Predictive Maintenance using Machine Learning

Predictive maintenance is mainly used to understand effect of component degradation on the performance of equipment. Machine learning plays a key role in predicting failures, by using suitable ML algorithm we can predict the failures thus lowering maintenance cost, by analyzing data and using required algorithms efficient failure prediction model can be constructed to minimize the effect of failure within the system. Analysis of data as per the requirements is an important step in constructing an efficient model followed by gathering and labelling of data, and obtaining a model with high accuracy using time series and Machine Learning algorithms. Following steps are involved in developing an accurate failure prediction model-

1) Collection and preprocessing of data

The data is collected with the purpose of providing failure specifics for system and components in as much detail as possible so that analysis might produce some useful findings. Data is stored in a database with required labelling for analysis of dependent variables and in depth and transforming our data for a machine learning model, selection of an appropriate failure window will always depend on the context of the problem and duration between machine failure. After importing data, analysis of data fields is done and dependent variables are examined to predict the failure rate and required transformation in datasets are performed.

2) Predicting Failure using Time series

It is basically based on number of failures occurred in the system over a period of time, a classical time series model called autoregressive integrated moving average (ARIMA) is developed to predict failure using time series it is a combination of AR (auto regression) and MA (moving average). In auto regression failure us predicting on the basis of past causes data is sorted on the basis of duration between failures and in moving average, erratic failures are analyzed by taking average of past analysis.

3) Expanding failure Scenario

Failure scenario of datasets cause imbalance in data thus reducing chances of efficient prediction. Firstly, datasets are categorized in test set, validating set and training set randomly to test the accuracy of required model, then we further proceed to create a balance training dataset using SMOTE (Synthetic Minority Oversampling Technique). this technique is used to have equal failure rate for efficient analysis of data, it simply duplicates examples from the minority class in the training dataset prior to fitting a model. This can balance the class distribution but does not provide any additional information to the model. An improvement on duplicating examples from the minority class is to synthesize new examples from the minority class. The approach is effective because new synthetic examples from the minority class are created that are plausible, that is, are relatively close in feature space to existing examples from the failure class, thus providing better analysis.

3) Using Machine Learning Techniques to Frame model

After obtaining balanced data set, machine learning techniques are used to frame accurate model. The methods considered are-

LDA (linear discriminant analysis)

It is a classification technique in machine learning which analyze data my performing dimensionality reduction. For ex if we use a single feature to predict failure and neglect other features the model so obtained won't be effective thus LDA technique is used to reduce multiple variables in single variable thus minimizing the variance. The representation of LDA is pretty straight-forward. The model consists of the statistical properties of your data that has been calculated for each class. The same properties are calculated over the multivariate Gaussian in the case of multiple variables. The multivariate are means and covariate matrix. Predictions are made by providing the statistical properties into the LDA equation. The properties are estimated from your data. Finally, the model values are saved to file to create the LDA model.

CART (Classification and Regression Tree)

It is based on Classification and Regression Trees (decision trees) which basically contains a root node of dependent variable thus further dividing into predictors, the end of the nodes contains prediction for the desired variable. It can be considered as the tree is traversed by evaluating the specific input started at the root node of the tree. A learned binary tree is actually a partitioning of the input space, each input variable can be considered as a dimension on a p-dimensional space. The decision tree split this up into rectangles or some kind of hyper-rectangles with more inputs. New data is filtered through the tree and lands in one of the rectangles and the output value for that rectangle is the prediction made by the model.

RF (random forest)

It is an algorithm used for both Classification and Regression tasks, the entire algorithm is based on combining decision tress to create a random forest, random is used as random datasets are used to create forest. In this case the prediction is made by averaging. It provides a more accurate and stable prediction as compared to that of decision tress.

SVM (Support Vector Machine)

Support Vector Machine is a supervised machine learning algorithm it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n-dimensional space with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes very well. Unlike LDA It mainly focuses on points that are difficult to classify, and it is discriminative, it is great when there is a clear margin of separation between classes. SVM is more effective in high dimensional spaces and in cases where the number of dimensions is greater than the number of samples and is relatively memory efficient.

