***Topic 1: Type 1 & 2 error***

**Type I Error (False Positive)**

In statistics, **Type I error** (also known as a false positive) occurs when you incorrectly reject a true null hypothesis. The **significance level** (denoted as α) is directly related to the probability of making a Type I error.

**Type II Error (False Negative)**

A **Type II error** occurs when you fail to reject a false null hypothesis, also known as a false negative. The probability of committing a Type II error is denoted by **β**.

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.

**Relationship between Type II error and significance level (α):**

* **Significance level (α)** and **Type II error (β)** are inversely related, but not directly. Lowering the significance level (reducing the chance of a Type I error) can increase the probability of a Type II error, and vice versa.
* **Power of the test**: The probability of correctly rejecting a false null hypothesis is called the **power of the test** (1 - β). If you reduce the Type II error rate (β), you increase the power of the test. To increase power, you might need to increase sample size or adjust the design of the experiment.

**In summary:**

* **Type I error (α)** and **Type II error (β)** are somewhat in balance. Lowering α reduces the chance of making a Type I error but can increase the chance of making a Type II error.
* Increasing **sample size** or improving study design can help reduce both Type I and Type II errors without having to choose between them.

***Topic 2: Resampling***

**What is Sampling?**

Sampling is an active process of **gathering observations** with the intent of estimating a **population variable**.

**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

* Bootstrap

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

Stratified bootstrapping involves generating multiple resamples from the dataset while ensuring that the class proportions (for both majority and minority classes) are maintained in each resample.

* 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**.

In **stratified cross-validation**, you divide the data into training and validation sets, but ensure that each set maintains the same class distribution as the overall dataset. This is particularly useful in model evaluation.

***Topic 3: Confusion Matrix***

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.

A screenshot of a computer

Description automatically generated

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

➔ 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

➔ 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.

A diagram of a negative result

Description automatically generated with medium confidence

***Topic 4: AUC-ROC and AUC-PR***

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



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

➔ TPR = TP / (TP + FN)

**False Positive Rate (FPR)**

➔ FPR = FP / (FP + TN)

An ROC curve plots TPR vs. FPR at different classification thresholds. Sometimes FPR will be replaced by precision, especially for imbalance dataset.

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

**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.

A diagram of a positive rate

Description automatically generated

***Topic 5: Handling Missing Values***

Deletion or Imputation

1. **A. Deletion Method**
2. • **Not a robust solution**. Recommended to **only use** this technique when the dataset contains **fewer** missing values. (either rows or columns)
3. **B. Data Imputation**
4. Impute the missing values with aggregated value like mean, median, mode of a column

* With mean when continuous data with no outliers.
* With median when continuous data with outliers
* With mode when categorical data
* With mean/median/mode based on other categorical columns: similar category may have similar properties.
* If there is a trend, we may need to use linear interpolation
* If there is a seasonality, we may need seasonal adjustment + linear interpolation
* Build a model to predict the missing value (linear regression, logistic regression, K nearest neighbors)

***Topic 6: Basic Statistics***

**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

A screenshot of a graph

Description automatically generated

***Topic 7: Bias-Variance Tradeoff***

A diagram of a graph

Description automatically generated with medium confidence

How to reduce variance/improve overfitting model?

Ensemble methods like bagging, boosting(it may not help on overfitting)

Feature selection and engineering: remove irrelevant or redundant features

Add regularization term.

***Topic 8: Imbalance Dataset***

Dealing with imbalanced datasets, where one class (the majority class) significantly outweighs another class (the minority class), is a common problem in many real-world applications, such as fraud detection, medical diagnosis, and customer churn prediction. Imbalanced data can lead to biased models that perform well on the majority class but poorly on the minority class. Several techniques can be used to address this problem, including data-level, algorithm-level, and evaluation-level approaches.

1. **Data-level: Resampling (over-sampling, under sampling, SMOTE)** 
   1. **Note: only for training dataset, but not for testing dataset**
   2. **For data split, we need to use stratified splitting or stratified K-fold cross validation.**
2. **Algorithm/Model level:** 
   1. **Update loss function (Tree-based and Support Vector Machines can be adapted by modifying their loss functions to include class-specific class)**
   2. **Logistic regression (adjust probability threshold)**
   3. **Ensemble (boosting – naturally helps with imbalanced data; random forest with under sampling data)**
3. **Evaluation level**

Recall, F1 score, precision, AUC-ROC (balanced classes), PR-ROC (prone to imbalanced classes.

