# Orderability Mismatch between Organizations for Products

Detailed Project Report (DPR)

| | |
|---|---|
| Created by : | **Jateendra Kumar Pradhan** <br><br> **Machine Learning Batch ( 21$^{st}$ Nov 2021 )** |
| Reviewed by : <br> *\*\*\*TBD* | **Sudhanshu Kumar** <br> **iNeuron Expert Team** |

## Q1 . Tell me about your current project ?

Generally a Product ( known as PID – Product Identifier ) can be ordered from a common Organization center globally . Products are nothing but Router , Switches , Servers , Interfaces , Machinery , etc . Meanwhile there is a concept where these Products can be ordered from their particular regional country locations ( Exp : For India , IND is the Organization code/location ) . As a whole any product can be ordered from Global location or any Country code locations .

I work for the client which is world's number one Supply Chain Industry ( as per Gartner's ranking of World top 10 Supply Chain Companies ) . There are products which shows Order failures when end customer tries to book order , because there is a difference in Order status and few other business attribute values differences between Global location and Country location . So we are building a model which will detect the differences of Order mismatch of a product and tell us if the product can be ordered or not . This is kind of early detection of order failure for the product to rectify early before end customer places any order .

## Q2 . What is the size of the data ?

We generally work with Products which is called as PID ( Product Identification Number ) which refers to a single unit of device . For example Laptop is a device , Desk phone is a device , Speaker is a device . Like wise number of PID/Devices for training purpose used is : 9.7 million .

## Q3 . What was the datatype

Every Device or PID consists of 27 types of attributes . They are combination of Int , Floats , Date , String types .

## Q4 . What was the team size & distribution

Below are the Structure of the team :

| Orderability Setup Team globally | 42 | Members who setup Orderability mechanism for Products globally . |
|---|---|---|
| Product Compliance Team | | |
| Delivery Manager | 1 | Onsite |
| Program Manager | 1 | Overall program cum Domain Expert . |
| Project Leads | 1 | |
| UI Developers | 1 | |
| Data Scientists | 2 | |
| Data Analyst | 2 | |

| | | |
|---|---|---|
| Deployment Engineer | 1 | |

## Q5 . Which databases you were using ?

We use Below DBs :

**Oracle Database 12c** : For Data Extractions from many Databases which uses Oracle DB.

Mongo DB ( DBaas ) : For storing and extracting data after clean-up for ML Algorithms .

## Q6 . What are the version of Databases

Here are the Databases versions : O**racle Database 12c** & **Mongo DB** ( DBaas )

## Q7 . What was the size of Infrastructure setup ?

We use Virtual Machine Infrastructure setup with below configuration  :

| | |
|---|---|
| Processor | Intel i7 |
| CPU | 8 Core Processor |
| Installed Memory ( RAM ) | 48GB |
| Storage Area Network | 256GB |
| System Type | 64-bit OS,x64-based processor |
| Mounted Virtual Machine | Over Client Network |

## Q8 . How many nodes were there in all the Dev , UAT and Prod environments ?

All developments are done in individual systems . Integration was done in one particular system and we use only one Dev environment server to deploy . Parallelly tuning and other enhancements were being done on other Development server . For training the model we use directly the Virtual Machine Server to make model file . The same model is being pushed to Staging environment for testing purposes . Once everything is fine in Staging environment , the model file is pushed to Prod environment .

We had below instances of servers .

| | |
|---|---|
| Development environment | 2 instances |
| Staging environment | 2 instances |
| Production environment | 2 instances |

## Q9 . How were you creating and maintaining the logs ?

We maintain the logs in various files within a folder . Logs will be maintained for 7 days atmost and will be achieved after that to another folder . There is a purging mechanism in place to delete all the archived logs after 90 days of timeframe . This is as per business requirement and as every day outputs/predictions are being sent to emails , so no need of keeping the details in logs for more days .

Below are the log file maintained :

- User logging and Transaction logs

- Data collection from Oracle Databases logs

- Data Validation logs

- Data Insertion to MongoDB logs

- Data Extraction to CVS/Excel file logs

- Model Training logs

- Model Prediction logs

- Memory logs

- Storage performance logs

## Q10 . What techniques were you using for data pre-processing for various data science use cases and visualization ?

We follow various data pre-processing techniques for all the raw data :

$\Rightarrow$ Oracle schema and Oracle DB connections to gather raw data .

$\Rightarrow$ Datatypes and valid values for columns , if not match to be taken to Orderability setup team .

$\Rightarrow$ Identify important and required columns by using Heat-map incase of any columns relations .

$\Rightarrow$ Handling NULL values .

$\Rightarrow$ Missing values handling by KNN Imputer .

$\Rightarrow$ Removing outliers by IQR technique , BOX plots .