KNN (K Nearest Neighbour)

It is an algorithm used for both Classification and Regression, for prediction the KNN algorithm will use the entire dataset. Indeed, for an observation that isn't part of the dataset and is the actual value we want to predict, the algorithm will look for the k instances of the dataset closest to our observation. In simpler words If KNN is used for a regression problem, the mean (or median) of the x variables of the k closest observations will be used for predictions. If KNN is used for a classification problem, it's the mode (value that appears most often) of the variables x of the k closest observations that will be used for predictions. It uses feature similarity to predict the values of any new data points. This means that the new point is assigned a value based on how closely it resembles the points in the training set.

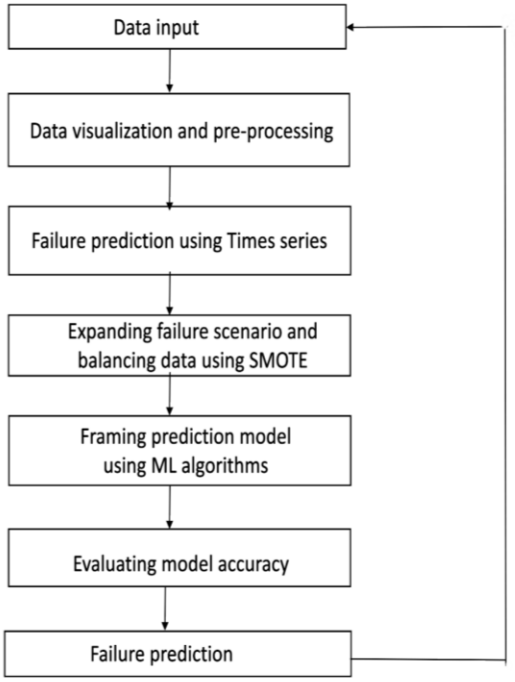
4) Model Accuracy Evaluation

The developed models are evaluated and compared on the basis of accuracy and feasibility of model using test data-sets. To evaluate accuracy of a classification model, confusion matrix is plotted consisting of following outcomes-True positives, True negatives, False positives, False negatives. For regression models explained variance, mean squared error metrics are used for accuracy evaluation.

5) Prediction

Thus, the most effective model on the basis of accuracy is used for the prediction of failure.

SCHEMATIC FOR FAILURE PREDICTION



Implementation of Machine Learning Models for machine failure prediction

To create an efficient Machine learning model, first and foremost process is preprocessing of data followed by implementing various ML algorithms to obtain model with high accuracy.

Following are the steps -

PREPROCESSING OF DATA

1. Importing the required dataset and exploring it to analyze main feature and dependent variables, our aim is to predict machine failure i.e., "Fail_tomorrow" in the given dataset.

2. Transforming the data, by replacing null values in numerical features with mean values as deleting columns will result in loss of data and decrease accuracy of model.

3. Encoding the data in all numeric values, as machine learning models completely work on numbers and for categorical data to avoid correlation issues dummy encoding is used.

4. Splitting dataset into training and test datasets to enhance performance of model, after splitting feature scaling of data is done to standardize variables in specific range to avoid domination of variables.

MACHINE LEARNING MODELS IMPLEMENATION and ACCURACY EVALUTION USING CONFUSION MATRIX

1. KNN-Using scikit module KNN using Kd tree algorithm is used for prediction and obtained accuracy rate is 83.9%.

2. LDA-using scikit module implementing linear discriminant analysis for prediction and obtained accuracy rate is 84.28%.

3. SVM- using scikit module implementing support vector machine with SGD classifier for prediction and obtained accuracy rate is 84.37%.

4. RF- using scikit module implementing random forest classifier technique for prediction and obtained accuracy rate is 86.01%.

5. Decision Tree- using scikit module implementing decision tree classifier for prediction and obtained accuracy rate is 78.66%.

Using confusion matrix to evaluate accuracy of model, Confusion matrices are used to visualize important predictive analytics like accuracy. Confusion matrices are useful because they give direct comparisons of values like True Positives, False Positives, True Negatives and False Negatives.

Based on above implementation of various ML algorithms and checking accuracy most accurate ML algorithm for Failure Prediction Model is Random Forest Classifier with accuracy rate of 86.01%.

Database source: https://www.kaggle.com/binaicrai/machine-failure-data