***Topic 9: Outliers***

How to detect outlier?

**Box plot, scatter plot, histogram, distribution plot.**

How to handle outlier?

**1. Remove the outliers**

o **Remove outlier values** from the dataset to stop them from skewing our analysis.

o May not be a good idea if we have a small dataset.

**2. Imputation**

o Alike with imputation of missing values, we can also **impute outliers**.

o Mean, median method can be used but **median** would be **more appropriate** as it is not affected by outliers.

**3. Transforming values**

o Transforming variables can also eliminate outliers and reduces the variation caused by extreme values.

o Scaling, Log Transformation, Cube Root Normalization, Box Transformation

***Topic 10: Linear regression***

Ordinary least square (OLS) = maximum likelihood

A close up of a number

Description automatically generated with medium confidenceA mathematical equation with numbers and symbols

Description automatically generated

A screenshot of a test

Description automatically generated

**Measurement**

In linear regression, **R-squared (R²)** is a statistical measure that represents the proportion of the variance in the dependent variable (the variable you're trying to predict) that is explained by the independent. A black and white math symbols

Description automatically generated

A screenshot of a computer

Description automatically generated

How we choose loss function depends on how we treat outliers. The outliers are closer to the model trained with MSE than to the model trained with MAE.

***Topic 11: Multicollinearity***

Multicollinearity exists whenever an **independent variable** is **highly correlated** with one or more of the other **independent variables** in a **multiple regression** equation.

**Why Multicollinearity is a problem?**

It will cause unreliable coefficients and increases the standard errors of the coefficients. It makes some variables statistically insignificant when they should be significant.

**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.

➔ 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?**

* 1. **Remove highly correlated** predictors from the model.
  2. Combine correlated variables
  3. 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
  4. Add penalty (L2 Ridge regression is particularly effective when the independent variables are highly correlated (multicollinearity))

L1 Regression (Lasso Regression) A black text on a white background

Description automatically generated

L2 Regression (Ridge Regression): A black text on a white background

Description automatically generated

**Differences between L1 and L2 regression:**

* **Penalty type:** L1 uses the absolute value of the coefficients (L1-norm), while L2 uses the square of the coefficients (L2-norm).
* **Feature selection:** L1 regression can result in sparse models with some coefficients exactly zero, performing feature selection. L2 regression will not shrink coefficients to zero, but it reduces the magnitude of all coefficients.
* **Use cases:** L1 is often used when you have many features and suspect that only a subset is important. L2 is used when you want to keep all features but reduce overfitting by shrinking coefficients.

***Topic 12: Feature Engineering***

**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.

**What are the processes of Feature Engineering?**

**Feature Selection**

✓ 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**

✓ Imputation techniques were used to handle missing values.

**Handling outliers**

✓ 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**

✓ It is a technique that converts the categorical data in a form so that they can be easily understood by machine learning algorithms

**Feature scaling**

A screenshot of a white sheet with black text

Description automatically generated

***Topic 13: Gradient Descent, Stochastic Gradient Descent***

Gradient Descent is an optimization algorithm used to minimize a function by iteratively moving towards the minimum of the function.

A math equations and formulas

Description automatically generated with medium confidence

A white background with black text

Description automatically generated

A math equations on a white background

Description automatically generated

A math equation with black text

Description automatically generated with medium confidence

A white text with black text

Description automatically generated

Stochastic Gradient Descent uses not all data but only one sample or a subset of samples (mini-batch) to run gradient Descent.

***Topic 14: Decision Tree***

The decision tree algorithm splits the dataset at each node by selecting the feature that best separates the classes (for classification) or minimizes variance (for regression). Several metrics can be used to evaluate the quality of the split, with **Gini Impurity** being one of the most common for classification tasks.

**Gini Impurity (other metrics: entropy or information gain)**

A number and a symbol

Description automatically generated with medium confidence

In general, the lower the Gini Impurity, the better the split. The goal of the decision tree algorithm is to minimize Gini Impurity at each split to create nodes that are as pure as possible.