$\Rightarrow$ Filtering out unnecessary data as per Business requirements using Pandas and Numpy libraries .

$\Rightarrow$ Scaling of data by Standard Scalar due to high difference in magnitude of data .

$\Rightarrow$ One hot encoding to convert categorical data to numerical data .

$\Rightarrow$ PCA ( Principal Component Analysis ) technique to reduce no. of columns and it's dimensionality

## Q11 . How were you maintaining the failure cases ?

If model is not able to predict or provide proper outcome , we have a exception email mechanism in place which will send email to our internal team . We shall take up those exception PIDs and start analyzing with the help of Orderability setup team and compliance team . Once all the criteria and root causes are found we fulfill all the requirements for that PID to retrain and predict the outcome again .

## Q12 . What kind of automation have you done for data processing ?

End to End automation is being implemented as part of data processing . This involves data extraction from Oracle , Segregation of all raw data to a particular Data Frame format , Keeping required columns , Filling unknown/blank values with statistical analysis , Handling special characters , attribute information ranges , invalid values in columns , datatypes of various columns , etc .

## Q13 . Have you used any scheduler ?

Yes , There is a scheduler we have set twice in a week – Monday and Thursday at 8.30 PM IST . This scheduler will identify all the PIDs which have undergone transaction for any changes in the field and then our model picks up them to predict incase of any Orderability mismatch between Global organization and Country code organization .

## Q14 . How are you monitoring you job ?

We have logging mechanism in place . If the PID doesn't give proper prediction or outcome , an automated email with the exception issue will be sent to our internal team . With this we analyze the issue and resolve with the help of Orderability setup team .

Also we have Memory logging , CPU consumption logging and Space logging mechanism in place . If the limit exceeds certain defined threshold value , it triggers an alert email to our Internal team .

## Q15 . What were your roles and responsibilities in the project ?

I work as a Program Manager cum Technical Architect and Domain Expert for the entire project . Client Interaction and cross functional team interactions were primary jobs . Also getting involved in Data gathering and segregation with Data Analysts , work with Data Scientists and Leads for all the Agile deliverables . Responsible for Cloud Infrastructure setup for Deployment . Co-ordinate with Orderability Setup team during UAT testing of the model . Post Deployment Support and Issue fixing support after Go-Live .

## Q16 . What was your day to day task ?

My daily work involves Data gatherings from Oracle Databases , guiding and reviewing Data Analyst's daily tasks . Also get engage with Data Scientists and Leads for Development guidance and progress on daily basis . Interaction with stakeholders and orderability team for Data Integration part . Involve in Architecture plan and act as Domain Export in each step . Also conduct meetings for agile mode of delivery .

| | |
|---|---|
| Scrum meeting with Data Analysts and Data Scientists | Daily |
| Entire Team | Twice a week |
| Meeting with Delivery Manager | Once a week |
| Client Meeting | Once a week |
| Meeting with Orderability setup team | Twice a week |

## Q17 . In which area you have contribute the most ?

I contributed most in Architectural plan , raw data gathering from Database , Domain Expertise for the whole project . Next level of support was Data validation , segregation and preprocessing . Also work for best ML Algorithm modeling , UAT support for Product owners . Also end to end deliverables & client interaction . Budget planning and works on different POCs as well .

## Q18 . In which technology you are most comfortable ?

I worked mostly in the field of Machine Learning & Natural Language Processing . As my project has been executed from Raw data gathering till Model deployment and model training , I am comfortable most in Machine Learning and Huge data handling .

## Q19 . How you rate your self in Databases and Big Data technology ?

I have my prior IT experience also in Database and PLSQL programming where I worked in acquisition and data migration projects . I have functional knowledge in Big Data as I use to interact with Big Data team in my other projects who pass on mass and huge data to Snowflakes which is being referred by many Oracle Databases  . But in our project most concentration is for Data Scientists and Data Analysts work .

## Q20 . In how many projects you have already worked .

I have worked for almost more than 20 projects in my whole career . Since last 3.5 years I have been working on Machine Learning and Model building projects . Also in parallel working on various POC projects which we showcase to customers and if approved by Customers will lead to full fledge delivery projects . This helps us to expand our team , growth and revenue to our current account . Recently we

have own 2 POC projects and Client agreed to fund them for next 2 years of Budget planning for FY22 and FY23 .

Our project and development mode of delivery is in Agile methodology . So we have milestones defined as part of our project planning and timeline , by which we deliver part wise projects . Firstly , we deploy our code in local server and do continuous internal testing for 7 days . We fix all issues in between and once ready will push them for UAT . Post UAT will deploy them to Virtual Machine server . Model files will be deployed by directly in Production environment .

