**DATA SCIENCE INTERVIEW PREPARATION (30 Days of Interview Preparation)**

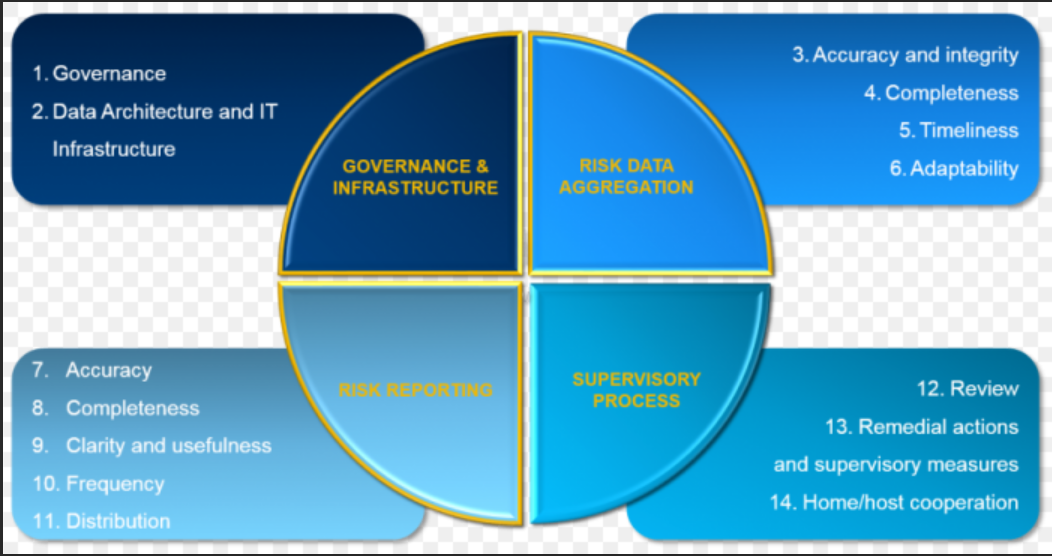
**Detailed Project Report**

**Project Title – Credit Default Risk Prediction (RDARR)**

**Q1. Tell me about your current project**

The current project is called Credit Default Risk Prediction. The global financial crisis revealed that many banks had inadequate practices for timely, complete, and accurate aggregation of risk exposures. These limitations impaired their ability to generate reliable information to manage risks, especially during times of economic stress. Responding to this pervasive systematic issue, the Basel committee on Banking Supervision (BCBS) issued the “Principles for Effective Risk Data Aggregation and Risk Reporting” (RDARR).

There are 14 prinicpals under RDARR –



**Introduction of Machine Learning in Project -**

**Business Requirement** – Default risk is the chance that companies or individuals will be unable to make the required payments on their debt obligations. In other words, credit default risk is the probability that if you lend money, there is a chance that they won’t be able to give the money back on time. Bank wanted to implement the 14 principals of RDARR to become BCBS complaint and also implement a model which can predict the risk for those counterparties which are more susceptible for getting defaulted.

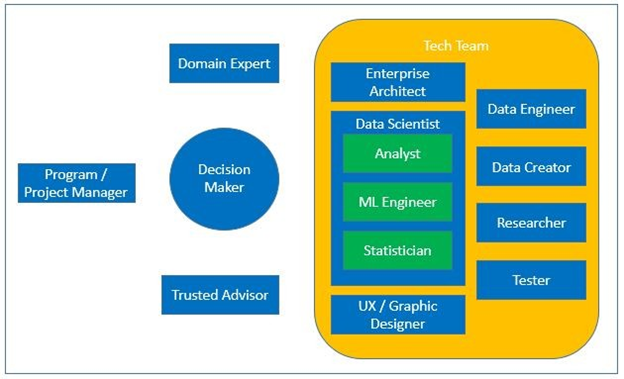
**Q2. What was the size of the data?**

The number of trades used as training data is around 250000.

**Q3. What was the data type?**

The data types used for this model are float, integer, char, blob as the data contains large number of columns.

**Q4. What was the team size and distribution?**



**The Team consists of -**

1. 1 Project Manager
2. 1 Technical Lead
3. 2 QA Engineers and 2 Developers (UI Developers)
4. 1 Enterprise Architect
5. 2 Data Scientists (Analyst+Statistician+ML Engineer)
6. 1 Data Engineer

**Q5. What was the size of the cluster?**

Need help from ineuron on this answer.

