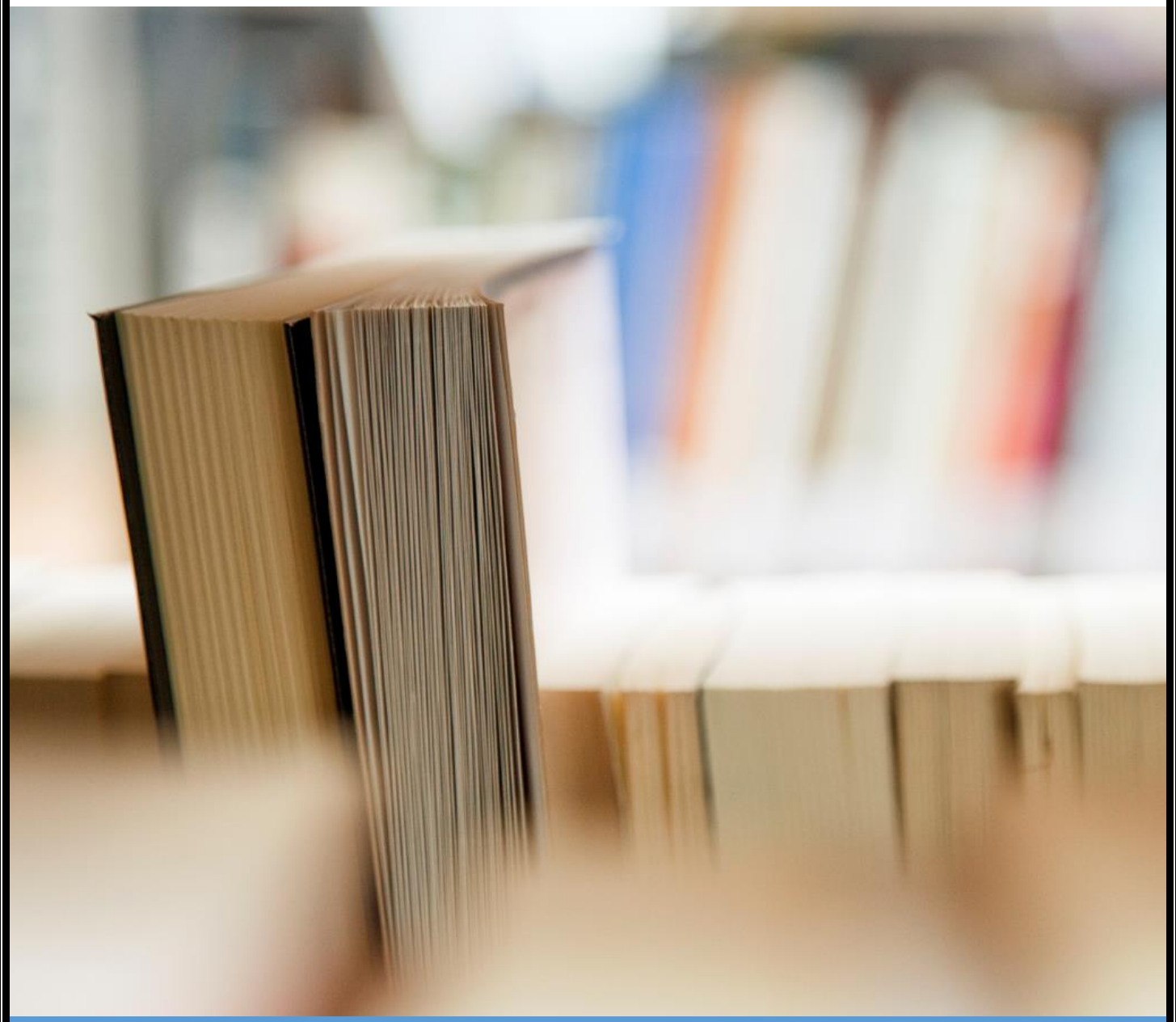

Product Failure Categorization by Natural Language Processing

Detailed Project Report (DPR)

Jateendra Kumar Pradhan 8/4/21 Project Details

Prepared for Internal Purpose Only

Created by :	**Jateendra Kumar Pradhan** **Machine Learning Batch (21st Nov 2021)**
Reviewed by : ****TBD*	**Sudhanshu Kumar** **iNeuron Expert Team**

Q1 . Tell me about your current project ?

This project is a POC which we have developed as a prototype to present to our Client .

Generally every product known as PID (Product Identifier) that is being sold to end customer directly or indirectly via Distributors might fail when installed at their working site . PID mostly here are servers , routers , Storage work stations , Interfaces & Modules , Optical Networking , etc . If these products will fail to get installed , then a Return Material Authorization to be submitted by the end customer and log a ticket to seller . A lot of conversation happens between seller and end customer via Issue logging tool . We analyses this data by text processing and find the product failure reason code with the help of ML Algorithm and Natural Language Processing .

Q2 . What is the size of the data ?

