SOULPAGE

IT SOLUTIONS

Predictive Maintenance Analysis of Machine Data

Predictive Maintenance Analysis of Machine data

Problem statement:-

Predictive maintenance techniques are designed to help anticipate equipment failures to allow for advance scheduling of corrective maintenance, thereby preventing unexpected equipment downtime, improving service quality for customers, and also reducing the additional cost caused by over- maintenance in preventative maintenance policies. Many types of equipment for e.g., automated teller machines(ATMs), information technology equipment, medical devices, e.t.c. like tracking run- time status by generating system messages, error events, and log files which can be used to predict impending failures.

We have data of 983 records of data available to do analysis and to predict will the system work if the error hits.

Solution:-

We can use Machine learning techniques to predict whether the system/machine works with respect to the error that machine faced at any time on the day.

Technologies used:-

We use Data science and Machine learning models to predict whether the machine fails or works on the respective day, if it faces an error. We use tools like

- python 3.6
- Numpy
- Pandas
- matplotlib libraries along with

machine learning models like

- logistic regression
- k- nearest neighbours
- Decision trees.
- Random forest

We check their accuracy using precision - recall metric to know how well the model predicting the target variable.

Dataset:-

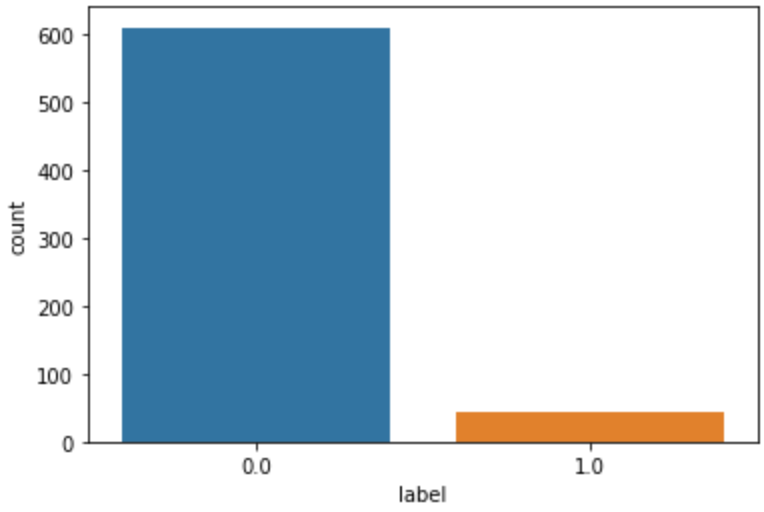
We have features that are the log data of target machine with 26 log errors. There are many different error types when the machine runs throughout the day. Each type of machine log error has some unique id (for example: 136088194). We have a target variable called label which is the failure record of target machine. If the value in the label feature is 0, then it represents that the machine fails to work in entire day where as the value in the label feature is 1, that represents the machine works fine in the day.

We have the dataset with 983 records with 26 features which are log errors that are in the int data type.

Process:-

First once after the data is given, check for Duplicates, Missing values, Outliers and data imbalance. If you find any duplicates in the dataset, then remove them. After that look for missing values, and try to find out how we got null values in the dataset. You can impute the missing values using numerical data such as Mean, median, and mode. You can also impute the missing values using text with upfilling/downfilling. And then check for outliers and try to remove them. Finally check for data imbalance in the target variable.

After checking all the above mentioned preprocessing techniques, we come to know that we have an imbalance in the target variable of our data.

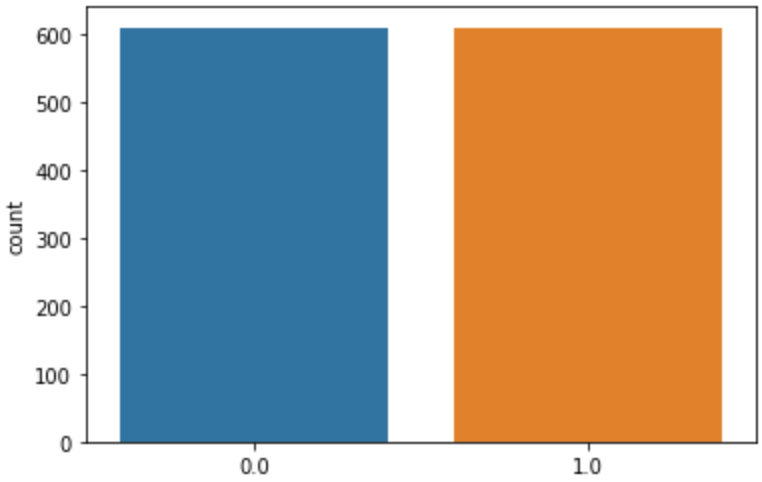


From the graph we can see that we have a huge imbalance in the data which is making our machine learning models to give wrong predictions. Even one percentage of change in predictions can make a big change in the real world.

Hence lets try to balance our data using data balancing techniques. We can make imbalance data:-

- Oversampling/upsampling
- undersampling/downsampling
- Changing accuracy metrics
- Ensemble models

Now we upsample the data using well known oversampling method known as SMOTE. After applying SMOTE , lets have a look at the data now



Now you can see that we got balance data. Lets apply above mentioned machine learning models to predict the label. Split the whole dataset to train and test sets. We also have to change performance metric from accuracy score to precision recall curve.

We applied all the available models to check the algorithm that predicts the best.
We have list of models such as KNN, D-tree,Log-Reg, Random forest, Support vector machine, XGBoost.

K- Nearest neighbour model gives us the accuracy of 83.
Decision tree classifier model gives accuracy of 68.
Logistic regression model gives accuracy of 66.
Random forest model gives accuracy of 94.
Support vector machine model gives 88.
XGBoost model gives 94 similar to random forest.

After using all lets fit the Random forest with 94% accuracy and see how the results would be when compared to Actual. Look at the table below to see results.

	Actual	**Predicted**
0	1.0	1.0
1	0.0	0.0
2	1.0	1.0

3	1.0	1.0