## Q22 . What kind of challenges you have faced during the project ?

Below are the challenges we find during the execution of the project :

➢ Fetching live data from Oracle DB was really challenging as the Schema and tabulation details were owned by Backend IT team and they have constraints of sharing live data information and direct interaction with DB . So we came to a conclusion of mutual understanding of placing live data into a separate table which will be of 4 days old data . So our raw data was of 4 days old which has least impact as there will be no changes in this short span of time .

➢ During UAT testing , we involved Customer Partner Team as per Orderability Team's request . We faced challenges to give them trainings , KT and socialization about the model plan . There were around 150 people involved for UAT testing from all team .

## Q23 . What will be your expectations ?

My expectation always is to give best support and guidance to our Customers for their Business growth and future prospects . Parallely to take my project account to a new height for our company revenue and forecast , so that we serve best and do best in current AI cutoff technologies .

## Q24 . What is your future objective ?

I have 5 top most objectives in my current role :

➢ Explore every corner of AI technologies and apply them in our future projects .

➢ Learn new things with Top Level Managements and also with all our subordinates .

➢ To push everybody of my team and make them learn towards leadership role and professionalism

➢ Most Customer engagements to deal with them very smoothly with their sentiments towards our company and services .

➢ At last learn , get learning from others and guide others .

## Q25 . Why are you leaving your current Organization ?

I have been with my current company for long years now and almost have worked with many projects , automations, , POCs , Project acquisitions , Deliveries and Program Engagements . The outside market is changing rapidly with many new technologies and want to be as par updated with new cut AI technologies and want to upgrade my skill set and want to explore with new team and Management . I can get the same opportunity in my company also , but want to see the larger world and environments .

## Q26 . How did you do Data validation ?

We have designed an automated Data Validation section to validate all raw data and if any set is not meeting the requirement will send them back to Data Analysts for further analysis . The automation involves all type of Data Mining in the larger dataset like – Special Char Handling , Description length handling , Missing values handling , Business Unit Distributions , processing to fit as per input template format . By this we shall align the valid data to be chosen for trainings .

## Q27 . How did you do Data enrichment ?

We do data appending to the current set of data from newly generated data , this happens every month , so as to make the dataset fully complete .

## Q28 . How would you rate yourself in Machine Learning ?

I still find everyday is a kind of learning phase for me . So I would rate myself as ~8.0 in a scale of 1-10 .

## Q29 . How would you rate yourself in distributed computation ?

I would rate myself as ~7.5 in a scale of 1-10 .

## Q30 . What are the areas of Machine Learning algorithms that you already have explored ?

Here I have listed the Machine Learning algorithms I have explored with :

- ➢ Linear regression
- ➢ Multi-linear regression
- ➢ Ridge & Lasso regression
- ➢ Polynomial regression
- ➢ Logistic regression
- ➢ Decision Tree

- Random Forest
- Ensemble Techniques
    - Bagging ( Bootstrap Aggregation )
    - Boosting – Ada Boost , Gradient Boost , XG Boost
- K Nearest Neighbors
- Naïve Bayes
- SVM , SVR
- Clustering :
    - K Means
    - Hierarchical
    - DBSCAN
- NLP ( Natural Language Processing ) :
    - TFIDF
    - Text Analysis
    - Bag of Words
- PCA
- Multicullinearity

## Q31 . In which part of Machine Learning you have already worked on?

I have an experience of working in both Supervised and Unsupervised Machine Learning algorithms. Based on client requirement we choose and decide the best model to be built in .

## Q32 . How did you optimize your solution ?

For optimizing solution and make the model work better , we spent more time on :
- More data for training of the model and increase the accuracy of the model .
- More insight into data to keep them in record instead of removing them with the help of Orderability team .
- Rigorous hyper-parameter tunings for models to get best accuracy and choose the best model out of it .
- Optimization of time by tracking & recording time for every model we build . We choose the best later on .
- Memory optimization & Space optimization techniques .
- Fixing UAT issues raised by Customer Partner team and Orderability team .

> ➤ Enhancements based on Business requirement and wider user socialization to use the model .

## Q33 . How much time did your model take to get trained ?

With 9.7 millions of records and with hyper parameter turning and all models predictions , it take around 14 hours to train the model .

## Q34 . At what frequency are you retraining and updating your model ?

We retrain our model biweekly once .

## Q35 . In which mode have you deployed your model ?

We have deployed the model both in stand alone system of Deployment engineer and post UAT phase in direct Virtual Machine system .

## Q36 . What is your area of specialization in Machine Learning ?

I am well versed with all the algorithms in Machine Learning . This is because we are working on a number of POCs for which we need to be expert on all area . But mostly we see client requirements on Unsupervised Machine Learning . But I am ready to talk on any area of interest .