Issue logging tool will have all the information of failure PIDs and transcript of chat/email/phone communication texts . The dataset we used from 2015 are : 2.3 millions of PID information .

Q3 . What was the datatype

Every Device or PID consists of 33 types of attributes . They are combination of Int , Floats , String , Date types .

Q4 . What was the team size & distribution

Below are the Structure of the team :

Failure Analysis Team		
Delivery Manager	1	Onsite
Program Manager & Domain Expert	1	I Basically Manages overall Program execution .
Project Lead	1	
UI Developers	1	
Data Scientists	2	
Data Analyst	2	
Deployment Engineer	1	

Q5 . Which databases you were using ?

There was a Hadoop team who has all the Master data of PID's Failure transcript logged via Logging tool . They pass on the data to Oracle Database what we plan for certain date range and defined no. of columns . We use O**racle Database 12c** as our database .

Q6 . What are the version of Databases

Here is the Database versions : O**racle Database 12c**

Q7 . What was the size of Infrastructure setup ?

For POC purpose we used Virtual Machine Infrastructure setup with below configuration :

Processor	Intel i7
CPU	8 Core Processor
Installed Memory (RAM)	16GB
Storage Area Network	128GB
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 servers . 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 . As this a POC project we never deploy anything to Production server .

We had below instances of servers .

Development environment	2 instances
Staging environment	1 instances
Production environment	POC , so not used

Q9 . How were you creating and maintaining the logs ?

While developing the model we keep track of all logging mechanism at every stage of the project .

Below are the log file maintained :

- User logging and Transaction logs
- Data collection from of Databases logs
- Data Validation logs
- Data Insertion to Oracle 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 :

⇒ Oracle schema and Oracle DB connections to various sources of tools to gather raw data .

⇒ Datatypes and valid values for columns .

⇒ Identify important and required columns by using Heat-map incase of any columns relations .

⇒ Handling NULL values .

⇒ Missing values handling by KNN Imputer .

⇒ Removing outliers by IQR technique , BOX plots .

⇒ Filtering out unnecessary data as per Business requirements using Pandas and Numpy libraries .

⇒ Scaling of data by Standard Scalar due to high difference in magnitude of data .

⇒ One hot encoding to convert categorical data to numerical data .

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

Q11 . How were you maintaining the failure cases ?

With this project model we mainly have used Clustering technique to categorize Product Failure reason .
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 . Wit this data our Data Analyst explore the failure reason and come up with solution . Once fixed will be included to training dataset , else will be sent back to Issue logging team .

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

Full fledge automation is in place as part of data processing . This involves data extraction from Oracle , Segregation of all raw data to a particular Data Frame format , Checking required number of 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 which is set at 8 PM IST time and it will pick all the PIDs information which are being logged as part of issue due to any installation failure at end customer side .

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 it with proper root causes .

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 . I interact mostly with Clients for all the requirement gathering , Hadoop team for data gathering , internal team for Model development . Responsible for Cloud Infrastructure setup for Deployment . Dash boarding and future forecasting of failures – This will enable how much budget to allocate for a particular section of Business Unit . Post Deployment Support and Issue fixing support after Go-Live .

Q16 . What was your day to day task ?

My daily work involves Interacting with Hadoop team , data gathering , guiding and reviewing Data Analyst's daily tasks . Also get engage with Data Scientists and Leads for Development guidance and progress on daily basis . 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 Hadoop team	Once a week

Q17 . In which area you have contributed the most ?

I contributed most in Architectural plan , raw data gathering from Hadoop team , Domain Expertise for the whole project . Next level of support was Data Mining and validation , segregation and preprocessing . Also work for best ML Algorithm modeling , UAT support for Product owners . Also end to end deliverables & client interaction .

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 mostly in Machine Learning .

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 Hadoop team as they are responsible for pushing Mass data to Oracle which is being referred by our team . 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 other 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 including this one and Client agreed to fund them for next 2 years of Budget planning for FY22 and FY23 .

Q21 . How were you doing Deployments ?

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 . 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 which all predictions will be carried out by those files directly in . Deployment to production is yet to be done for this project as it's in POC phase .

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

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

- Data Mining and understanding texts entered by different analysts into Case Logging tools .
- Mass processing of Text and text analytics takes a lot of time during preprocessing and actual training . So later we enhanced our Infrastructure setup to handle large amount of data .
- During UAT testing phase we were getting many Unknown failure reasons without proper failure categorization . So we take them to Client , understand their failure categorization and then retrain the model again .

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 we park them aside 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 , Failure reasons categorization , 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 Text Transcript of PIDs and this happens every 15 days , so as to make the dataset fully complete . The raw data which we get directly from Hadoop team will have all the information .

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 Product Owners .
- 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 Product Owners .
- 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 2.3 million of PID records transcript and with hyper parameter turning and all models predictions , it take around 11 hours to train the model .

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

We retrain our model twice in a month on : 12th & 27th .

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 .