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

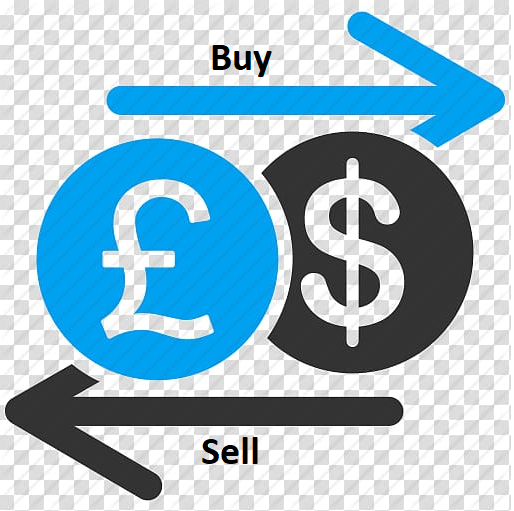
**Detailed Project Report**

**Project Title - Trade volume prediction during heightened awareness period**

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

The current project is called Wallstreet Settlements System (WSS). It is responsible for the settlement of foreign exchange, cash flows and money market trades. WSS has the ability to process transactions from multiple front-end systems simultaneously. WSS has the capability to settle funds both externally via SWIFT and internally via International. It also creates Cash/Nostro accounting for reconciliation and ledger control. It is also a confirmation generation system for all FX & MM transactions via multiple mediums SWIFT/CLS/FAX/Paper/Email.

**Foreign Exchange Market -** The foreign exchange market is a global decentralized or over-the-counter market for the trading of currencies. This market determines foreign exchange rates for every currency. It includes all aspects of buying, selling and exchanging currencies at current or determined prices. In terms of trading volume, it is by far the largest market in the world, followed by the credit market.



**Money Markets –** Money market is the market where the buying and selling, lending and borrowing of short- term instruments/funds occur. It is a highly liquid market with maturities of up to one year. The money market is typically seen as a safe place to put money due to the high liquid nature of the securities and short maturities. Examples of money markets includes loans and deposits, commercial papers, banker’s acceptance etc.



**Introduction of Machine Learning in Project -**

**Business Requirement** – Since the foreign exchange and money markets are volatile and susceptible to ongoing social and political events, there was an ask from the bank to predict trade volumes of FX trades during Heightened Awareness Period. Heightened Awareness Period will be US/UK elections, Brexit, Financial Year end, Covid Pandemic, Counterparties getting default etc. During this period trade volumes are pretty high and business wants a model which can predict trade volumes based on the past volume data so that the operations team can prepare in advance for the number of manual/automated payments which needs to be done on a particular day. Manual payments are done as system takes time to process payments pertaining to high volumes of trading data which can lead to exception or some other glitches. The Model will predict hourly wise expected volumes of trades on the basis of past data.

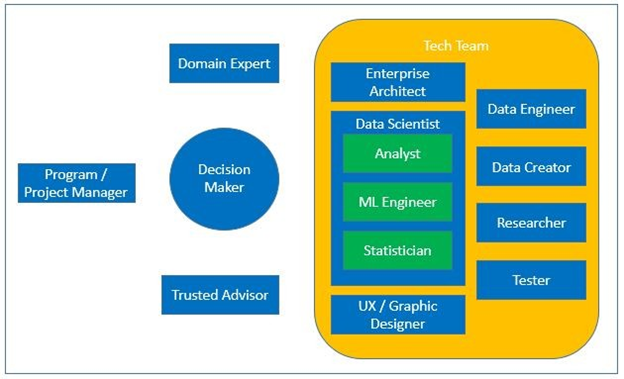
**Q2. What was the size of the data?**

The number of trades used as training data is around 125000.

**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?**

Need help from ineuron on this answer.

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

1)Two Database production and two backup secondary server located in Hetahcote and Beachcroft

2)Two application server and two back up secondary server located in Heathcote and Beachcroft.

