**TELL US ABOUT YOURSELF (Background & Experience)**

First of all, thank you very much for giving me this opportunity to be interviewed for the Data Scientist 1 position at your company. My Name is Trishad Phogole. I hold a **BSc in Mathematical Sciences** and a **Honours degree in Computer Sciences** from the **University of Limpopo**. I am a **Data Scientist at Mintek** (if current employee) or can just say I am a **Data Scientist**, with over one/two years hands-on work/internship experience in **data analytics**, **machine learning**, and **business intelligence within the mining sector**.

In my current/previous role at Mintek, I developing, test, and deploy machine learning models to optimise organisational operations and enhance decision-making. I also, led the development and maintenance of organisational Key Performance Indicators (KPIs) dashboards for real-time tracking and reporting.

My key strengths include **strong problem-solving skills**, **attention to detail**, and the **ability to translate complex data into meaningful insights** that drive business value. I am passionate about data and its power to influence strategy in a business. Lastly, I am committed to ensure that my contributions will have a meaningful and valuable impact on your organisation.

**Why do you want to work here?**

I have been aware of your organisation for many years, as you know you’re a market leader your innovative and creative, and what I have seen you put your customers and clients first. And what I understand during my research you have very ambitious plans for the future, and I would like to be a part of those, and that is great for my long-term career future plans. So that are the main reasons I see myself working here for a long period of time, and I would be able to grow as your company grows as well.

**Tell us about your Certifications. Also, the projects for each certification, how did you use them for? What have you learned there**

**Understand and Remember all your Github Projects**

**Machine Learning Projects**

**Project:** Credit Risk Assessment using Logistic Regression, Random Forest, and Decision Tree Classifier

**Project:** Credit Scoring using Decision Tree Classifier

**Project:** Loan Eligibility Prediction using Logistic Regression, RandomForest and DecisionTree classifier

**Project:** Credit\_Fraud\_Detection using Logistic Regression

**Project:** Customer Segmentation Using K-Means and PCA

**AI Project**

**Project:** LLM Chatbot Using Google Gemini Pro

**Microsoft Power BI Projects**

Car Sales Dashboard

HR Dashboard

Sales Dashboard

**SQL Projects**

Job Analysis: Top Highest Paying Jobs

Library Management System

Retail Sales Analysis

**What do you understand about a Bank or Banking?**

**Bank**

Is a **financial institution allowed/licensed** to accept deposits, provide loans & credits, facilitate payments, & manage investmets **from customers**.

**Banking**

Refers to all the activities carried out by banks to **provide financial services** — such **as lending money**.

**What is Mintek all about, what do they do?**

* Mintek is a national mineral research organisation and a global leader in **mineral processing** & **metallurgy**

Metallurgy is the **science and technology of metals** – it deals with how we **extract, process, refine, and use metals**.

* **Mineral Processing** = “physical separation” (getting valuable minerals out of ore).
* An **ore** is a type of rock from which we can extract metals or minerals that are valuable for industrial and economic use.
* **Extractive Metallurgy** = “chemical separation” (getting pure metal out of minerals).
* Ores are mined, then go through **mineral processing** (crushing, grinding, concentration) and **extractive metallurgy** (smelting, leaching, electrolysis) to obtain pure metals.

**RESEARCH OUTPUTS**

Are the tangible results (**journal papers, conf papers, etc**) that come from scientific investigations, experiments, or applied studies. Document findings on ore deposits, mineral processing, metallurgy, automation, sustainability, etc.

**What are your core values?**

My core values are **curiosity**, **accountability and collaboration**. **Curiosity drives** me to continuously learn new analytical techniques and stay updated in the rapidly evolving data science field. **Accountability** ensures I take ownership of my work and deliver accurate, actionable insights that the team and clients can rely on.

**Collaboration** – I believe that great insights and innovative solutions come from teamwork. I value open communication, knowledge sharing, and learning from others’ perspectives to build models and strategies that create real impact.

**BANKING VS FINTECH**

**What do you understand about FinTech?**

**FinTech**

The use of technology to improve financial services. Online banking, mobile wallets

**Statistical Foundation: A deep understanding of statistical concepts, experimental design, and modelling techniques. Which are?**

**Statistical Concepts**

* Descriptive statistics → mean, median, variance, standard deviation, distributions.
* Probability theory → random variables, probability distributions (Normal, Binomial, Poisson, etc.).
* Hypothesis testing → null/alternative hypotheses, p-values, t-tests, chi-square tests, ANOVA.
* Confidence intervals → estimation and margin of error.
* Correlation & causation → Pearson/Spearman correlation, confounding variables.
* Bayesian statistics → prior, posterior, likelihood.

**Experimental Design**

* Sampling methods → random, stratified, cluster sampling.
* Bias & variance → sources of error in experiments.
* Randomized controlled trials (RCTs) → gold standard for causal inference.
* Factorial design → studying effects of multiple variables at once.
* A/B testing → comparing two versions of a system/product.

**Modelling Techniques**

* Regression analysis → linear regression, logistic regression, polynomial regression.
* Classification models → decision trees, SVM, Naïve Bayes.
* Clustering → k-means, hierarchical clustering.
* Time-series analysis → ARIMA, exponential smoothing, seasonal decomposition.
* Survival analysis → hazard functions, Kaplan–Meier estimator.
* Multivariate methods → PCA (Principal Component Analysis), factor analysis.

**What do you understand about Credit Risk, Credit Risk Modelling & Credit Scoring**

**Credit risk** is the risk of financial loss that arises when a borrower (individual, company, or institution) fails to meet financial obligations (e.g., repay a loan).

**Example 1: Personal Loan (Retail Banking)**

**Typical Example:** A bank lends R10,000 to an individual for a car loan. The individual loses their job six months later. **This is a Credit Risk** because now the bank faces a potential loss of the remaining loan interest.