**Total Gini impurity = weighted average of Gini Impurities for the leaves**

For regression tree, a commonly used metric is **sum of square of residuals (SSR).**

In general, the lower the sum of square of residuals, the better the split.

1. **Pruning the Tree**

**a. Pre-Pruning (Early Stopping)**

**Stopping Criteria**: Maximum depth, Minimum samples per leaf node, Minimum samples per split, Minimum impurity decrease

b. **Post-Pruning**

**Cost Complexity Pruning (CCP)**: This method prunes the tree by introducing a penalty for complexity. (Tree score = SSR (or total gini impurity) + aT) -- a is the hyper parameter we will tune through cross validation.

A screenshot of a computer

Description automatically generated

***Topic 15: Logistic Regression***

Logistic Regression is a “Supervised machine learning” algorithm that can be used to model the probability of a certain class or event.

A math equation with black text

Description automatically generated

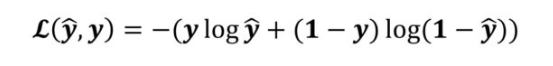
A math equation with black text

Description automatically generated

A white background with black text

Description automatically generated

The cross-entropy loss function is used to measure the performance of a classification model whose output is a probability value.



***Topic 16: Statistics Concept***

Bayes’ Theorem (Bayes’ Rule)

A white paper with black text

Description automatically generated

Law of total probability

A white background with black text

Description automatically generated

Random Variables:

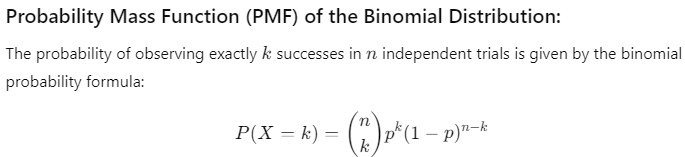
A white background with black text

Description automatically generated

***Topic 17: Probability Distribution***

A white background with black text

Description automatically generated





A white background with black text

Description automatically generated



A math equation with black text

Description automatically generated

A black text on a white background

Description automatically generated

Common Questions and Answers

1. Can you explain P value?

The P-value is a number that helps you understand how likely it is that the results you see in your data could have happened by random chance. A small P-value means that your results are less likely to be due to chance, and you may have found something significant. A larger P-value suggests that your results could easily be explained by random chance, so there may not be anything special going on.

For example, you're flipping a coin. You hypothesize that the coin is **fair**, meaning it will land heads or tails with equal probability (50%). Now, you perform an experiment by flipping the coin 10 times, and you observe 9 heads and 1 tail. P value will be the probability that we have 9 heads or even more extreme results if we flip a fair coin, which is around 1%. If we set up significant level at 5%, we are confident to say this coin is not fair.

1. Can you explain confidence Interval?

A confidence interval gives us an estimated range that is likely to contain the true value of what we are trying to measure, along with a degree of certainty. Typically, the certainty is expressed as a percentage, like 95% or 99%. If we have a **95% confidence interval**, it means that if we took many samples from the population and calculated a confidence interval from each one, **95% of those intervals would contain the true population value.** For example, imagine you’re conducting a survey to estimate the average height of adults in a city. Since it’s impossible to measure everyone’s height, you randomly select a sample of 100 adults and calculate confidence interval. This means that you are **95% confident** that the true average height of all adults in the city lies in the confidence interval. if you repeated this survey multiple times with different samples of 100 people, about 95% of the intervals you calculate would contain the true average height.

1. What is central limit theorem?

The central limit theorem (CLT) states that the distribution of sample means approximates a normal distribution as the sample size gets larger, regardless of the population's distribution. Sample sizes equal to or greater than 30 are often considered sufficient for the CLT to hold.

1. How to you decide sample size?

***Math Review #1:***

**Arithmetic Sequence (Progression)**

A math equations on a white background

Description automatically generated

Sum of an Arithmetic Progression:

A white background with black text

Description automatically generated

**Geometric Sequence (Progression)**

A white background with black text

Description automatically generated

Sum of a Geometric Progression:



A math equation with black text

Description automatically generated

***Math Review #2:***

Counting:

A white background with black text

Description automatically generated

Special case:

A white background with black text

Description automatically generated

A white background with black text

Description automatically generated

A white paper with black text

Description automatically generated