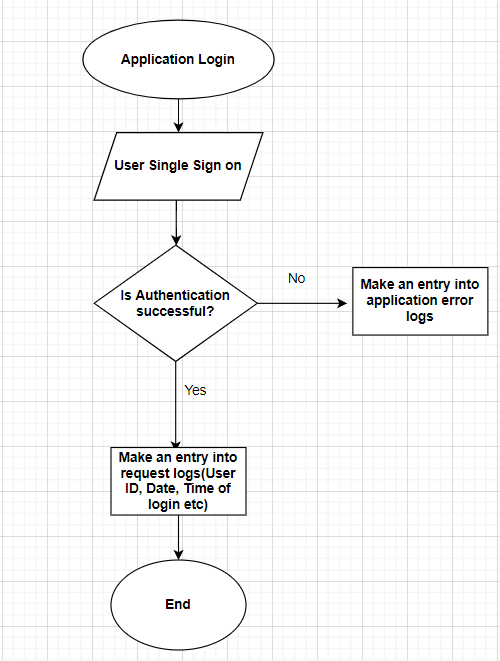
3)Golden Gate

4)MicroServices

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

We are creating and maintaining database and application logs.

Application Logs –



Database Logs –

We have also maintained database audit logs –

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?**

It is a data mining technique that transforms raw data into an understandable format. Raw data(real world data) is always incomplete and that data cannot be sent through a model. That would cause certain errors. That is why we need to pre-process data before sending through a model.

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 of trade volumes, for example – Due to heavy load some sort of trades has not flown into system correctly or expelled out then we have maintained a table in which a trigger will fire which will store all the failed scenarios and the team will inspect all the failure and will retrain the model for better performance.

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

Upstream systems send the data via argon tool which then loads into our intermediate table via ODI. Preprocessing task we do in our intermediate table and once data is finalised data is moved to main table. Trades which are already entered into Wallstreet system can be easily used after doing pre-processing for training the Volume model.

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

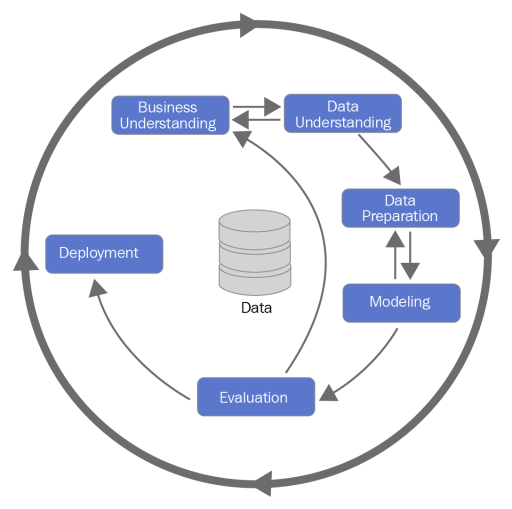
Yes, a scheduler whenever there is a failed scenario the scheduler runs and retrain the model after a fixed time(every day).

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

Logging are in place whether an event is related to application or database. During any error support team gets notified via email to take action or support team also manually checks the logs on daily basis. We have classified error color coding as Red, Amber and Yellow so that support team should know the criticality of the error and how fast action needs to take on that particular issue. If some error is related to business then those have SLA to rectify it with in some particular timeline so that there should be no impact to the business.

**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?

My day to day task is having meeting with ops business if any change modification they want in the model, debugging issues and mentoring prod support team how to debug issues, Working with DBA team also for pulling AWR report for the daily performance, data validation, data labelling and helping provide UAT support also.

Q16.In which area you have contributed the most?

As a database specialist I was involved in the table/trigger creation, Storage planning, AWR analysis, finding out the best algorithm which will suit the model, cloud deployment etc.

Q17.In which technology you are most comfortable?

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

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, Risk.

Q19. How were you doing deployment?

The mechanism of deployment depends on the client's requirement. For example, some clients want their models to be deployed in the cloud, and the real-time calls they take place from one cloud application to another. On the other hand, some clients want an on-premise deployment, and then they do API calls to the model. Generally, we prepare a model file first and then try to expose it through an API for predictions/classifications. The mechanism in which he API gets called depends on the client requirement.

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 different data categories such as FX , cashflow and Money Market trades. 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. Also business used to change their requirement in between so to manage those scenarios was also a challenging task. Failed scenarios was also need to be taken care on the same day as the data is real time and we need to retrain the model everyday.

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 – Our deal number(Trade number) is a combination of ddmmyy+counterparty id. Sometimes during publishing the deal number counterparty id is missing. In such we have one more trusted source from which we confirm the counterparty id of particular trade number and then on the bssis of their confirmation we fill that missing counterparty id in our data.

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 mainly used during Heightened Awareness period so at that model needs to train on daily basis.

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.