimated insights.

Only need user's historical preference on a set of items to recommend from.

Suffer from cold-start problem (If an item is not seen during training, the system cannot generally create an embedding for it. Hence, it cannot query the model with this item.

Example:

Last.fm recommends tracks that other user with similar interests play often.

Additional Source:

https://analyticsindiamag.com/collaborative-filtering-vs-content-based-filtering-for-recommender-systems/

Day 56: Recommender Systems – Content Based (18/4/2022)

Topic: What is Recommender System? (Content based)

Content Based approach uses additional information about users / items.

Uses item features to recommend other items similar to what the user likes and also based on their previous actions or explicit feedback.

For example:

In movie recommendation system, age, gender, job, category could be all be the “items”

Main Idea of Content-Based Method

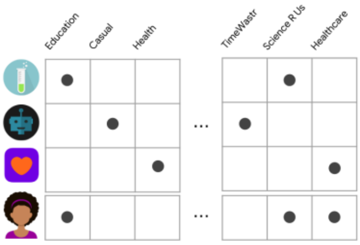
To try to build a model based on the available “features” that explain the observed user-item interactions.

For example, looking at the profile of this user and based on its information to determine relevant movies to suggest.

Content Based filtering model does not need any data about other users. This makes it easier to scale down the same to a large number of users.

Utility Matrix can help signify the user’s preferences for certain items. Data about the relationship between the items which are liked by the user as well as those which are disliked could be gathered with the help of Utility Matrix.

Using dot product as a similarity measure. A high dot product then indicates more common features, thus a higher similarity.



Challenges

Content Based approach requires a good amount of information about items’ features, rather than using the user’s interactions and feedback.

Content-based model can only make recommendations based on the existing interests of the user and the model hence only has limited ability to expand on the users’ existing interests.

Additional Source:

https://developers.google.com/machine-learning/recommendation/content-based/basics

https://analyticsindiamag.com/collaborative-filtering-vs-content-based-filtering-for-recommender-systems/

Day 57: Data Analyst vs. Data Scientist (19/4/2022)

Topic: Data Analyst vs. Data Scientist

Data Analyst

Data Analyst typically work with structured data to solve tangible business problems using tools like SQL, R, Python programming languages, data visualisation, statistical analysis.

Common tasks include:

- Identify information needs
- Acquiring data from primary and secondary sources
- Cleaning and reorganizing data for analysis
- Analysing data sets to spot trends and patterns that could be translated into actionable insights
- Present findings in an easy-to-understand way

Data Scientist

Data Scientist often deal with the unknown by using more advanced data techniques to make predictions about the future.

Might automate machine learning algorithms or design predictive modelling processes

Common tasks include:

- Gathering, cleaning, and processing raw data
- Designing predictive models and machine learning algorithms to mine big data
- Developing tools and processes to monitor and analyse data accuracy
- Building data visualisation tools / dashboard / reports
- Write programs to automate data collection and processing

Additional Source:

https://www.coursera.org/articles/data-analyst-vs-data-scientist-whats-the-difference

Day 58: SQL Commands - Data Definition Language (20/4/2022)

Topic: What is data definition language (DDL)

DDL

DDL consists of the SQL commands that can be used to define the database schema.

Used to create and modify the structure of database objects in the database.

DDL is a set of SQL commands used to create, modify and delete database structures but not data.

These commands are normally not used by a general user, who should be accessing the database via an application.

List of DDL commands:

Create: Used to create the database or its object (table, index, function, views, store procedures etc.)

Drop: Used to delete objects from the database

Alter: Used to alter the structure of the database

Truncate: Used to remove all records from a table, including all spaces allocated for the records are removed.

Rename: Used to rename an object existing in the database

Additional Source:

https://www.geeksforgeeks.org/sql-ddl-dql-dml-dcl-tcl-commands/

https://www.w3schools.in/mysql/ddl-dml-dcl

Day 59: SQL Commands - Data Manipulation Language (21/4/2022)

Topic: What is data manipulation language (DML)

DML

DML deals with data manipulation and includes most common SQL statements such as SELECT, INSERT, UPDATE, DELETE etc.

In other words, DML deals with the manipulation of data present in the database.