**Q6. How many nodes were there in all the Dev, UAT, and Prod environments?**

3 instances for Dev.

3 instances for UAT

3 instances for Prod

Need more help on this answer from ineuron

**Q7. What was the Server and infra used in the project?**

1)Two Database production and two backup secondary server located in Edinburgh and Manchester

2)Two application server and two back up secondary server located in Edinburgh and Manchester.

3)Golden Gate

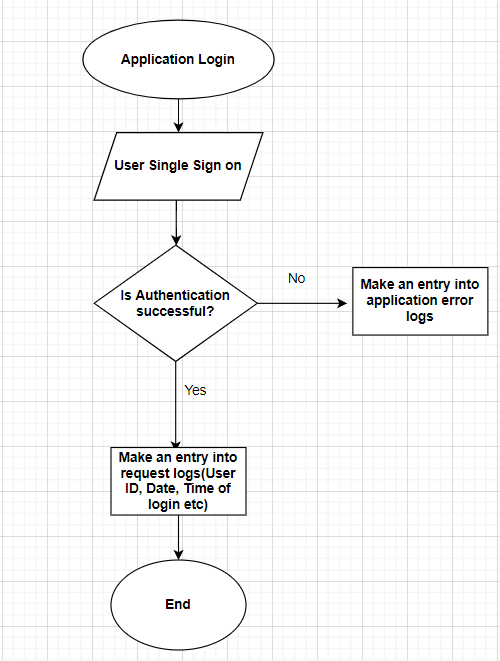
4)Unix Server to host application and database server.

**Q8. How were you creating and maintaining the logs?**

We are creating and maintaining database and application logs.

Application Logs –

Applications logs are stored in OBIEE.



Database Logs –

We have also maintained database audit logs but for any log we need to reach out DBA team by raising a ticker for any logs required–

a) Capturing authenticated user has login to schema

b) Capturing all the DML and DDL events

c) Archiving the logs and deleting it in every 60 days through housekeeping as per the request from business stakeholders.

d) Capturing all the errors in DB error logs.

**Q9. What techniques were you using for data pre-processing for various data science use cases and visualization?**

Steps in Data Pre-processing in Machine Learning

1. Acquire the dataset
2. Import all the crucial libraries
3. Import the dataset
4. Identifying and handling the missing values
5. Encoding the categorical data
6. Removing the outliers
7. Splitting the dataset
8. Feature scaling

**Q10. How were you maintaining the failure cases?**

If the model is not able to predict the correct information then we must gather more varied, non-conventional data to further refine and improve our approaches to accessing risk. For example – We are getting data from source 1, however source 1 is getting data from source 2 after remodeling and refining the data so rather taking data from source 1 we can take it from source 2 directly.

**Q11. What kind of automation have you done for data processing?**

We had a full-fledged ETL pipeline in place for data extraction. We load the data into staging area. The staging area is used to combine data from multiple data sources, transformations, validations and data cleansing.

**Q12.Have you used any scheduler?**

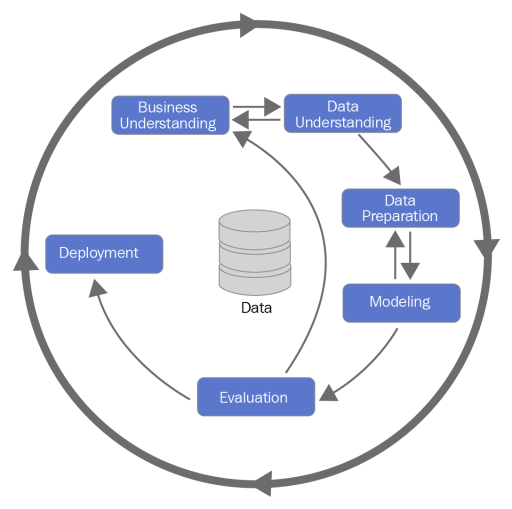
Yes, a scheduler whenever there is a failed scenario the scheduler runs and retrain the model after a fixed time(20 days).

**Q13.How are you monitoring your job?**

We have our autosys scheduler which runs at 4 am UK time for our data batches and checks for any error or failures, incase of any failures alerts were generated and notified via email.

