PREDICATIVE MAINTENANCE

JINENDRA SONTAKKE

OBJECTIVES

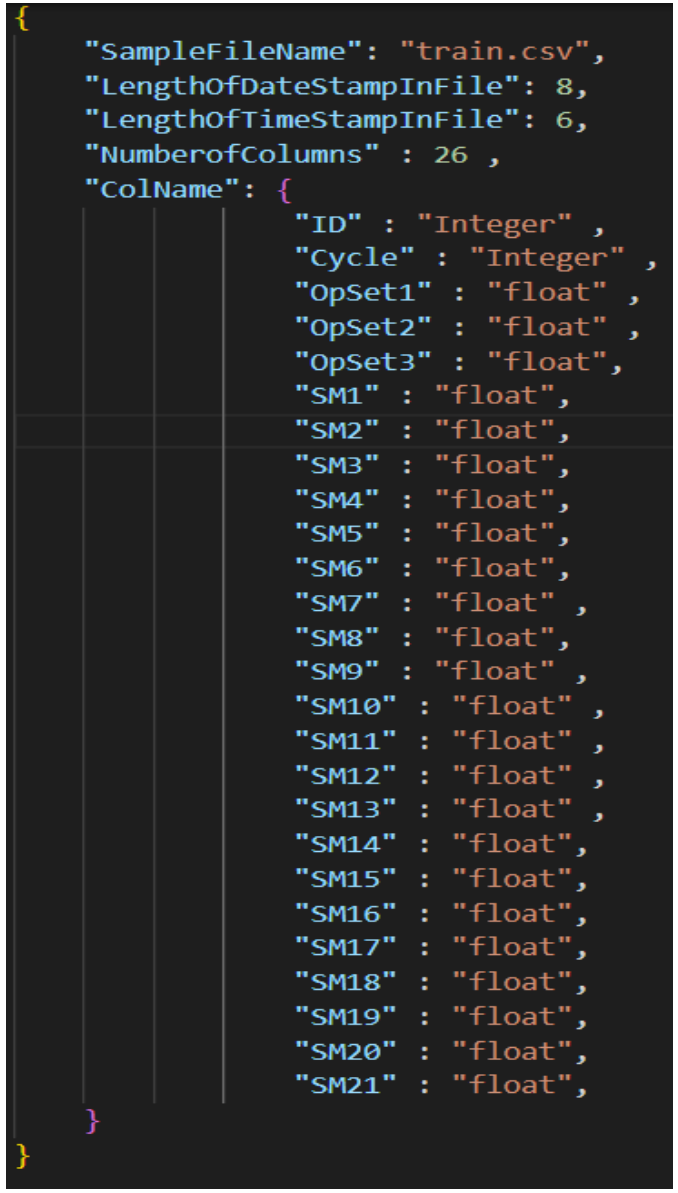
- A jet engine is the most important part of the plane, and it helps to move the airplane forward with a great force that is produced by a tremendous thrust Maintenance of equipment is a critical activity for any business involving machines. And causes the plane to fly very fast. During operation, degradation occurs in each of the components. If the degradation level is any component exceeds a threshold the engine is said to have failed. Therefore, the jet engines are inspected before every take-off and maintenance of equipment is critical activity for any business involving machines.
- Predictive Maintenance is the method of scheduling maintenance based on the prediction of the failure time of any equipment. The prediction can be done by analyzing the data measurements from the equipment. Machine learning is technology by which outcomes can be predicted based on a model prepared by training it on past input data and its output behavior. The model developed can be used to predict machine failure before it happens.

BENEFITS

- Predictive Maintenance is the method of scheduling maintenance based on the prediction of the failure time of any equipment. It reduces the possibility of system failure.
- predictive Maintenance help to reduced maintenance costs.
- It also help us to increase the equipment lifespan. Due to better management the life expectancy also increases.