Component of SQL statement that controls access to data and to the database.

It is mainly used to store, modify, retrieve, delete and update data in a database.

List of DML commands:

SELECT: Retrieve data from a database

- SELECT: This command is used to retrieve rows from a table. The syntax is SELECT [column name(s)] from [table name] where [conditions]. SELECT is the most widely used DML command in SQL.

INSERT: Insert data into a table.

- This command adds one or more records to a database table. The insert command syntax is INSERT INTO [table name] [column(s)] VALUES [value(s)].

UPDATE: Updates existing data within a table

- UPDATE: This command modifies data of one or more records. An update command syntax is UPDATE [table name] SET [column name = value] where [condition].

DELETE: Delete all records from a database table

- DELETE: This command removes one or more records from a table according to specified conditions. Delete command syntax is DELETE FROM [table name] where [condition].

Additional Source:

https://www.techopedia.com/definition/1179/data-manipulation-language-dml

https://www.geeksforgeeks.org/sql-ddl-dql-dml-dcl-tcl-commands/

Day 60: SQL - Entity Relationship Diagram (22/4/2022)

Topic: What is Entity Relationship Diagram (ERD)?

ERD is a snapshot of data structures. An ERD shows entities (tables) in a database and relationships between tables within that data.

A good database design it is essential to have an ERD.

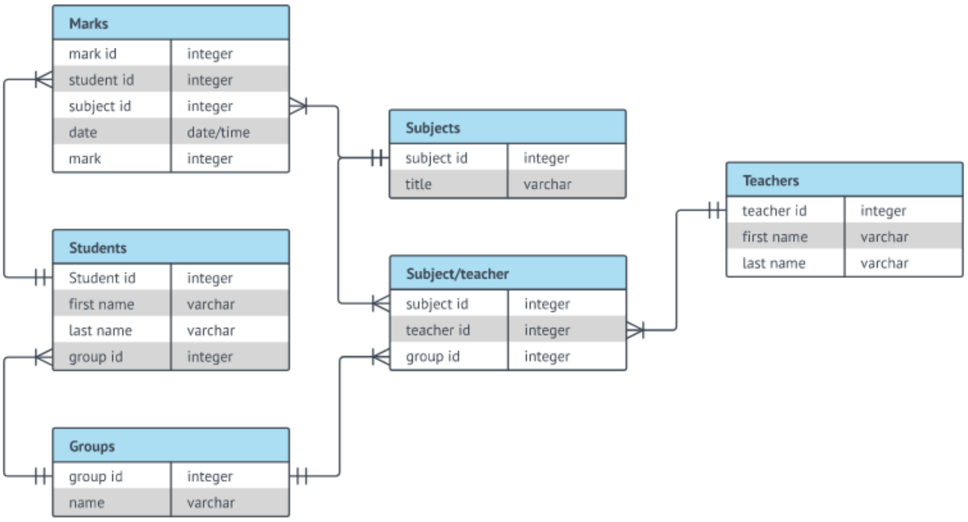
A type of flowchart that illustrates how “entities” such as people, objects or concepts relate to each other within a system.

Most often used to design or debug relational databases in the fields of software engineering, business information systems, education and research.

Three basic elements in ER-Diagrams:

Entities are the “things” for which we want to store information. An entity is a person, place, thing or event.

Attributes are the data we want to collect for an entity.



Marks	
mark id	integer
student id	integer
subject id	integer
date	date/time
mark	integer

Subjects	
subject id	integer
title	varchar

Teachers	
teacher id	integer
first name	varchar
last name	varchar

Students	
Student id	integer
first name	varchar
last name	varchar
group id	integer

Subject/teacher	
subject id	integer
teacher id	integer
group id	integer

Groups	
group id	integer
name	varchar

Additional Source:

https://www.datanamic.com/dezign/erdiagramtool.html

https://www.lucidchart.com/pages/er-diagrams

Day 61: SQL – Types of Relationships (23/4/2022)

Topic: How many types of relationships are there?

One to Many (1:M) relationship

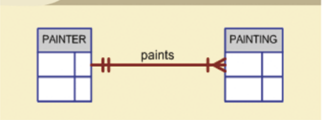
Should be the norm in any relational database design and is found in all relational database environments

Example:

One department has many employees.

