**­Interview questions**

1. About yourself
2. About your current work
3. Technical implementation
4. SQL questions
5. Python questions
6. Any question to interviewer

# Tell me about yourself

First of all, I am thankful that you choose me for this interview.

Hi, I am Rohan Srivastwa. I have a total industry experience of 5.3 years. Currently, I am working as a Data Engineer with Tredence Analytics. I have worked in CPG and retail domain and I have a good understanding of these domain.

I have worked on

I am skilled in PySpark, ADF, PowerBI, SQL.

I have implemented CI/CD pipelines using Azure DevOps for scheduling the ADF pipelines.

I have a good understanding of DQ

Apart from this, I am an avid reader, write blogs and I enjoy cycling.

Lastly, I am a person who value sincerity & integrity and follow an organized approach towards the problem.

That is me.

I am looking forward to work for your company.

# Tredence Project



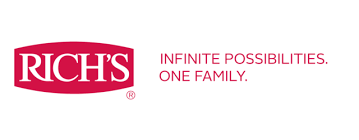
1. AB-InBev

* Worked as data engineer for refactoring tables from legacy system to modern system based on Medallion architecture using Azure Databricks, Databricks workflows for orchestration.
* Created feature branches to develop code from GitHub repos, committed the changes and merged the PR using GitHub.
* Created silver and gold procedures using Databricks for assigned tables using PySpark, Spark SQL based on standard and user defined libraries.
* Validated completeness and equivalency check for each table migrated to Unity Catalog of databricks.
* Migrated tables to snowflake from Unity Catalog as an alternate Data Warehouse.

1. Kelly Service



* Worked on Data Ingestion Framework to ingest data from multiple sources (SQL Server, Oracle, SFTP, SAP) to ADLS Gen2 using code driven framework build on databricks.
* Created delta table for each table ingested to ADLS Gen2
* Worked on creating Data Quality Framework to ensure data ingested have no discrepancy.
* Implemented logging & monitoring in the code for auditing and performance monitoring.
* Email alert sending system in case of any code failure and also a final data quality report.
* Orchestrated the pipeline using Databricks workflows.
* Used Azure DevOps for productionizing code by creating pull request.
* **Tools used** – Azure Databricks, Azure Log analytics – App insights, Azure DevOps.

1. Rich’s Food

* Worked as a data engineer to build pipeline to ingest data from API sources – Facebook Graph API to ADLS gen2 based on Medallion model – Bronze, Silver & Gold layer.
* Responsible for creating Solution architecture for Data Engineering workflow
* Performed API testing for data using POSTMAN.
* To facilitate Data Science team with Solution Architecture and ad-hoc requirements related to data engineering. Ex- Databricks setup, Cluster configuration etc.
* Spearheaded Client calls for requirements and updates with Weekly Presentations.
* **Tools used** – Azure Databricks, Azure Synapse analytics, Azure data factory.

Requirement

1. Good in Azure cloud
2. SQL coding
3. Python coding

**Question to the interviewer?**

Kind of work you do

Any recent challenge you encountered and how did you handle it?

**Why do you want to leave your current organization**

It’s not that I want to leave it but currently the company is going through organizational changes in management, so most people who were hired are looking for project. The same case is with me. I want to start working early.

I find fractal as a fast-evolving company.

**How do you rate yourself in SQL, Python?**

**Tell me about your family and graduation details.**

I have my father who is working in a central government job and a younger brother who is working in a cloud computing firm as a cloud engineering.

**What is the compensation you are looking for ?**

Since I have been loyal to TCS from last 3 and half years, It will be a big decision to change to a new firm. What I am looking for a mix of learning and a decent compensation as per the market standards.

**Why do you want to leave your current organization?**

Growth prospect and I can’t deny I am looking for a decent compensation at this point of my career.

