FRAUD CLAIM DETECTION

Objective:

Development of a predictive model for monitoring fraud insurance claims. The model will determine whether a customer is placing fraudulent insurance claim or not.

Value of Project:

- Give better insight of customer base
- Detection of upcoming frauds
- Helps in easy flow for managing resources
- Prior information about the faulty sensor in wafer

Data Sharing Agreement:

- Sample file name(ex fraud_detection_0003201_0101010)
- Length of data stamp(8 digits)
- Length of timestamp(6 digits)
- Number of columns
- Columns name
- Columns data type

Data Validation and Data Transformation:

- **Name Validation-**Validation of File name according to our created regex as per DSA. If it validates then we move these files to GOOD_DATA_FOLDER, else in BAD_DATA_FOLDER.
- **Number of Columns-**Validation of no. Of columns as per by DSA. If it validates then we move it in GOOD_DATA_FOLDER, else in BAD_DATA_FOLDER.
- **Name of Columns-**Validation of name Of columns as per by DSA. If it validates then we move it in GOOD_DATA_FOLDER, else in BAD_DATA_FOLDER.
- **Data type of Columns-**The data type of columns is given in schema files and will be validated at the time of insertion of files into Database. If the data type is wrong then we will move the data into BAD_DATA_FOLDER.
- **Null Values in Columns-**If any of the columns in file having all values as NULL or Missing then we will move such files into bad data folder.

Data insertion into Databases:

- Database Creation-Create a database and if database is already present according to proposed name then establish a connection to Database.

- Table Creation-Creation of a table as GOOD_DATA and if already present then we have to move our new files which are validated according to given schema into it for training along with old present files in the table.

- File Insertion-Insertion of all valid files having same data type from Good Data Folder to the GOOD_DATA table. And those files which are invalid shall not be loaded into table they will be moved to Bad Data Folder.

Model Training:

- Data Export from Database-The data stored in database is exported in csv file format for model training.

Data Pre-processing:

- Checking of null values present into the columns. Impute them using KNN imputer.
- Checking for zero standard deviation of any columns and removing those columns because a column having zero standard deviation does not provide any useful information in the process of model training.

Clustering:

- KMeans model created during training is loaded and clusters for the pre processed prediction data is predicted.

Model Selection – After the clusters have been created, we find the best model for each cluster. We are using two algorithms, “SVM” and "XGBoost". For each cluster, both the algorithms are passed with the best parameters derived from GridSearch. We calculate the AUC scores for both models and select the model with the best score. Similarly, the model is selected for each cluster. All the models for every cluster are saved for use in prediction.

Prediction :

- Based on the cluster number, the respective model is loaded. and is used to predict the data for that cluster.
- Once the prediction is made for all the clusters, the predictions along with the wafer names are saved in a csv file at a given location and the location is returned to the client.

Deployment:

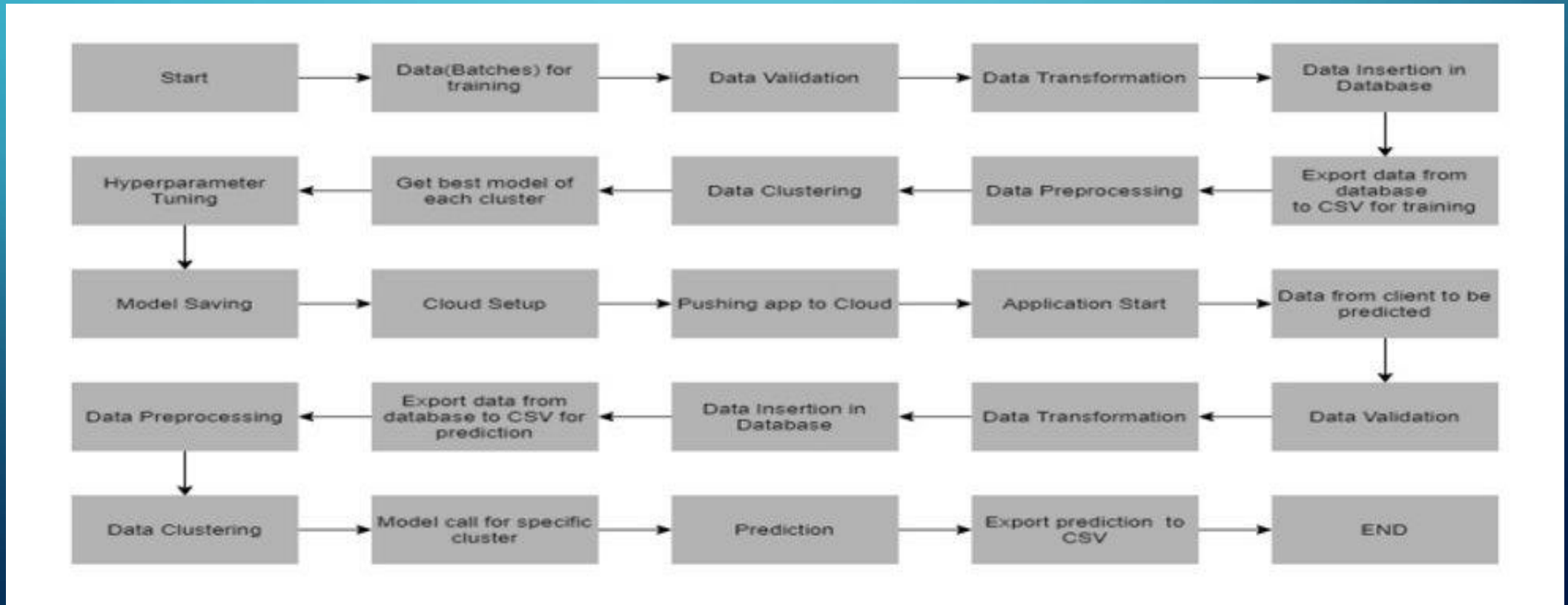
We will be deploying the model to the Pivotal Cloud Foundry platform.