One student register for multiple courses, but all those courses have a single line back to that one student.

FIGURE 3.18 The 1:M relationship between PAINTER and PAINTING

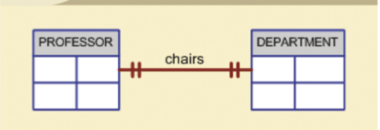


One to One (1:1) relationship

A one to one (1:1) relationship is the relationship of one entity to only one other entity. It should be rare in any relational database.

Example: One student associated with one mailing address

FIGURE 3.22 The 1:1 relationship between PROFESSOR and DEPARTMENT

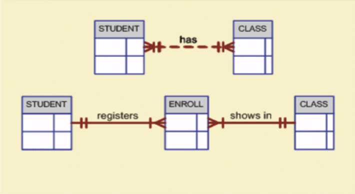


Many to many (M:N) relationships

Can be implemented by breaking it up to produce a set of 1:M relationships.

Can avoid problems inherent to M:N relationship by creating a composite entity or bridge entity.

FIGURE 3.27 Changing the M:N relationship to two 1:M relationships



Additional Source:

https://opentextbc.ca/dbdesign01/chapter/chapter-8-entity-relationship-model/

Day 62: SQL – Normalization Process (24/4/2022)

Topic: What is normalization in SQL?

Normalization

The process of organising data in a database. This includes creating tables and establishing relationships between those tables according to the rules designed.

Makre database more flexible by eliminating redundancy and inconsistent dependency.

Redundant Data: Wastes disk space and creates maintenance problems.

Ex: A customer address change is much easier to implement if that data is stored only in “Customer Table”.

Converting from Un-normalized Form (UNF) to First Normalized Form (1NF)

UNF: Data stored in a form that is not suitable to be implemented in the database.

- Present data in tabular / table format where each cell has single value.
- Identify Primary Keys (PKs)

Converting from First Normalized Form (1NF) to Second Normalized Form (2NF)

- Identify all dependencies and remove partial dependency
- Records should not depend on anything other than a table's primary key

Example:

Consider a customer's address in an accounting system. The address is needed by the Customers table, but also by the Orders, Shipping, Invoices, Accounts Receivable, and Collections tables.

Instead of storing the customer's address as a separate entry in each of these tables, store it in **one place**, either in the Customers table or in a separate Addresses table.

Converting from Second Normalized Form (2NF) to Third Normalized Form (3NF)

- Identify all dependencies and remove partial dependency
- Identify transitive dependency (if any) then remove it, else the tables are already in third normal form (3NF)
- Transitive dependency occurs when a non-key attribute is dependent on another non-key attribute

Additional Source:

https://docs.microsoft.com/en-us/office/troubleshoot/access/database-normalization-description#:~:text=Normalization%20is%20the%20process%20of,eliminating%20redundancy%20and%20inconsistent%20dependency.

Day 63: Data Management – Part 1 (25/4/2022)

Topic: What are the 11 key areas of Data Management?

Data Management Body of Knowledge (DMBoK)

Data Management is an overarching term that describes the processes used to plan, specify, create, acquire, maintain, archive, retrieve and control data.

Why DMBoK?

Standardisation of data management disciplines will help data management professionals perform more effectively and consistently.

11 Data Management Knowledge Areas:

NO. 1 Data Governance

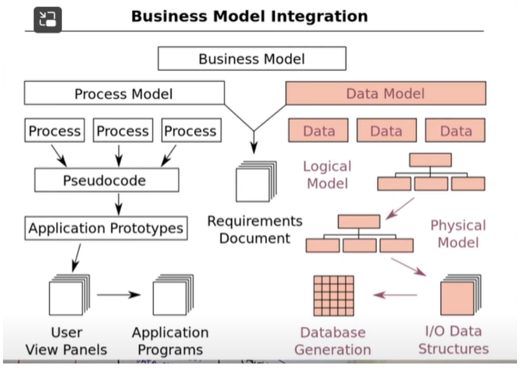
- Planning, oversight and control over management of data and the use of data and data-related resources.
- Generally, comprises guaranteeing an organisation's data availability and security.
- The purpose is to ensure that data is managed properly according to policies and best practices.
- Data Governance sits at the heart of Data Management and is all about strategies, policies and rules.