**Where do you find yourself in next 5 years?**

I am more of a technical role loving person; I wish to be remained more of a technical geek than manager. I would love to work on cutting edge technology in the field of data science, machine learning and AI.

Learning has always been

I want to keep myself updated with the latest technology in the IT industry.

**What are your strengths?**

SWOT analysis Growth Mindset

**About the company applying to**

**What current role are you doing ?**

**Situation faced/ challenges in your current company that you solved.**

**Personal questions -**

**Interests/hobbies** in security

# TCS Project Details

Campaign management - to target the customers who can be converted into potential buyers

M&S - to get the process in-house (within the same business) - opposite of outsourcing

Email marketing, CRM, Retention & Acquisition

**Marketing Budget**

Role as Data Engineers - Mark & Spenser

Customers based on master eligibility criteria with premium customers (sparks, dotcom).

1. Merged the bank data of customers based on eligibility criteria to centralize the customers repository data.
2. Worked on migration to bring the process in-house for Mark and Spenser from Planning Inc. (Third party)

ACQUISITION AND RETENTION

Growth of a company depends on A&R.

First a company acquires a customer and then retention happens.

Acquisition

* Increases the customer’s base
* Metric for success of Acquisition campaign is more instant. Ex- Increase in the number of customers as a metric
* Ex- More visitors on web-page, more daily active users

Ways to acquire customers : -

There are various marketing tactics, digital channels which involves cross-channel marketing

1. Push
2. Email Marketing
3. Social Media
4. Content Marketing
5. Affiliate Marketing

Retention

* Customers LTV(Long term value) - it is the project value of that customer over the entire course of his relationship with the company
* Frequent buyers are more loyal than one-time buyers

How does a company should spend on retention and acquisition?

* Acquisition is more expensive -5 times costlier than retention
* Retention has better ROI
* It would depend on the type of company if a company wants to sell one-time buy products, then acquisition should be their priority
* SaaS companies focus on retention side -> a subscription-based fee for source of income. It has been estimated that 65% of revenue of a company comes from its existing customers

Keywords to sort list the resume, models you have worked on,

Data Engineer - Developer role - business logic into codes , azure databricks

Data analyst -

Data portfolio -

Design

1. Brief introduction about you.
2. Python related question
3. SQL related question
4. Big data/ miscellaneous questions

Apache Kafka, TIPCO, WCS, Pluto, OMP

Database knowledge

SQL, NoSQL, Azure

Relational paradigm

Spark - Streaming

OOPS - python

CI/CD

SSO

Docker, Kubernetes

Data structures

Problem solving skills

**Cluster Analysis -**

A statistical method to group similar objects into respective categories. It is also segmentation analysis, taxonomy analysis, or clustering.

Objective -

to group data or data points into groups so that degree of association between 2 objects is high if they belong to same group.

Low if they belong to different groups.

It is unsupervised learning technique.

Propensity model - used to identify those most likely to respond to an offer or to focus retention activity on those most likely to churn (customer most likely to stop doing business).

# Personal Projects

### Blood Donation Prediction

A Taiwan company named blood transfusion service center collects blood in every 3 months for its blood banks in the city.

**To predict** - if a blood donor will donate in a given time window

**Data -**  The data is collected based on RFM model extension - RFMTC

**RFM model** - Recency, Frequency & monetary value is a **marketing analysis** tool used to identify a company’s best customers by measuring & analyzing spending habits.

This analysis helps in predicting when a customer is going to buy and how to convert an occasional buyer into a habitual buyer.

Based on Pareto principle - “80% of outcomes(profit)comes from 20% of effort(potential spending customers).”

Generally, the customers are marked based upon the rank given to them in a scale from 1 to 5.

This way we can better filter out potential buyers and least buyers and how to target them based on the rank provided to them.

**Recency** - The more recent a customer has made purchase the more chances they will continue business.

**Frequency** - The cycle of purchase depends on frequency and type, size and replenishment of product. Ex- groceries purchase can be predicted based on the consumption of product.