**Credit Risk Modelling**

So, banks use credit risk modelling to **predict likelihood** **that a borrower will default (missing scheduled payments, etc) and to measure the potential financial losses that the bank will face in the future.**

**These models help financial institutions evaluate and manage credit risk by estimating key parameters such as:**

* **Probability of Default (PD):** The likelihood that a borrower will fail to meet their debt obligations within a specific timeframe.
* **Loss Given Default (LGD):** The estimated percentage of loss the lender incurs if the borrower defaults.
* **Exposure at Default (EAD):** The amount of exposure or outstanding loan balance at the time of default.

**Example:** At this stage the problem knows it has a problem, but it does not know how big the problem is. So, it uses the credit risk model to measure the potential financial loss.

**Credit Risk** is the fundamental business problem.

**Credit Risk Modelling** is the quantitative solution to **understand** and manage that problem.

**What is IFRS 9?**

It is an international financial reporting standard that many banks comply to, it requires a forward-looking approach. Banks must now recognize **Expected Credit Losses (ECL)** = **PD\*LGD\*EAD** before a **default** actually happens. So that they can improve their financial stability.

**The Three Stages of IFRS 9**

**Stage 1: Performing Assets**

* **Definition:** Loans or financial assets that are new or have not shown a significant increase in credit risk since origination.
* **Loss Measurement:** Entities recognize **12-month Expected Credit Loss (ECL)** — losses expected over the next 12 months only.

**Stage 2: Underperforming Assets**

* **Definition:** Assets where credit risk has increased significantly since origination but there is no actual default yet.
* **Loss Measurement:** Entities recognize **Lifetime ECL** — losses expected over the entire life of the asset.

**Stage 3: Credit-Impaired Assets**

* **Definition:** Assets that are in **default** or are **credit-impaired** (e.g., borrower is unlikely to pay or has missed payments).
* **Loss Measurement:** Entities recognize **Lifetime ECL** (like Stage 2).

**Simplified Example:**

* **Stage 1:** A customer takes a new loan and pays on time → bank sets aside small 12-month expected loss.
* **Stage 2:** The same customer starts missing payments or shows signs of financial stress → bank must now set aside lifetime losses.
* **Stage 3:** Customer defaults or loan becomes impaired → lifetime losses recognized, and interest is only on what the bank realistically expects to recover.

**What is Credit Scoring?**

Is mainly used by organisations (banks,etc) to **assess whether a customer qualifies for a credit** (loan or credit card) based on their income, spending behaviours etc

**What is Impairments?**

Requires organisations /banks or businesses to recognise expected credit losses from the time a financial asset is originated. This rule means that as soon as a bank (or any company) gives out a loan or credit, it must immediately set aside some money to cover the risk that the borrower might not pay back in the future.

So instead of waiting until the borrower actually defaults, the bank has to **recognize the “expected” loss upfront**. This makes financial reporting more realistic and ensures banks are prepared for possible losses.

**What is Impaired?**

An asset (like a loan, bond, or receivable) is **impaired** when there is a high likelihood that the borrower will **not be able to pay back** the full amount owed.

**What is IFRS 9 in Credit Risk Modelling and Credit Scoring**

IFRS 9 changed how banks build models. Instead of recognizing losses only when a borrower defaults, banks now must use models to estimate future expected losses.

Credit risk models now need to be forward-looking, not just historical.

**What do you understand about Insurance?**

Insurance is a **financial arrangement** where an individual or organization pays a fee (called a premium) to an insurer in exchange for protection against financial loss from uncertain events.

To reduce financial risk and provide security in case of accidents, illness, property damage, or other losses.

**Take us through your CV/Resume**

I hold a bachelor’s degree in mathematical science majored in Statistics and Computer Science and a Honours degree in Computer Sciences from the University of Limpopo.

My **Technical skills** include:

I am a **data scientist** at mintek

**What is your salary expectation?**

I am open to discussing a fair market-based salary for this role. Based on my qualifications, certifications and experience, I’d expect a range around 250 000 to 500 000, but I am flexible depending on the overall package and growth opportunities.”

My priority is finding a role that’s a strong fit for my skills and career goals. I am confident we can agree on a competitive salary that reflects my experience and the responsibilities of the role.

If they ask for an exact number, ask them the salary range for this position or how much people earn for this position in your organisation.

**Do you have any questions for us?**

**Yes,**

**Why are you leaving your current/previous role?**

Well, for me Is all about growth, I am looking for a role where I can work on more advanced data science projects and continue to grow my skills in the data science field and obviously gain more experience as well. So, I think this opportunity at FNB aligns perfectly with my goals.

**The Projects you did ( Mintek and Me)**

**Science Vote Project**

**Science Vote** is a government-funded allocation from the Department of Science and Innovation (DSI) that supports Mintek’s core research and development activities.

Duty:

**Before (Reporting Focus) – Science Vote Dashboard**

This dashboard focuses on research projects funded under mintek. My role involved monitoring and reporting on this project progress (Quartely Reporting). The dataset contained **Total Overall Revenue Approved** per PI (Principal Investigators), **Project title**, **Planned Revenue**, **Planned cost**, **Actual cost**, **Actual Revenue**, **Total amount spent to date**, and **funding utilisation rate**.

**After (The Machine Learning Focus) – Profit or Loss for the past 10 years**

They also wanted me to come up with a way that they can measure performance per project for the past years whether they runned profit or loss. So, I had to come up with a plan to make a machine learning model that can help us predict whether a project runned for a loss or profit, and the figures (How much) for the past 10 years.

So, I developed a machine learning model using logistic regression.

**Input features(X):** Projects,Planned Revenue**,**Planned Costs**,**Actual revenue, Actual cost

**Target (Y):** Profit or Loss

Divided data into two classes:

Profit =0

Loss = 1

**Dealing with Imbalances**

Created a **sample data** containing data from profit, to ensure profit and loss both have the same count of classes. Created a **new dataset** containing the sample dataset and loss Using **pd.concat ( sample, loss)**

Planned Revenue

**The total amount of funding that has been formally approved for a project**, combining all sources of funding (e.g., Science Vote allocations, external grants, and other contributions) over the lifetime of the project.

**Planned Revenue**

* **What it is:** The **expected income** or funding amount approved for the project at the beginning. It's a forecast or a goal set in the official budget.