**Q14. What were your roles and responsibilities in the project?**

My responsibilities are analyzing, processing, and interpreting data. Work on SQL optimization and python scripts whenever required. Using advanced statistics to derive insights from it and communicate their findings to business stakeholders. Labeling the model and training the model on the prepared dataset. Deploying the model to cloud and monitoring the model for any issues. Provding support to QA and Prod support team before deployment and then providing the hypercare support post deployment



15.What was your day to day task?

1. Updating Jira tasks assigned to me
2. Attending daily scrum meetings and requirement call with business
3. Working on sql and python scripts
4. Data validation, Data Cleaning and Pre processing
5. Working on the train and test data and retraining the model as and when required.

Q16.In which area you have contributed the most?

1. Data validation, Data Cleaning and Pre processing
2. Working on the train and test data and retraining the model as and when required.
3. Finding out the best algorithm which will suit the model
4. Cloud Deployment

Q17.In which technology you are most comfortable?

Machine Learning, Python, SQL, PL/SQL ,AWS cloud, GCP, AZURE,Heroku, unix, pycharm

Q18. In how many projects you have already worked?

Well I have worked on multiple projects but in case of Machine Learning I have worked in basically two projects eg : trade volume prediction during heightened awareness period, Storage prediction for next 6 months for our databases, Credit Risk Prediction (RDARR).

Q19. How were you doing deployment?

The mechanism of deployment depends on the client's requirement. For this project deployment was done on AWS cloud.

Q20.What kind of challenges have you faced during the project?

Challenges we faced in the terms of a good dataset as we are having different upstream which are sending limits, utilisations, current exposure and potential exposure data for the same counterparties. This raises an issue of discrepancies as we are getting the different data from different sources for the same counterparties. Cleaning the data and analysing the data properly so that data can fit for a model. Labeling the data was the tedious task which took ample amount of time to get it completed.

Q21.What will be your expectations?

I expect to showcase all my skills and learnings into the projects which are assigned to me and also enhance my knowledge by learning new things.

Q22. What is your future objective?

As it is said that data science is the future ahead so I would like to touch base all the aspects of machine learning, AI and deep learning. As the technologies are evolving at a very fast pace so would like to learn new technologies and skills and want to keep myself updated and upskilled.

Q23. Why are you leaving your current organization?

My learning curve has come to a halt as there is not much scope of Machine Learning in my current organization. I am looking for a change to upgrade my knowledge and skills in the field of AI.

Q24. How did you do Data validation?

Data validation is done on many scales. Mandatory fields should not be null or missing. There should not be special characters in the data. Reconciliation data after loading in our application with source. Also keeping a check in each and every new data entry that is going to add to the training data.

Q25.How did you do Data Enrichment?

One of the ways in which we do data enrichment is that if any column value is not available we check our existing data if we can get some information from there.

For example – We have legal entities and they are demonstrated using ciscodes, now in few cases it might happen that ciscodes for a particular legal entity is missing then in that case we check the facility Id as the facility id is a combination of a unique sequence number and ciscode. We can take the ciscode from there and fill our missing value.

Q26. How would you rate yourself in machine learning?

Well I have many experience in different technologies and domain but when it comes to Machine Learning I can rate myself 7.5 out of 10.

Q27. What are the areas of machine learning algorithms that you already have explored?

I have explored various machine learning algorithms like Linear Regression, Logistic Regression, L1 and L2 Regression, Polynomial Regression, Multi Linear Regression,Decision Trees, Random Forests, Extra Trees Classifier, PCA, TSnE, UMAP, XG Boost,CAT Boost, ADA Boost, Gradient Boosting, Light Boost, K-Means,K-Means ++,LDA, QDA, KNN, SVM, SVR,Naïve Bayes, Agglomerative clustering, DBScan, Hierarchical clustering, TFIDF, Word to Vec, Bag of words, Doc to Vec, Kernel Density Estimation are some of them.

Q28. In which part of machine learning have you already worked on?

I have worked on both supervised and unsupervised machine learning approaches and building different models using the as per the user requirement.

Q29.How much time did your model take to get trained?

Need help from ineuron team on this answer.

Q30. At what frequency are you retraining and updating your model?

Model is retrained on every 21 days after the sprint call.

Q31. In which mode have you deployed your model?

I have deployed the model both in cloud environments as well in the on-premise ones based on the client and project requirements.