Detailed Project Report on Scania Truck Failure

What is the objective of the project?

The objective of the project is to build an end to end system which will be able to predict the failure of the truck’s APS (Air Pressure System)

Tell me about your current project?

The project is called as a Scania Truck Failure System. The goal of the project is to predict the failure of the APS (Air Pressure \System) based on the parameters. The parameters name or columns names are actually encrypted for security reasons. The dataset consisted of 60000 rows and 171 columns of training data. For the sake of simplicity we have split the data into 3000 rows per file, and created multiple batches of data.\These batch data can even come from a Big Data Pipeline which would be pick up the sensor data and put them in csv files which is stored in S3 buckets. So based a particular frequency, data will be continuously be stored in S3 buckets. Now data is stored in S3 buckets we can retrieve data from S3 buckets to perform data validation, data transformation, data pre-processing and model training, etc. Once model training is done MLFlow is used for experiment tracking and monitoring of metrics, parameters and model itself and transition them to production and staging, or staging and archived based on the condition.. After the model training is don, for a new batch of data prediction data\the prediction data validation, data transformation and data pre-processing steps are done, are the based on the cluster number a model is picked up and then prediction is done. The system is containerised using Dockers and deployed in AWS ECR and ECS.

What is the size of your data?

The size of the data in terms of MB is 45, with 60000 columns and 171 columns,

Column names are encrypted for security reasons

What was the data type?

The columns consisted of both float and string values with missing values and some invalid values like “?”

How are creating and maintaining logs?

The logs are maintained inn MongoDB. The logging starts with the start of application and for custom function which have created enrry and exit logs are created to know about the time complexity of the function with respect to the operations, the function was supposed to do There are loggings for error scenarios and except block as well.

What are the techniques were you using for data pre-processing for various data science use cases and visualizations?

To ensure that the data is given to us correct format, we have to sign a data sharing agreement with client to ensure that the right numbers of files are coming from their system, filename should be according the data sharing agreement column length etc. Based on this given from functional team, we will prepare the master management which is nothing but the schema training json file and schema prediction json file. This acts source of ground truth to us for our data validation process. Any file or data column which does not agree to master data management will be pushed to bad data folder in S3 buckets.

For the pre-processing, we used standard scaler and PCA transformation for dimensionality reduction. Replaced invalid values with nan values, encoded the target values, imputed the missing values, removed columns which had no meaning at all.

What kind of the automation was done for data pre-processing?

We had a full-fledged data pipeline which would extract the data from the S3 bucket, which had come from Big Data pipeline created by the client side.

How are monitoring your job?

There are logging setup done. We can regularly monitor the logs to see for any error messages or scenarios.

What were your roles and responsibilities in the project?

My responsibilities included writing code to extract the data from S3 buckets, and perform data validation operations, data transformation operations and database operations in MongoDB for the first half of the project. Then I was involved in MLFlow integration and setup in cloud also in CI-CD operations with ECR and ECS deployment.

What was day to day task?

My day to day task involved writing code the data validation pipeline, performing database operations and involved in MLFlow integration and cloud setup of MLFlow

In what areas you have a contributed the most?

I have contributed the most in data operations like validation, transformation and in setting up MLFlow in cloud and writing code related MLFlow on how our application we communicate MLFlow to perform experiment tracking and monitoring of metrics, parameters and models.

In which technology you are most comfortable?

I am comfortable in Machine Learning, Deep Learning and little bit of AIOps. But if we are talking personal preference Natural Language Processing and Machine Learning.

In how many projects you have worked on?

I continuously experiment my learning on doing small projects and some medium level projects. But to give a rough idea, I have worked on object detection using Detectron2, some machine learning projects both classification and regression problem statement along with AIOps integration.

\

How are you doing the deployment?

Before the deployment was done, the entire application was containerised using Docker and deployed in AWS ECS and CI-CD integration tool being GitHub Actions.

What are the challenges have you faced during the project?

The biggest challenge is that getting good data for and building and robust model and would not take a lot of time to load and more accurate in the terms predictions and reduce more errors. Coming to code side, figuring out get the data S3 bucket without downloading the csv file from S3, and then keeping the entire pipeline run on cloud setup without local system interface, only thing will be in local is code developed. Even the trained models are saved to S3 buckets and loading of the model for prediction is also from the S3 bucket itself. Coming to AIOps part setting up MLFlow in cloud was challenging and as it involved creating Nginx configuration and Linux service management and attaching an elastic IP so that the endpoint url does not change whenever we restart the MLFlow application also ensure the security of the MLFlow server using Nginx and apache utils.

How did you the data validation?

Whenever we try to solve any problem using AI, with the client we sign a data sharing agreement involved type of data, which will be receiving for example let’s say we csv files are given to us, we should sign off the filename pattern column length, date time stamp in file, etc. with functional team. So based on this data a master data management file is created, according that the data validation is performed in our case nothing but the schema file.