**Planned Costs (or Budgeted Costs)**

* **What it is:** The **approved budget** for the project—the detailed plan for how much can be spent on various activities (salaries, equipment, materials) to achieve the project's goals.

**Actual Costs**

* **What it is:** The **real, exact amount of money that has already been spent** on the project. This comes from invoices, payroll, and receipts.

**Actual Revenue**

* **What it is:** The **funding the project has actually generated or received**.

Staff members & those studying

**Strategic Objectives (KPIs):** Understanding Research Outputs, etc

Science Vote Dashboard

**HR:**

SET-HCD Development Dashboard - Staff members & those studying

Interns Dashboard

Career Ladder Dashboard

MACHINE LEARNING PROJECTS:

**Target vs Feature in ML**

**A feature** is an **input variable** that the model uses to make predictions.

**The target** is the **output variable** the model is trying to predict

* **Credit Fraud Detection System**

**Aim:** To detect transaction as either Fraud or legitimate. Using **logistic Regression**

* **Customer Segmentation**

Customer Segmentation – Mall Customer Data, to analyse mall customer data to segment shoppers into distinct groups based on their demographics, spending habits. Using clustering techniques like K-Means, I was able to identify high-value, loyal customers, which can help the mall target marketing strategies more effectively.

* **Predicting House Prices**

To predict house prices based on input features such as number of rooms, size of the room. Using **linear regression**

* **Loan Eligibility Prediction**

Predicting loan eligibility based on applicant data

* **Credit modelling & Credit Scoring**
* **IFRS 9 Model development and Validation**

**Why pursued your studies under computer science and why do you want to be a Data Scientist or Data Engineer?**

I have always been interested in problem-solving and technology. So, data science just gives me that platform to express my problem-solving skills

**Why should we hire you?**  
Well, I am sure that other candidates brings a lot to the table also, but what sets me apart is, when I am put head to head with anyone, I out work them everytime, I follow through, I follow up and when I mess something up I own it, and learn from it.

In my previous role as a **Data Scientist at Mintek**, I worked on the **Science Vote Project**, my role focused on developing a **machine learning solution** that could evaluate **project profitability over the past 10 years** ,predicting whether a project ran at a **profit or loss**, and by how much. The final model successfully classified project outcomes with high accuracy, allowing leadership to quickly assess financial performance trends and make better funding allocation decisions for future research projects.

So, if you are looking for a Data Scientist who can turn complex datasets into meaningful, measurable impact , improving accuracy, decision-making, and operational efficiency that is exactly what I bring to the table.

**What is your biggest weakness?**I am a perfectionist, I tend to focus a lot on making sure everything is perfect, which can sometimes slow me down. However, I have learned to balance my perfection and efficiency by prioritising tasks.

**What are your biggest strengths?**I have strong analytical and problem-solving skills, especially in working with data. I am also highly adaptable and flexible, which allows me to quickly learn and apply new technologies.

**What are you passionate about?**

I am passionate about using technology to solve real-world problems. Because technology is evolving verything now is about AI, and I like to stay up to date with emerging technologies.

**What are your goals for the future?**

My goal is to expand my knowledge in my field, step into leadership positions within my field, and play a key role in developing innovative solutions that address real-world challenges.

**How do you handle stress and pressure?**

I manage stress by staying organized and prioritizing tasks effectively. It helps me stay productive and deliver quality work even in challenging situations.

**What motivates you?**

I have always been self-motivated, I know that if I want to reach my goals and make progress, I have to take action, nothing will happen on its own, that mindset has pushed me to stay disciplined , keep learning and follow through even if its challenging

**What are your goals for the future? Or where do you see yourself in 5 years? What are you hoping to gain from this role**

In five years, I see myself growing within your organisation, having developed a deep understanding of its goals, culture, and strategic priorities. My career progression plan is firmly aligned with this organisation, and I aim to take on a more senior role where I can lead projects, mentor junior team members, and contribute to innovative solutions that address key business challenges.

I want to be part of initiatives that drive measurable impact, whether through improving efficiency, enhancing data-driven decision-making, or supporting the company’s broader vision. Ultimately, I see myself as a trusted and valuable contributor who helps the organisation achieve sustainable success.

**In this role I hope** to **gain practical experience**, expand my technical skill set, and deepen my understanding in the Data Science field.

**Why do you want to be a Data Analyst or What motivates you? Personally, and Professionally**

I want to be a data analyst for **three reasons**:

First, it is a role that utilises my best skills: First, Analysis, problem-solving, critical thinking and SQL. Second, it is a role that empowers me to have a positive impact on an organisation and that brings me a great deal of job satisfaction. Finally, I want to be a Data Analyst because it is a role that is constantly evolving.

**Why would you like to join FNB ( )?**

As a data scientist, I believe it is essential to work in an environment that is both supportive and forward thinking, qualities I found consistently highlighted in my research about your company.

Your organisation value their customers; you always want to ensure that your customers gets quality services. And also, what stood out to me was your strong industry reputation, consistent track record of success.

**Why should we hire you? It must be related to the job**

What sets me apart is because I bring a strong mix of **technical expertise** and **business impact**. Also, my ability to deliver end-to-end solutions from **data cleaning** and **feature engineering** **to model deployment and monitoring** while always **aligning projects to business strategic objectives.** And lastly, I’m confident I can add value to your team by not just **building models**, but by **ensuring they drive measurable improvements for your organization.**

**How do your skills and experience align with this role?**

My skills and experience align closely with this role in several ways. I have hands-on experience in data analytics, machine learning, and business intelligence, which allows me to extract insights from complex datasets and build predictive models.

In addition, I have experience with data visualization and reporting, creating dashboards that communicate insights effectively to both technical and non-technical stakeholders. My proficiency in Python, SQL, and relevant libraries like Pandas, NumPy, and Matplotlib, combined with practical exposure to Alteryx and Power BI, equips me to handle the end-to-end data science workflow, from data preparation to model deployment.

**QUESTIONS ACCORDING TO CV**

**General / Background**

* Can you walk us through your career journey and how you transitioned into data science?
* Why did you choose to specialize in data analytics, machine learning, and AI?
* What project in your portfolio are you most proud of and why?