DATA SHARING AGREEMENT FILE FOR TRAINING DATA SET

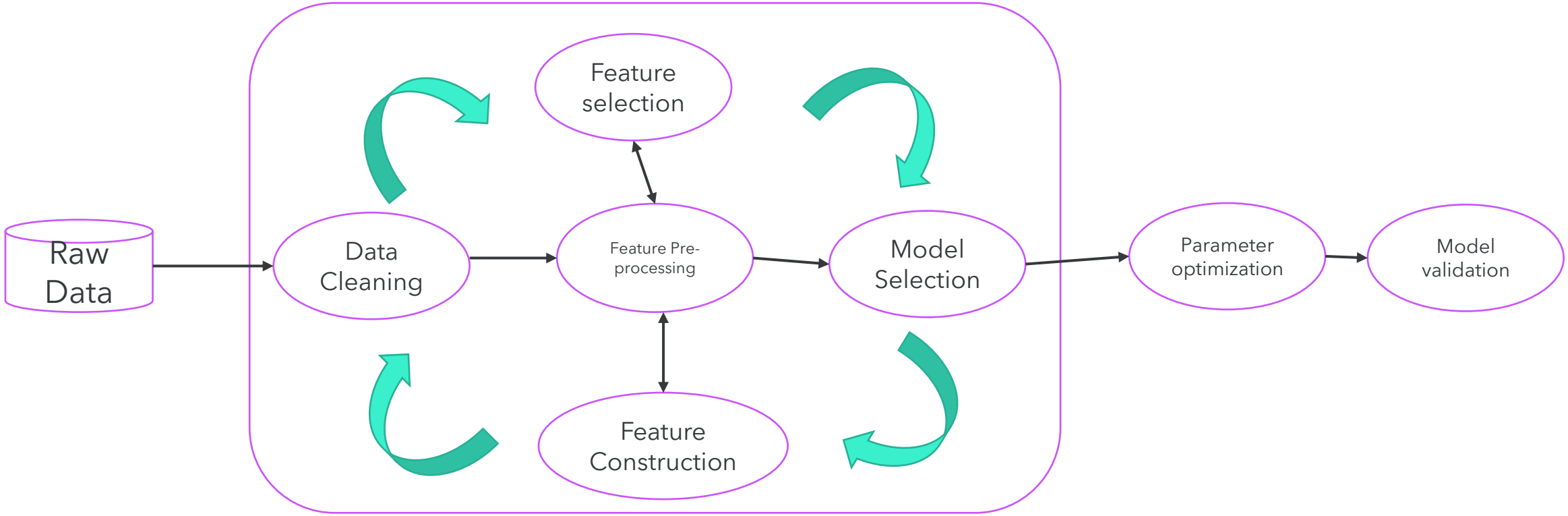
- Sample File Name (Ex train.csv)
- Length if data stamp(8 digits)
- Length of time stamp(6 digits)
- Number of columns
- Number of columns names
- Columns Data type
- Columns details



DATA DESCRIPTION

- ID : Engine ID
- Cycle : Engine Cycle
- OpSet : Operational setting (OpSet1 to OpSet3)
- SM : Sensor Measure (SM1 to SM21)

APPLICATION ARCHITECTURE



DATA VALIDATION & TRANSFORMATION

- **File Name Validation :** File name validation as per the DSA. We have created regular expression pattern for validation with time stamp format. IF this file satisfy the pattern criteria it will go to valid_data_file folder or else it will go to bad_data file folder.
- **Name and no. of columns :** It will check for number of columns and Name of the columns. If it will pass the validation criteria then it will goto valid_data_file or else bad_data_file .
- **Data types of column :** The datatype of columns is given in the schema file. This is validated when we insert the files into database. If the datatype is incorrect, then the file is moved to "bad_data_folder".
- **Null values in Columns :** If any columns in file contain null values or data ids missing then this file is going to bad_file_folder.

MODEL TRAINING

- The data is exported as A CSV file to be used for model training.
- **Exploratory Data Analysis & Data Pre-processing:**
    1. Missing value count
    2. No of rows and columns(Shape)
    3.) categorical/Numerical columns
    4. Correlation heat map
    5. Null value handling(impute null value)
    6.) Outlier detection and remove
    7. Perform standard scalar for scaling down

MODEL SELECTION

- In hyperparameter tuning, we have implemented randomized search cv or grid search cv and from that, we also implemented cross-validation techniques for that. From that, we have chosen the best parameters according to hyperparameter tuning and the best score from their accuracies
- Finally, , we fit the random forest model with optimal tuning parameters on the entire dataset. We then could use this model to predict whether the engine condition is good or not.

PREDICTION

- The testing files are shared in batches, and we perform the same validation operations, data transformation, and data insertion on them.
- The accumulated data from the database is exported in CSV format for prediction.
- We perform data pre-processing techniques in it.

QUESTION & ANSWERS

Question 1: Explain about the Project and your day to day task :

Answer : The normal average of flight operations per day is around 1,75,000 according to the Flghttradar24 and flight generates gigabytes of data. With help of this data, we can predict when a component failure might occur and secondly prevent the occurrence of the failure by performing maintenance. It minimizes the time and also minimizes the cost of spare part.

As a data scientist I am involving in each an every phase of the project. My responsibility consisted of gathering the dataset ,labelling the data for the model, training the model on the prepared dataset , deploying the training model to the cloud, monitoring the deployed model for any issues. Mixed in are calls, stand ups and the attending Scrum meeting.

Question 2 : How Logs Are Managed?

Answer : We Are Using Different Logs As Per The Steps That We Follow In Validation And Modeling Like File Validation Log , Data Insertion ,Model Training Log , Prediction Log Etc.

Question 3 : What is the source and size of data ?

Answer : The data for train is provided by client in batches . Size of the data usually in MB

Question 4 : How Prediction Was Done?