**Monetary Value -**Natural inclination of a business should focus on increasing the monetary value of a customer’s purchase.

RFMTC is a variation of the RFM model. Below is a description of what each column means in our dataset:

* R (Recency - months since the last donation
* F (Frequency - total number of donation)
* M (Monetary - total blood donated in ml)
* T (Time - months since the first donation)
* C (Churn rate) a binary variable representing whether he/she donated blood in March 2007 (1 stands for donating blood; 0 stands for not donating blood)

Steps involved in model implementation :-

We check a precise **summary of our dataset** using dataframe.describe() method.

Checking our **target variable** -> we want to check if a person who has donated earlier will donate again or not if the donation camp is setup again. This model is a case of binary classifier

0 - the donor will not give blood, 1-the donor will give blood

**Target incidence check** - to check how balanced or imbalanced is the data for target variable. 76% of target variable has 0 - means no donation happened while 24% of target variable has 1 -means donation happened.

Split the data into training set and test set for evaluation in the same fashion such that each split sample has same 76% of 0 and 24% of 1 using the stratify parameter = target variable. The split has X\_train, X\_test , y\_train and y\_test samples.

Selecting the model -> TPOT is an AutoML library which optimizes the data fetched to it and based on CV score it proposes the best possible model for classifier

AUC score - One of the performance measure/evaluations for classification model, it is basically a measure to understand how well the classification model has performed while predicting 0s and 1s.

Higher the value of AUC curve, the better the model has performed. Here, in our case to predict more accurately if a person is going to donate or not.

A confusion matrix can also help in understanding AOC - ROC value. True positive and True negative are not errors in predicting since they are the correct predictions of a model.

***True prediction*** - means the prediction is true.

**True positive -** prediction value is positive (1) and actual value is also positive (1)

**True negative -** prediction value is negative (0) and actual value is also negative (0)

***False prediction*** - means the prediction is wrong

**Type 1 error - False positive** - your prediction tells true(positive) but in reality, that is false(False). Ex- model predicting a male is pregnant but in reality, this cannot happen.

**Type 2 error - False negative** - your prediction tells false(negative) but in reality, it is true (positive) .Ex - model predicting that a woman is not pregnant but in actual she is.

Type 2 error is very critical as its value can create a lot of problems if not justified properly. Ex - In case of a disease prediction if False negative happens then someone will be predicted as not diseased but in reality, that person is having that disease.

Actual value

|  |  |  |
| --- | --- | --- |
|  | Positive (1) | Negative (0) |
| Positive (1) | **TP** | FP |
| Negative (0) | FN | TN |

Predicted value

Recall = TP/(TP+FN) -> should be always high

Precision = TP/(TP+FP) -> should be always high

Accuracy = (TP+TN)/Total Prediction -> should be always high

F-score = 2\*Recall\*Precision/(Recall+Precision) -> should be always high

AOC - ROC is a curve of plot between **Total Positive Rate vs. True Negative Rate** based on various threshold values of classification.

Models available for Classification ->

### Heart Disease Prediction

303 rows, 13 columns

No null values

Features

1. Age - continuous
2. Sex - categorical
3. Blood pressure (BP) - at the time of hospitalization in mm Hg - continuous
4. Cholesterol - in mg/dl - continuous
5. fbs -Fasting blood sugar (> or < 120 mg/dl) - categorical
6. Resting ECG
7. Max\_heart\_rate - Maximum Heart Rate
8. exang - exercise induced angina
9. Old peak - ST depression induced by exercise relative to rest
10. Slope - slope of the peak exercise ST segment
11. cv - no of major vessels (0-3) colored by fluoroscopy
12. thal - type of defect in 3 ranges 0-3 - normal, 3-6 - fixed, 7 - reversable defect
13. Target - 0,1

# Business questions

1. You have banks credit and debit card transaction data. How do you approach a customers (filter 100 out of 1000) for loan offers?
2. As a bank how do you target to establish the bank’s ATM in a particular location?