**Technical Skills (Python, SQL, ML, AI)**

* How have you applied Python libraries such as Pandas, NumPy, Scikit-learn, or TensorFlow in your projects?
* Can you explain a machine learning model you built at Mintek and how you evaluated its performance?
* You mentioned achieving a 20% improvement in model selection through A/B testing. How did you design and implement that test?
* In your experience, what are the trade-offs between SQL and NoSQL databases for big data analytics?
* Can you explain how you optimized query performance in SQL to achieve 99% data accuracy in reports?
* Describe a time you had to fine-tune a machine learning algorithm. What techniques did you use (e.g., hyperparameter tuning, feature engineering)?
* How would you explain the difference between supervised and unsupervised learning to a non-technical stakeholder?

**🔹 Data Engineering & Cloud**

* What is your experience with ETL processes, and how did you streamline reporting using multiple data sources (SharePoint, Excel, databases)?
* You’ve worked with Hadoop and Spark—can you describe a scenario where distributed computing was essential?
* How have you leveraged AWS cloud services in your projects?

**🔹 Business Intelligence & Impact**

* Tell us about the KPIs dashboard you developed. How did it help the organization reduce manual reporting time by 40%?
* How do you approach translating complex datasets into actionable business insights?
* Give an example of how your work directly contributed to operational efficiency or performance improvement in your previous role.

**🔹 Statistics & Analysis**

* How would you design and conduct an A/B test for a new product feature?
* Can you explain the role of hypothesis testing in your work?
* How would you detect and handle bias in your datasets or machine learning models?

**🔹 Behavioral / Soft Skills**

* Describe a challenging data problem you faced and how you solved it.
* How do you collaborate with cross-functional teams when stakeholders may not fully understand technical details?
* Give an example of when you had to balance accuracy with efficiency in your work.

**🔹 Future & Continuous Learning**

* What emerging AI or ML technology excites you the most right now?
* How do your certifications (IBM, Udemy, AWS) influence your approach to projects?
* Where do you see yourself in the next five years in your data science career?

**TECHNICAL QUESTIONS**

**WHAT IS YOUR UNDERSTANDING OF DATA SCIENCE/ OR BEING A DATA SCIENTIST/ MACHINE LEARNING DATA ENGINEERING/DATA ENGINEER/ BI ANALYST/BI / DATA ANALYST? AND WHY DO YOU THINK WE NEED MORE DATA SCIENTIST OR DATA ENGINEERS?**

Data Science / Data Scientist

Data Science is all about extracting insights and knowledge from data using machine learning to drive business decisions. A data scientist develops predictive and prescriptive models, performs advanced analytics, and communicates actionable insights to stakeholders. They often work end-to-end: from data collection and cleaning, exploratory analysis, feature engineering, model development, to deployment and visualization

Data Analysis / Data Analyst  
Data analysis is the process of examining, cleaning, and interpreting data to extract meaningful insights that support business decisions. A data analyst collects and organizes data from various sources, identifies trends or patterns, performs statistical analyses, and visualizes results using tools like Excel, Power BI, or Tableau.

Data Engineering / Data Engineer  
Data engineering is about building and maintaining the infrastructure and pipelines that allow data to flow efficiently from sources to storage and analytics platforms. A data engineer ensures data is clean, reliable, and accessible, often working with databases, ETL processes, and cloud storage solutions.

Machine Learning / ML Engineer  
Machine learning

Machine Learning (ML) is a field of Artificial Intelligence (AI) that focuses on building systems/models that can learn patterns from data without being explicitly programmed with rules. An ML engineer implements these models in production, optimizing for scalability, reliability, and performance, often integrating them into applications or business processes.

BI / BI Analyst (Business Intelligence)  
Business Intelligence is the process, technology, and practice of collecting, analyzing, and presenting business data to support better decision-making. A BI analyst collects, cleans, and visualizes data using tools like Power BI or Tableau, creating dashboards and reports that support strategic decision-making. The emphasis is on understanding trends, monitoring KPIs, and supporting business stakeholders.

Data Analysis / Data Analyst  
Data analysis is the process of examining, cleaning, and interpreting data to extract meaningful insights that support business decisions. A data analyst collects and organizes data from various sources, identifies trends or patterns, performs statistical analyses, and visualizes results using tools like Excel, Power BI, or Tableau.

We need more Data Scientists and Data Engineers because organisations are generating massive amounts of data, and that data is only valuable if it can be processed, analysed, and interpreted effectively. Data Engineers build and maintain the infrastructure and pipelines that make data accessible and reliable, while Data Scientists use that data to generate insights, optimise processes, and support strategic decisions. Together, they help businesses make data-driven decisions, innovate, and stay competitive.

**What is Machine Learning?**

**Machine Learning (ML)** is a field of Artificial Intelligence (AI) that focuses on building systems that can **learn patterns from data** **without being explicitly programmed with rules**. **Example 2** – **Classification:**

Email spam detection: ML models learn to classify emails as “spam” or “not spam” based on words, sender info, and past patterns.

**Understand when to say a problem is a Classification or Regression**

**MACHINE LEARNING WORKFLOW**

**Machine Learning Workflow**

1. **Problem Definition & Goal Setting**
   * Clearly define the business problem and translate it into a machine learning problem.
   * **Example:** Reduce customer churn → Predict probability of churn **(classification).**
2. **Data Collection**
   * Gather relevant datasets from databases, APIs, sensors, or other sources.
   * Ensure data is **representative of the problem.**
3. **Data Exploration & Preprocessing (Data Wrangling)**

**Data Exploration -** This is the **first step** in **data analysis**, where you try to understand the dataset you’re working with. Such as how many **rows** and **columns**

**Data Preprocessing / Wrangling -** This is the process of **cleaning, transforming, and organizing raw data** into a usable format.

1. **Model Selection & Training**

This step includes choosing an appropriate machine learning model based on the problem type and data characteristics. After model selection, the model is trained on the processed data, adjusting internal parameters to learn patterns and relationships

1. **Model Evaluation and Tuning**

After training, the model’s performance is assessed using predefined metrics like accuracy, precision, recall, or F1 score on validation/test datasets. If performance is unsatisfactory, **hyperparameter tuning** is done to optimize model performance for better results.