NO.2 Data Architecture

- The overall structure of data and data-related resources as an integral part of the enterprise architecture.
- Specifies how the overall data structure fits into the overarching enterprise architecture and how it will be embedded in standard operations.
- Data Architect designs the blueprint of enterprise data architecture.

NO.3 Data Modelling and Design

- To define and analyse data requirements.
- Design physical and logical models
- Should be aligned with the data architecture.
- Data Modeller.



Day 64: Data Management – Part 2 (26/4/2022)

Topic: What are the 11 key areas of Data Management – Part 2?

NO.4 Data Storage and Operations

Hardware storage, distribution and management of structured physical data assets.

Owner of how the data is acquired, maintained, recovery and ensuring end-to-end data pipeline are running on-time.

Data engineer will be pretty much responsible for this kind of activities.

NO.5 Data Security

Ensuring privacy, confidentiality and appropriate / proper access to individuals' private data.

Compliances, data authorisation and auditing process.

Cloud computing has made data security much more visible.

NO.6 Data Integration and Interoperability

Managing the integration of data between different lines of business applications.

Acquisition, extraction, transformation, movement, delivery, replication and virtualisation.

Extraction, Transform, Load (ETL)

Data Engineer's role: Integration pipeline (How the data will be acquired from the source.)

NO.7 Documents and Content

Storing, protecting, indexing and enabling access to data found in unstructured sources and making this data available for integration and interoperability with structured (database) data.

Manage, control, index, and secure access to data and information.

Managing all the unstructured kind of data (Images, files etc.)

Day 65: Data Management – Part 3 (27/4/2022)

Topic: What are the 11 key areas of Data Management – Part 3?

NO.8 Reference and Master Data (Master Data Management)

Managing shared data to eliminate redundancy and improve data quality by standardising the definition and use of data values.

Master data: Dimensions / entities which defines all business processes.

For example: System of records / Customer data (Managing all customers record)

NO.9 Data warehousing and Business Intelligence

Coordinating the processing of data management for analytics and providing access to decision support data for reporting and analysis.

Reporting & Analysis / Artificial Intelligence & Machine Learning

Data Analyst / BI experts / Data Scientist

NO.10 Metadata

Metadata collection, categorisation, maintenance, integration, control, management and delivery.

All the data of the data.

Data lake environment.

NO.11 Data Quality

Identifying, monitoring, maintaining and enhancing data integrity and quality

Ensure data integrity is in place.

Good data vs Bad data.

Additional Source:

What are the 11 key areas of Data Management and specific data roles? - YouTube

https://www.dama-dk.org/onewebmedia/DAMA%20DMBOK2_PDF.pdf

https://theecmconsultant.com/what-is-data-management/

https://www.linkedin.com/pulse/data-management-knowledge-areas-jay-zaidi/

Day 66: Data Science & Cloud Computing (28/4/2022)

Topic: What is Cloud Computing?

Essentially, Cloud Computing allows companies to access different computing services like databases, servers, software, data analytics etc. over the internet (Cloud).

Companies would be able to run their applications on the best data centers in the world with minimal costs.

In turn, this ensures that small companies can use this technology for ambitious and complex projects that would otherwise be quite costly.

Advantages / Benefits of Cloud Computing

Companies can use the cloud to host their data and they don't need to worry about servers anymore.

Companies can now access server architecture in the cloud according to their needs and even save money by only paying as much as the data they are using on the cloud.

Cloud Computing Platforms for Data Science:

- **Amazon Web Services (AWS)**

Subsidiary of Amazon. AWS provides various products for data analytics:

Amazon QuickSight – Business Analytics Service

Amazon Redshift – Data Warehousing

Amazon EMR – Big Data Processing

Amazon Aurora / DynamoDB – Relational Database / NoSQL Database

Etc.

- **Google Cloud Platform**

BigQuery - Data Warehouse

Dataflow - Streaming Analytics

Dataproc - Apache Spark clusters

Etc.

- **Microsoft Azure**

Additional Source:

https://www.geeksforgeeks.org/why-cloud-computing-is-important-in-data-science/