**CASE STUDY - BUSINESS ROUND**

Factors to choose ATM on a location of a bank

How will you decide whether to provide a loan to person if you have the transaction details.

1. **How many cups of tea were consumed in Mumbai last month?**

First, clarify the question. Then, start solving.

As a first step, inform the interviewers that each day of the week is being considered equally. Tea consumption might likely decrease during the weekend as people do not go to the office—so you might consider that as well. We shall go with the first assumption. The population of Mumbai is 18 crore; we shall round it up to 2 crores. 20% of this population is assumed to be children who do not drink tea. Another assumption is that of the remaining population, 20% are habitual drinkers, 30% are regular drinkers, 20% are occasional drinkers, and 10% are non-drinkers. The habitual drinkers may be said to have three cups of tea in a day. Regular drinkers may be said to have one cup of tea in a day. The tea consumption of occasion drinkers maybe once a week, and that of non-drinkers none at all.

Calculating proportions-

Habitual – 3 x 0.2 x 7 = 4.2 Regular – 1 x 0.3 x 7 = 2.1

Occasional – 1 x 0.2 x 1 = 0.2 Non – 0

Total = 6.5

Total cups per week = 6.5 x 1.6 crore = 10.4 crore

**2. How many iPhones are currently being used in India?**

Clarify with the interviewers whether the question is about only a single version of the iPhone or all versions put together. Here, we shall assume that all iPhones put together are being talked about.

The first step toward solving this query will be segmentation. There are many ways in which India’s population can be segmented. Here, we shall first assume that only people who have attained a working age and are under the age of retirement own an iPhone. Children and old citizens do not own an iPhone. This removes 20% of the population as children and 20% as senior citizens.

The next assumption will be that only the upper stratum of India’s income range can afford an iPhone. This metric assumes that only 5% of the eligible citizens from the previous filter can own an iPhone.

Now, it is not necessary that every member of this upper stratum will own an iPhone.

Other options, such as OnePlus, Samsung, etc., are also available. However, a fair assumption would be that 50% of the eligible population from the previous filter owns an iPhone.

Calculating the proportion of the population that owns an iPhone –

0.6 x 0.05 x 0.5 = 0.015

Total iPhones in India = 0.015 x 130 crore = 1.95 crore

**SCENARIO BASED QUESTION**

# Job Application

## Email to recruiters

Hi,

I hope you are doing well.

As per the LinkedIn conversation/post, I have shared my resume.

Please find my attached resume for reference.

## Salary negotiation

1. Experience in TCS - 3.5 year - promoted to IT analyst with **45K/month** in hand ,

1. onsite opportunity in TCS, UK -Working for Mark and Spencer account.
2. A very good HIS policy which cover all my family members upto 15 lakhs , other benefits - mobile, internet & transport allowances
3. Moreover, I get Saturday, Sunday as holiday. (52\*1000) -> **52k**
4. I will also be eligible for my gratuity in next 9-10 months -> approx **80K**
5. Leaving the brand name of TCS ecosystem is also a big decision for my career.
6. On an average market standard, a DE is gets 12-15 lpa.

I have worked in multiple domains in TCS be it Banking & Finances, Retails , Security which has enriched my learning and provided me with better understanding of the market. After all these years of experience I am expecting a good compensation.

I am very much excited to work for xyz organization, with has a steep learning curve. I hope this conveys my view clearly.

I hope to hear back from you.

Thanks

## Other

**Note to connection**

Hi, thank you for accepting my connection request.

I am interested in putting myself forward for any data analytics//ML roles. Feel free to check out my LinkedIn profile to see if I may be a fit for any of your current openings.

## Resume building techniques

Project and its objective

Key points - campaign experience

ATS filters -

What exactly have you done?

OK so like should I talk about the previous project on my current project