**Model Evaluation Metrics:**

**Regression:**

Root Mean Square Error (RMSE):  
RMSE measures the average magnitude of the errors between predicted and actual values,

Mean Absolute Error (MAE)

R² coefficient of determination

**Classification: Use Examples to Explain**

**Column matrix -** Is a table used to evaluate the performance of the classification model. It compares the predicted labels with the true labels.

**Accuracy** - measures the **overall correctness** of a **classification model.**

**Precision** - measures the quality of **positive predictions**. Out of all instances predicted positive, how many are actually positive?

**Recall** - measures the ability of the model to **find all positive cases.**

**F1-score** - It balances the **trade-off between precision and recall**, giving a single metric that considers both. It is especially useful when dealing with imbalanced classes.

**ROC-AUC(**Area Under the Curve**)** - measures the ability of the model to **discriminate between positive** and **negative classes** across different classification thresholds.

**ROC (Receiver Operating Characteristic) Curve:**

* A graph that shows how well a classification model separates positive and negative classes.
* It plots True Positive Rate (Recall) against False Positive Rate at different thresholds.

**AUC (Area Under the Curve):**

* A single number that summarizes the ROC curve.
* It represents the model’s ability to distinguish between classes across all possible classification thresholds.

**Hyperparameter Tuning**

* + Optimize model performance using techniques like **Grid Search**, **Random Search**, or **Bayesian Optimization**.

1. **Model Deployment**

Getting the model into production

1. **Monitoring & Maintenance**

Continuous monitoring is necessary to track performance, ensure it remains effective over time, and detect model drift or degradation. The model may require retraining or updating as new data becomes available**.**

**Explain the concept of Overfitting and Underfitting**

**Underfitting**

* Underfitting occurs when a model is too simple (the model is unable to learn the complex or non-linear relationships in the data) to capture the underlying pattern or relationship in the data.
* It performs **poorly** on the **training data** and on **new data**.

**Overfitting**

* **Overfitting** occurs when a model is too complex (excessive capacity or flexibility to learn not only the true underlying patterns in the training data but also the noise, outliers). It learns not only the underlying pattern but also the noise and random fluctuations in the training data. Does not generalize well
* It performs exceptionally well on the **training data** but very poorly on new, unseen data. It has essentially **memorized the training set** instead of learning to generalize.

**Why Standardize data**

* Some machine learning algorithms are sensitive to scaling (if measures are on different scale), where features with larger values dominate the model. Such as **Logistic regression**, **Neural networks**
* **Dominate the model:**

When one feature has much larger numerical values or scale than others, the algorithm gives it more weight or influence, even if it’s not the most important feature.

**MACHINE LEARNING FRAMEWORKS, LIBRARIES, ALGORITHMS**

**Frameworks:**

**Tensorflow -** Open-source framework by Google for ML and DL. Supports large-scale training. Used in image recognition, etc

**Keras** - High-level API for TensorFlow; easy model building and prototyping. Used for Rapid prototyping of neural networks

**PyTorch -**

|  |
| --- |
|  |

|  |
| --- |
| Open-source DL framework by Facebook. Known for dynamic computation graphs and flexibility. Used in NLP, computer vision, etc  **Scikit-learn** **-** General-purpose ML library; supports classification, regression, clustering, etc.  **Libraries:**  **Numpy** - Efficient numerical computations and handling multi-dimensional arrays (matrices, vectors).  **Pandas** - Data manipulation and analysis. Handles structured/tabular data easily. E.g reading/writing data, data cleaning, merging, joining data  **Matplotlib** - Data visualization – creating static, publication-quality plots.  **Seaborn** - Statistical data visualization built on top of Matplotlib. Simplifies complex plots. |

**GIVE AND EXPLAIN TYPES OF MACHINE LEARNING**

**Supervised Learning**

The model is trained on labeled data, meaning the input comes with the correct output

**Types of Supervised Learning**

1. **Regression**
   * **What it does:** Predicts a **continuous numeric value**.
   * **Examples:**
     + Predicting stock prices.

**Algorithms:**

**Linear Regression**: Predict a continuous numerical outcome based on one or more input variables (features).

Linear Regression, Ridge/Lasso Regression, Decision Tree Regression, Random Forest Regression, Gradient Boosting, XGBoost

Project: Predicting House Prices

1. **Classification**
   * **What it does:** Predicts a **discrete class label** (categorical outcome).
   * **Examples:**
     + Email spam detection (spam vs. not spam).
     + Loan approval (approve or reject).

Project: Credit Fraud Detection System

Loan eligibility prediction

**Algorithms:**

**Logistic Regression:** Predict a categorical outcome, usually binary (0/1, yes/no).

Logistic Regression,

K-Nearest Neighbors (KNN),

Decision Trees,

Random Forest,

Gradient Boosting, XGBoost,

SVM, Naive Bayes.

**Unsupervised Learning**

The model is trained on unlabeled data, finding patterns or structures without predefined outcomes.

**Example:** Customer segmentation in marketing. The algorithm groups customers with similar behaviors without knowing in advance which group they belong to.

**Types of Unsupervised learning:**

**Clustering**

Groups similar data points together based on similarity. **Examples:**

* Customer segmentation.

**Common Algorithms:**

* K-Means Clustering
* Hierarchical Clustering

**Project**: **Customer Segmentation** – Mall Customer Data, to analyse mall customer data to segment shoppers into distinct groups based on their demographics, spending habits. Using clustering techniques like K-Means, I was able to identify high-value, loyal customers, which can help the mall target marketing strategies more effectively.

**Dimensionality Reduction**

Reduces the number of features while retaining important information. Often used for visualization or speeding up computation. Common Algorithms:

**Principal Component Analysis (PCA)**

**Reinforcement Learning**

The model learns by interacting with an environment and receiving feedback in the form of rewards or penalties.