Q&A Section

1. What is the source of Data ?

The data for training is provided by the client in multiple batches and each batch contains multiple files.

2. What is the complete flow you followed in project?



3. What was the type of Data?
The data was the combination of numerical and categorical values.

4. How did you validate the data?
We validate the input files as per the DSA. We create regex as per given schema file on that basic we validated the data.

5.How did you deal with invalid Data?
We create a file as Good data and Bad Data . If data validates then we move such files to good data folder for further insertion into Good_Data table in database. Otherwise we will move such files to Bad Data folder and resend it to client for refinement.

6. How logs are managed?
We are using different logs as per the steps we follow in validation and modeling like file validation log, data insertion, model trainin log etc.

7. What techniques were you using for data pre processing?
- Checking of null values present into the columns. Impute them using KNN imputer.
- Checking for zero standard deviation of any columns and removing those columns because a column having zero standard deviation does not provide any useful information in the process of model training.

8. what models were used in Training ?

- SVM
- XGBoost

9. How did you select the model for each cluster set?

We provide both the algorithms for each clusters with the best parameters derived from GridSearch. And we calculate AUC scores for each model. We select models on the basic of higher AUC score.

10.How pediction was done?

Based on the cluster number, the respective model is loaded and is used to predict the data for that cluster. Once the prediction is made for all the clusters, the predictions along with the Wafer names are saved in a CSV file at a given location and the location is returned to the client.

11. What are the different stages of of Deployment?

- When the model is ready we deploy it to the Pivotal Cloud Foundry platform.

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

My responsibilities consisted of Understanding the Problem statement, Creating HLD and LLD ,Data Cleaning, Data clustering, Model selection for training.

13.**What was your day to day task?**

My day to day tasks involved completing the JIRA tasks assigned to me, attending the scrum meetings, participating in design discussions and requirement gathering, doing the requirement analysis, data validation, Unit test for the models, providing UAT support, etc.

14.In which area you have contributed the most?

I contributed the most to HLD and LLD areas . Also, we did a lot of brainstorming for finding and selecting the best algorithms for our use cases. After that, we identified and finalized the best practices for implementation, scalable deployment of the model, and best practices for seamless deployments as well.

15-In which technology you are most comfortable?

I have worked in almost all the fields viz. Machine Learning, Deep Learning, and Natural Language Processing, and I have nearly equivalent knowledge in these fields. But if you talk about personal preference, I have loved working in Machine Learning the most.

16.How you rate yourself in big data technology?

I gave worked often in the big data computing technology with ample knowledge in distributed and cluster-based computing. But my focus and extensive contribution have been as a data scientist.

.17 In how many projects you have already worked?

It's difficult to give a number. But I have worked in various small and large scale projects,e.g., object detection, object classification, object identification, NLP projects, chatbot building, machine learning regression, and classification problems.

18.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.

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

The biggest challenge that we face is in terms of obtaining a good dataset, cleaning it to be fit for feeding it to a model, and clustering the dataset . Then comes the task of finding the correct algorithm to be used for that business case.

Then that model is optimized. If we are exposing the model as an API, then we need to work on the SLA for the API as well, so that it responds in optimum time

20-What will be your expectations?

It's said that the best learning is what we learn on the job with experience. I expect to work on new projects which require a broad set of skills so that I can hone my existing skills and learn new things Simultaneously.

21-What is your future objective?

The field of data science is continuously changing. Almost daily, there is a research paper that changes the way we approach an AI problem. So, it really makes it exciting to work on things that are new to the entire world. My objective is to learn new things as fast as possible and try and implement that knowledge to the work that we do for better code, robust application and in turn, a better user/customer experience.

22. Why are you leaving your current organization?

I was working on similar kinds of projects for some time now. But the market is rapidly changing, and the skill set required to be relevant in the market is changing as well. The reason for searching a new job is to work on several kinds of projects and improve my skill set.

23-How would you rate yourself in machine learning?

Well, honestly, my 10 and your 10 will be a lot different as we have different kinds of experiences. On my scale of 1 to 10, I'll rate myself as an 8.2.

24. How would you rate your self in distributed computation?

I'd rate myself a 7.7 out of 10.

25.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.

26.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.

27.How did you optimize your solution?

Well, model optimization depends on a lot of factors.

- Train with better data(increase the quality), or do data pre-processing steps more efficiently.
- Increase the quantity of data used for training.
- If you are not using transfer learning, then you can alter the number of hidden layers, activation function.
- Change the function used in the output layer based on the requirement. The sigmoid functions work well with binary classification problems, whereas for multi-class problems, we use a sigmoid model.
- Try and use multithreaded approaches, if possible.
- Reduce Learning Rate in plateau reasons optimizes the model even further.

28.How much time did your model take to get trained?
MY model did take 15 minutes for training.

29.At what frequency are you retraining and updating your model?
The model gets retrained at every 20 day time interval.

30. 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.

31. What is your area of specialization in machine learning?
Answer:
I have worked on various algorithms. So, It's difficult to point out one strong area. Let's have a discussion on any specific requirement that you have, and then we can take it further from there.