Answer : The Testing Files Are Shared By The Client .We Perform The Same Life Cycle Till The Data Is Clustered. Then On The Basis Of Cluster Number Model Is Loaded And Perform Prediction. In The End We Get The Accumulated Data Of Predictions.

Question 5 : What is AUC Curve ?

Answer : AUC stands for "Area under the ROC Curve" .AUC measures the entire 2D area underneath the entire ROC curve.

Question 6 : What Is The Type Of Data?

Answer : The Data Is The Combination Of Numerical And Categorical Values.

Question 7 : What techniques r you using for data pre-processing ?

Answer : 1) Removing unwanted attributes.
2) Visualizing relation of independent variables with each other and with dependent variable.
3) Removing Outliers.
4) Cleaning data and imputing if null values are present.
5) Convert Categorical data to numerical data.
6) Scaling the data

Question 8 : Does Your Dataset Show Normally Distributed Or Not? If Not Then Which Techniques You Will Use To Make It Normal?

Answer : No, These Data Set Does Not Show Normal Distribution Behavior. I Used Reciprocal, Square, Log, Exponential Techniques To Make It Normally Distributes.

Question 9 : Which Tool You Are Used For Implementation This Model?

Answer : 1) Ide : Visual Studio Code
2) Cloud : Azure

Question 10 : How were you maintain the failure cases?

Answer : If our model is not predicting correctly for data then that dataset goes to database . There will be a report triggered to the support team at the end of the day with all failure scenarios where they can inspect the failure. Once we have a sufficient number of cases we can label and include those data while retraining the model for better performance.

Question 11 : In which technology you are most comfortable?

Answer : I have worked o machine learning ,Deep learning and NLP. But personally I prefer deep learning.

Question 12: What Kind of challenges have u faced during the project?

Answer : The biggest challenge I face in project is in obtaining good dataset , cleaning it to be fit for model and then labeling prepared dataset. Labeling is a time consuming task and it takes lots of our. Then comes the task of finding the correct algorithm to be used for business case.

Question 13 : What Is Accuracy ?

Answer : Accuracy Is One Metric For Evaluating Classification Models. Accuracy = Number Of Correct Predictions /Total Number Of Predictions

Question 14 : What will be your expectations?

Answer : I expect to work on different projects to enhance my technical skill and learn new things simultaneously.

Question 15 : What is your future objective?

Answer : My future objective is to learn new things in AI field because it changes continuously, and my aim is to pursue my career as a solution architect near future.

Question 16 : How did you optimize your solution?

Answer : 1) Model optimization depends on various factors
2) Train with better data or do data pre-processing in efficient way.
3) Increase the quantity of training data etc.
4) Try and use multithreaded approaches

Question 17 : At what frequency are u retraining and updating your model?

Answer : The model gets retrained every 30 days

Question 18 : How did you optimize your solution?

Answer :

1) Model optimization depends on various factors
2) Train with better data or do data pre-processing in efficient way.
3) Increase the quantity of training data etc.
4) Try and use multithreaded approaches

Question 19 : At what frequency are u retraining and updating your model?

Answer : The model gets retrained every 30 days

Question 20 : Which is more important to you model accuracy or model performance?

Answer : Well, you must know that model accuracy is only a subset of model performance. The accuracy of the model and performance of the model are directly proportional and hence better the performance of the model, more accurate are the predictions.

Question 21 : What is Overfitting, and How Can You Avoid It?

Answer : Overfitting is a situation that occurs when a model learns the training set too well, taking up random fluctuations in the training data as concepts. These impact the model's ability to generalize and don't apply to new data.
When a model is given the training data, it shows 100 percent accuracy technically a slight loss. But, when we use the test data, there may be an error and low efficiency. This condition is known as overfitting.
There are multiple ways of avoiding overfitting, such as:

- Regularization. It involves a cost term for the features involved with the objective function
- Making a simple model. With lesser variables and parameters, the variance can be reduced
- Cross-validation methods like k-folds can also be used
- If some model parameters are likely to cause overfitting, techniques for regularization like LASSO can be used that penalize these parameters

Question 22 : Explain the Confusion Matrix with Respect to Machine Learning Algorithms ?

Answer : A confusion matrix (or error matrix) is a specific table that is used to measure the performance of an algorithm. It is mostly used in supervised learning; in unsupervised learning, it's called the matching matrix.
The confusion matrix has two parameters:

- Actual
- Predicted

It also has identical sets of features in both of these dimensions.

Question 23 : How Will You Know Which Machine Learning Algorithm to Choose for Your Classification Problem?

Answer : While there is no fixed rule to choose an algorithm for a classification problem, you can follow these guidelines:

- If accuracy is a concern, test different algorithms and cross-validate them
- If the training dataset is small, use models that have low variance and high bias
- If the training dataset is large, use models that have high variance and little bias

THANK YOU

Jinendra Sontakke