**Example:** Training a robot to navigate a maze or teaching an AI to play a game. The system improves its strategy based on trial-and-error feedback.

**Semi-Supervised Learning**

Uses a combination of labeled and unlabeled data. It’s helpful when labeling data is expensive or time-consuming.

**Example:** Image recognition where only a small subset of images is labeled, and the model uses patterns in the unlabeled data to improve learning.

**SQL (Standard Query Language)**

* What is DBMS? Which one do you know?
* MySQL, PostgreSQL and SSMS differences with current date/time as an example

**In short:**

Use LIMIT N – **MySQL, PostgreSQL**, TOP N – **SQL Server**, or FETCH FIRST - **Oracle**→ to get top 3 overall (**Top N**).

Use window functions (**RANK, ROW\_NUMBER**) → to get top 3 within each group (**Top N per Group**).

* What are the different types of SQL commands?
* What is a Database? Types of Databases
* SQL VS NoSQL
* **Composite key**: Is used when one column alone cannot uniquely identity a row
* WHERE VS HAVING
* Types of Datatypes in SQL
* What do you understand about SQL and what is used for? And where have you used it?
* What is Data Modelling? Key Components of Data modelling, Levels of Data modelling (Explain them with examples where necessary)
* What is a Data Warehouse? List types of Schemas with Understanding, ETL vs ELT
* What is ERD? Key Components of ERD, why use ERD also give examples, different **notations** are used to represent entities
* What are Relationships? Give types
* Define Keys in relationships
* What are stored procedures?
* Query Optimization or Indexing

**Query Optimization**

* **Definition:**  
  Query optimization is the process of **making SQL queries run as efficiently as possible** by reducing execution time and resource usage using **Indexing**

**Indexing**

**Is a technique used to speed up the retrieval of data from a database table.**

* **How it works:**  
  Instead of scanning the entire table row by row (**full table scan**), the database uses the index to **jump directly** to the location of the required data.
* **Types of Indexes:**
  + **Clustered Index:** Physically sorts the table data by the index key (only one per table).
  + **Non-Clustered Index:** Creates a separate lookup structure pointing to the actual data rows.
  + **Composite Index:** An index on multiple columns.

**Examples:**

**-- Creating an index on CustomerID**

CREATE INDEX idx\_customerid ON Orders(CustomerID);

**-- Now queries filtering on CustomerID run much faster**

SELECT \* FROM Orders WHERE CustomerID = 12345;

* Types of Joins (Use Table A and B as typical example and syntax) and used for what
* Union and Union all
* Projects you did that involved SQL

**SQL Data Manipulation**

* Create Table
* Insert
* Alter table
* Update
* Types of Joins and Explain with examples

**ADVANCED SQL**

What are window functions?

* Window functions perform calculations **across a set of rows related to the current row** without collapsing the result into a single row (unlike aggregate functions).
* They are often used for **ranking, running totals, moving averages, and comparisons within groups**.
* They are called “window” functions because they operate **over a defined window of rows**.
* **Aggregate**, **Rank**, and **Value (Analytics) functions** in Window
* **Top/Bottom N Analysis**
* **Distribution Analysis**
* Group by vs Partition by
* Over clause and Partition by
* **What are Window frames?** Types
* Case Statement in SQL, and why use them
* Temporary Table
* View vs Table and use cases
* Views vs CTAs and use cases
* SQL CTE vs Views and use cases
* Why CTEs
* What is Stored Procedures? And use cases
* **How do you find duplicate records in a table?** Using GROUP BY and HAVING COUNT (\*) > 1

**Examples:**

**Find duplicate emails in a users table**

**Customers with multiple orders**

**Products with multiple categories**

* What are Null functions? Types

**Practical Questions:**

**Basic SQL Interview Questions**

* What is SQL and how is it different from NoSQL?
* What are the different types of SQL statements? (DDL, DML, DCL, TCL)
* What is the difference between DELETE, TRUNCATE, and DROP?
* Explain PRIMARY KEY, FOREIGN KEY, COMPOSITE KEY, UNIQUE, and CHECK constraints.
* What are indexes in SQL? Why are they used?
* Difference between WHERE and HAVING clauses.
* What are INNER JOIN, LEFT JOIN, RIGHT JOIN, and FULL OUTER JOIN?
* What is a SELF JOIN and a CROSS JOIN?
* Difference between UNION and UNION ALL.
* What is the difference between CHAR, VARCHAR, and TEXT data types?

**Intermediate SQL Interview Questions**

* Explain GROUP BY and ORDER BY. Can they be used together?
* What are **window functions**? (Examples: ROW\_NUMBER(), RANK(), DENSE\_RANK(), NTILE())
* What is the difference between EXISTS, ANY, and ALL?
* What is a **subquery**? Difference between correlated and non-correlated subqueries.
* What are **views** in SQL? Advantages and disadvantages.
* Explain **stored procedures** and **functions**.
* What are **transactions** in SQL? What do COMMIT, ROLLBACK, and SAVEPOINT do?
* What is **normalization**? Explain 1NF, 2NF, 3NF, and BCNF.
* What is **denormalization**? When would you use it?
* Explain CASE WHEN statements with an example.

**🔹 Advanced SQL Interview Questions**

* What are **CTEs (Common Table Expressions)** and when to use them?
* Explain the difference between RANK() and DENSE\_RANK().
* What is **query optimization**? How can you improve SQL query performance?
* Explain **ACID properties** in SQL databases.
* What are **triggers** in SQL? Give an example.
* How do you handle **many-to-many relationships** in SQL?
* Difference between **OLTP** and **OLAP** systems.
* Explain the difference between **clustered** and **non-clustered** indexes.

