Problem Statement

Many different industries need predictive maintenance solutions to reduce risks and gain actionable insights through processing data from their equipment.

Although system failure is a very general issue that can occur in any machine, predicting the failure and taking steps to prevent such failure is most important for any machine. This Capstone project is aimed at predicting the machine breakdown by identifying the anomalies in the data.

The data we have contains about 18000+ rows collected over few days. The column 'y' contains the binary labels, with 1 denoting there is an anomaly. The rest of the columns are predictors.

Data source

'AnomaData' provided with project was used for modelling purpose.

Data preprocessing steps

1. Data info, columns detail, length of data was checked to understand the shape of data.
2. Repeated column ie y was dropped to avoid confusion.
3. Scatter plot was created but since predictor variables are very large in number hence it was difficult to identify any trend.
4. Correlation matrix and heat map was also created for the given variables.
5. Items with high correlation were removed from the data for higher credibility
6. Missing value was checked which was null.
7. Next values were assigned to 'x' and 'y' variables for modelling purposes.
8. Lastly data was split into train + test + k-fold (for cross validation).

Feature engineering

Feature selection was conducted using k-fold and was bar chart was plotted. From the plot 15 top features were identified and used for modelling.

Model selection

After data preprocessing, 3 models were tested manually.

1.linear regression

2. linear regression using selected features.

3. Gradient boost regressor method.

4. Hyper parameter tuning was conducted to identify best modelling technique.

Evaluation metrics.

Three metrics used for evaluation of models are-

1. Mean absolute error
2. Mean squared error
3. Root mean squared error

Findings

Regression model with Gradient boosting regressor method gives best R2 value. The R2 value of 0.8 was achieved which is a good result.

Model Deployment plans

To get predictions, we can also deploy our model to Amazon EC2 using Amazon Sage Maker. To host a model through Amazon EC2 using Amazon SageMaker, we can deploy the model that we trained in Create and Run a Training Job by calling the `deploy` method of the `xgb_model` estimator.

Future recommendations

This model can be used in factories and plants to predict the machine failures and decide the future plans.