**🔹 Scenario-Based SQL Questions**

* You need the **second highest salary** from an Employees table — how would you query it?
* A table has **duplicate rows**, how would you remove them while keeping one instance?
* Write a query to get the **top 5 customers** by revenue.
* How would you get employees who **do not belong to any department**?
* How do you find customers who **purchased in 2024 but not in 2025**?
* How would you **pivot data** in SQL (rows to columns)?
* You have a **large table (100M+ rows)** that queries very slowly — how would you optimize it?
* Write a query to return **running totals** of sales per month.
* How would you detect **missing values/dates** in a time series stored in SQL?
* Suppose you need to join **three tables with conditions** — how would you design the query for efficiency

**Top/Bottom N analysis with window functions is a very common pattern in SQL when you want, for example:**

* **“**Top 3 products per category by sales”
* “Bottom 5 students per subject by grade”
* “Top 10 recent transactions per customer”

**This is done using ranking window functions like:**

* **ROW\_NUMBER() –** unique sequential numbering
* **RANK()** – ranking with gaps for ties
* **DENSE\_RANK()** – ranking without gaps for ties

**Example: Top 3 products per category by sales**

**SELECT \***

**FROM (**

**SELECT**

**category,**

**product,**

**sales,**

**RANK() OVER (**

**PARTITION BY category**

**ORDER BY sales DESC**

**) AS rnk**

**FROM products**

**) t**

**WHERE rnk <= 3;**

* **PARTITION BY category → restart ranking per category.**
* **ORDER BY sales DESC → highest sales first.**
* **rnk <= 3 → top 3 products in each category.**

**Example: Bottom 2 students per subject by grade**

**SELECT \***

**FROM (**

**SELECT**

**subject,**

**student,**

**grade,**

**ROW\_NUMBER() OVER (**

**PARTITION BY subject**

**ORDER BY grade ASC**

**) AS rn**

**FROM results**

**) t**

**WHERE rn <= 2;**

* **Using ASC for bottom analysis.**
* **ROW\_NUMBER() ensures exactly 2 students per subject (even if tied).**

**🔹 Top N with Ties**

**If you want to include ties (e.g., two students tied for 1st place), use RANK() or DENSE\_RANK() instead of ROW\_NUMBER().**

**SELECT \***

**FROM (**

**SELECT**

**subject,**

**student,**

**grade,**

**RANK() OVER (**

**PARTITION BY subject**

**ORDER BY grade DESC**

**) AS rnk**

**FROM results**

**) t**

**WHERE rnk <= 3;**

**Top/Bottom N by Date (e.g., latest 5 orders per customer)**

**SELECT \***

**FROM (**

**SELECT**

**customer\_id,**

**order\_id,**

**order\_date,**

**ROW\_NUMBER() OVER (**

**PARTITION BY customer\_id**

**ORDER BY order\_date DESC**

**) AS rn**

**FROM orders**

**) t**

**WHERE rn <= 5;**

**✅ Summary:**

* **Use ROW\_NUMBER() when you need strict Top/Bottom N.**
* **Use RANK() / DENSE\_RANK() when ties should be included.**
* **Always combine with PARTITION BY for per-group analysis.**

**PYTHON**

* What is the difference between a list and a tuple?
* How do you handle missing data in Pandas?
* 27. How do you handle missing values in Pandas?
* 28. What are the different data types in Python?
* 29. Explain the difference between append() and extend() in lists.
* 30. What is the difference between NumPy arrays and Python lists?
* 31. Write a Python script to read a CSV file using Pandas.
* 32. How do you filter rows in a Pandas DataFrame?
* 33. What is the use of the apply() function in Pandas?
* 34. Explain the difference between loc[] and iloc[].
* 35. How do you merge two DataFrames in Pandas?
* 36. What is the difference between Lambda functions and normal functions?
* 37. How do you handle outliers in Python?
* 38. What is the difference between map(), filter(), and reduce()?
* 39. Write a Python function to count the frequency of words in a list.
* 40. Explain the use of the groupby() function in Pandas.
* 41. What are list comprehensions, and why are they used?
* 42. How do you create a pivot table in Pandas?
* 43. What is the difference between deepcopy() and shallow copy()?
* 44. What are Python decorators, and how do they work?
* 45. Explain the difference between correlation and covariance.

**MICROSOFT POWER BI**

**Basic Power BI Interview Questions**

* What is Power BI and how is it different from Excel?
* What are the different views in Power BI? Report View (visuals), Data View (tables), Model View (relationships).
* What is a DAX function? A formula language in Power BI used for creating measures and calculated columns.
* What are the main components of Power BI? (Power BI Desktop, Service, Mobile, Gateway, Report Server, etc.)
* Explain the difference between Power BI Desktop and Power BI Service.
* What are Power Query, Power Pivot, and Power View?
* What are the different data sources supported in Power BI?
* What is DAX in Power BI? Give examples of simple functions.
* Explain the difference between measures and calculated columns.
* What is a Dashboard in Power BI? How is it different from a report?
* What are Filters, Slicers, and Drill-through in Power BI?
* What are relationships in Power BI? Types of relationships supported?
* How do you optimize a Power BI report? Reduce visuals, use aggregated data, enable query reduction, avoid excessive calculated columns.

**Intermediate Power BI Interview Questions**

* What are Power BI dataflows, and why are they useful?
* What is Row-Level Security (RLS) in Power BI, and how do you implement it?
* Explain the difference between DirectQuery and Import mode.
* What is Composite mode in Power BI?
* How do you optimize a Power BI report for performance?
* What are bookmarks and how are they used?
* What is the difference between ALL, ALLSELECTED, and REMOVEFILTERS in DAX?
* What are hierarchies in Power BI, and how do you create them?
* How can you handle many-to-many relationships in Power BI?
* What is a KPI in Power BI and how do you create one?

**Advanced Power BI Interview Questions**

* Explain the difference between calculated tables and measures.
* How do you schedule data refresh in Power BI Service?
* What is the difference between Power BI Pro and Power BI Premium?
* How do gateways work in Power BI? (On-premises data gateway vs personal gateway)
* What are Aggregations in Power BI?
* How do you implement Incremental Refresh in Power BI?
* Explain the difference between star schema and snowflake schema in data modeling.
* How do you use DAX to calculate running totals or cumulative values?
* What are some best practices for Power BI data modeling?
* How do you share reports in Power BI and what are the licensing considerations?

**DATA ENGINEERING**

* ETL vs ELT

**Which one is better?**

**When ETL is better**

* You use a **traditional on-premises data warehouse** (like **Teradata**, **Oracle**, **SQL Server**).
* Data must be strictly validated and cleaned before storage (e.g., banking, healthcare compliance).
* You’re working with smaller, mostly structured data.
* The warehouse has limited compute power, so transformations outside are more efficient.

**When ELT is better**

* You use a **modern cloud warehouse** (**Snowflake**, **BigQuery**, **Redshift**, **Databricks**).
* You handle large-scale or unstructured/semi-structured data (JSON, IoT, logs, social media).
* You want faster pipelines with scalable compute for transformations.
* You need real-time or near real-time processing.
* You want to store raw data for reprocessing later (data lakehouse strategy).

**ETL:** Clean before load **→** great for traditional warehouses

**ELT:** Load raw, then clean **→** great for cloud, big data, and scalable analytics

* Batch vs Streaming Data pipelines

**Types of Indexes**

1. **Clustered Index**
   * Determines the physical order of data in a table.
   * Each table can have only one clustered index.
   * Usually, the primary key creates a clustered index automatically. The data in the table is physically stored in order of EmployeeID.
   * **Example:** If you cluster on EmployeeID, rows in the table are physically sorted by EmployeeID.

**-- Explicit clustered index (optional if primary key already exists)**

CREATE CLUSTERED INDEX idx\_employeeid

ON Employees(EmployeeID);

**Query that benefits from clustered index:**

SELECT \* FROM Employees

WHERE EmployeeID = 101;

1. **Non-Clustered Index**
   * A separate structure that references the table rows.
   * A table can have many non-clustered indexes.
   * Does not change the physical order of the table.
   * **Example:** Index on LastName in an employee table to quickly find employees by last name.

CREATE INDEX idx\_lastname

ON Employees(LastName);

**Query that benefits from non-clustered index:**

SELECT \* FROM Employees

WHERE LastName = 'Smith';

* Instead of scanning all rows, the database uses the index to jump to the matching LastName.

1. **Unique Index**
   * Ensures that indexed columns do not have duplicate values.
   * Often used on columns that must remain unique but are not primary keys.

* Ensures a column has **no duplicate values**.

CREATE UNIQUE INDEX idx\_email

ON Employees(Email);

* Trying to insert the same email twice will fail:

INSERT INTO Employees (EmployeeID, FirstName, LastName, Email)

VALUES (201, 'John', 'Doe', 'john.doe@example.com');

**-- Another insert with same email will throw an error**

INSERT INTO Employees (EmployeeID, FirstName, LastName, Email)

VALUES (202, 'Jane', 'Smith', 'john.doe@example.com'); -- Error

1. **Composite Index**
   * Index on multiple columns.
   * Useful when queries filter by more than one column.
   * **Example:** (DepartmentID, HireDate).

CREATE INDEX idx\_dept\_hire

ON Employees (DepartmentID, HireDate);

**Query that benefits from composite index:**

SELECT \* FROM Employees

WHERE DepartmentID = 5

AND HireDate >= '2023-01-01';

* The database can efficiently filter by both department and hire date using this index.

1. **Full-Text Index**
   * Special index type used for searching text within strings.
   * Useful for keyword searches.

**Useful for searching in long text columns like descriptions or notes.**

**-- Suppose we have a table with employee notes**

CREATE TABLE EmployeeNotes (

NoteID INT PRIMARY KEY,

EmployeeID INT,

NoteText TEXT

);

**-- Create full-text index**

CREATE FULLTEXT INDEX ON EmployeeNotes(NoteText);

**Query using full-text search:**

SELECT \* FROM EmployeeNotes

WHERE CONTAINS(NoteText, 'promotion');

* Finds all notes containing the word “promotion” quickly without scanning the whole table.

1. **Filtered Index**

**A filtered index is a non-clustered index that only includes rows meeting a specific condition (a WHERE clause).**

* This makes it smaller and faster than a full index.
* Great for sparse columns or queries that only need a subset of the data.

**Syntax:**

CREATE NONCLUSTERED INDEX index\_name

ON table\_name (column\_name)

WHERE <filter\_condition>;

**Suppose most employees are active, but some are inactive. You often query only inactive employees.**

**Example**

CREATE NONCLUSTERED INDEX idx\_inactive\_employees

ON Employees(LastName)

WHERE IsActive = 0;

**Query that benefits:**

SELECT \* FROM Employees

WHERE IsActive = 0

AND LastName = 'Smith';

* The index only contains rows where IsActive = 0, so the database scans fewer rows and returns results faster.

**JOINS**

**INNER JOIN**

👉 Returns only the rows with **matching values** in both tables.

📌 **Example use case:**

* **Employees and Departments**: You only want employees that actually belong to a department.

**LEFT JOIN (Left Outer Join)**

👉 Returns **all rows from the left table**, plus matched rows from the right. If no match → NULLs.

📌 **Example use case:**

* **All Customers & Their Orders**: You want to see all customers, including those who **haven’t made an order yet**.

**RIGHT JOIN (Right Outer Join)**

👉 Returns **all rows from the right table**, plus matched rows from the left. If no match → NULLs.

📌 **Example use case:**

* **All Products & Orders**: You want to see all products, including those that **haven’t been ordered yet**.

**FULL OUTER JOIN**

👉 Returns **all rows from both tables**, matched or unmatched. If no match, fill with NULLs.

📌 **Example use case:**

* **Customers vs Newsletter Subscribers**: You want to see people who are **customers only, subscribers only, or both**.

**SELF JOIN**

👉 A table joined with itself.

📌 **Example use case:**

* **Employees and Managers**: Find each employee’s manager from the same employees table.

**CROSS JOIN**

👉 Returns the **Cartesian product** = every row of one table with every row of the other.

📌 **Example use case:**

* **Menu & Drinks**: You want to create **all possible meal + drink combinations** for pricing or promotions.

**What is Data Modelling?**

**Data Modelling** is the process of **defining, structuring, and organizing data** for use in databases or information systems.

It provides a **blueprint** that shows:

* What data is needed,
* How it should be stored,
* How different data elements relate to each other,
* How it will be used and accessed.