Introduction

The aim of this project is to understand the machine event rates to improve machine maintenance and provisioning. Classification models namely, KNN, Logistic Regression, and Decision Tree were applied. F1-score was the metric chosen for measuring the performance of the model because as per the problem statement we have to reduce the false Positive and false negative rates.

Methodology

Dataset Information

The dataset used for this problem contains machine event data where 11 independent features are provided and binary target features for event (event = 0 or 1). The dataset contains both numerical and categorical variables with 124494 observations.

Data Exploration

There were two main purposes to perform data exploration. Firstly, to understand the characteristic of the data i.e the type of the values a feature contains, range into which value falls and the distribution of the values in the dataset across the range. The second goal is to understand whether data suffers from any kind of issues like outliers, or missing values. Data quality is checked by measuring the central tendency and variation in each numerical feature. Also, various graphs are generated to understand the distribution of features.

	event	feature1	feature2	feature3	feature4	feature5	feature6	feature7	feature8	feature9
count	124494.000000	124494.000000	124494.000000	124494.000000	1.244940e+05	124494.000000	124494.000000	124494.000000	124494.000000	124494.000000
mean	0.000851	9.940455	12.451524	260172.858025	1.223868e+08	14.222693	0.292528	1.741120	159.484762	0.438792
std	0.029167	185.747321	191.425623	99151.009852	7.045960e+07	15.943021	7.436924	22.908507	2179.657730	11.155386
min	0.000000	0.000000	0.000000	8.000000	0.000000e+00	1.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	221452.000000	6.127675e+07	8.000000	0.000000	0.000000	0.000000	0.000000
50%	0.000000	0.000000	0.000000	249799.500000	1.227957e+08	10.000000	0.000000	0.000000	0.000000	0.000000
75%	0.000000	0.000000	0.000000	310266.000000	1.833084e+08	12.000000	0.000000	0.000000	0.000000	0.000000
max	1.000000	24929.000000	18701.000000	689161.000000	2.441405e+08	98.000000	832.000000	1666.000000	64968.000000	1248.000000

Fig 1: Data Description

From the Fig: 1, it is clearly seen that there were no missing values in the dataset. Comparing the gaps between the median, min value, max value, 3rd Quartile and 1st quartile, it is seen that for the feature3 column there is a noticeably large gap between 1st Quartile(25%) and median. It can be considered as an outlier. But, as we don't have the domain knowledge or know the meaning of each feature, So the outliers are not removed from the dataset. Also, It is difficult to perform

feature engineering without understanding the meaning of the features. So “Date” and “Machine” columns are removed from the dataset.
While exploring it is also observed that the number of events are biased i.e. the number of events containing 1 are far less than event 0. Moreover, Pairplot was generated to see the relationship among the variables. From Fig:2, it is observed that there was a strong positive relationship between feature6 and feature9.

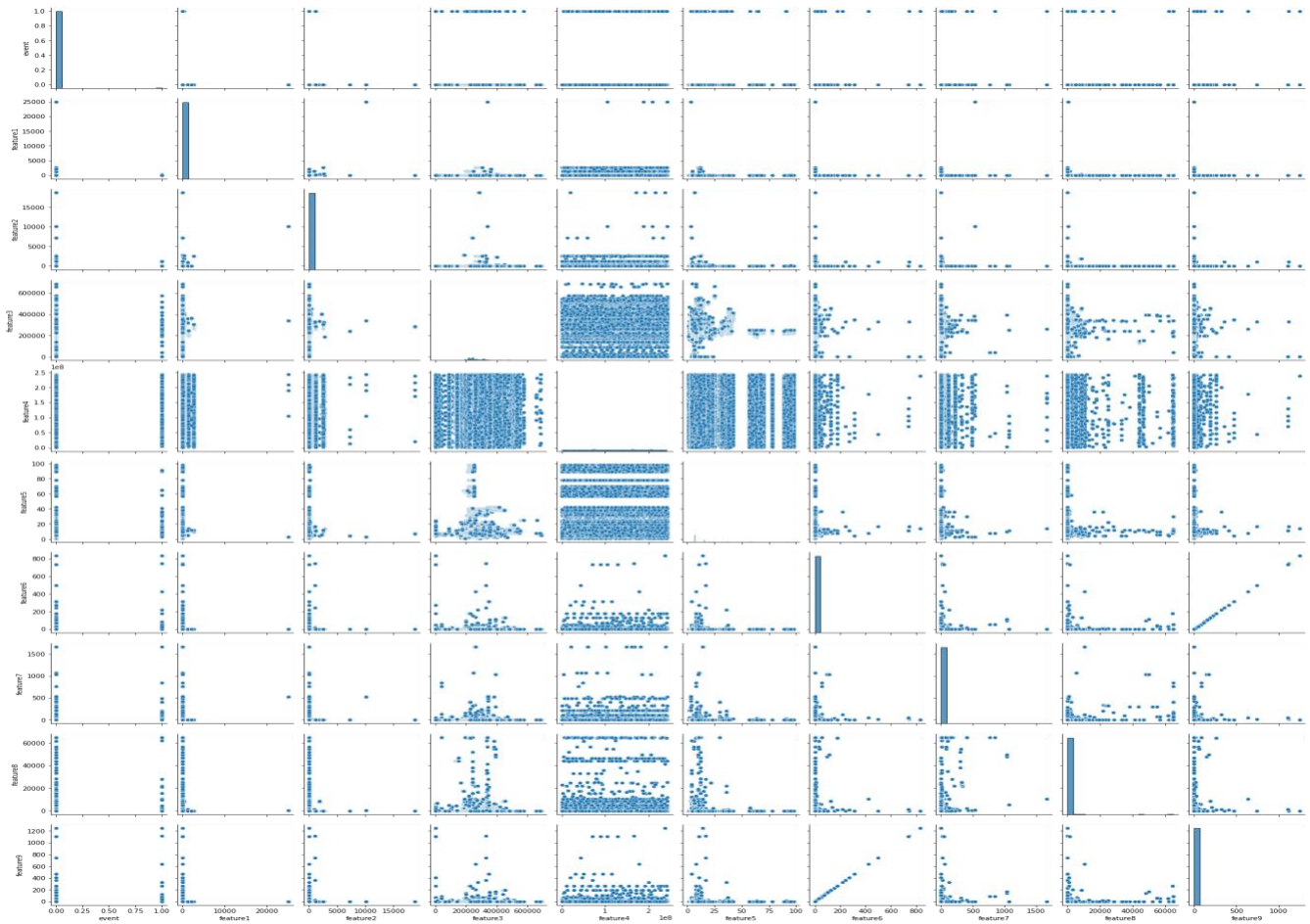

Fig: 2 (Please refer python notebook for high dimensional view)

Data Preparation

Before performing modelling, we need to do Feature Scaling. Scaling of the data is done specifically for the KNN classification algorithm because in KNN the distance between the feature values are measured using Euclidean distances and the scale or units of each feature is very different that may lead biases toward one feature which may affect models performance.

Modelling

As we have seen earlier that the number of class labels are biased toward event 0 which may lead to overfitting. So, we can't directly fit a model on the dataset. We used two approaches to solve this problem namely, up sampling and down sampling. Up Sampling involves randomly duplicating examples from the minority class and adding them to the dataset. Up sampling made the number of target labels of equal size. Three kinds of classification algorithms are applied namely, KNN classification, Logistic Regression, and Decision Tree. For the Up Sampling case, we split the data into train and test sets in a ratio of 70:30. For each split, all the three classifiers were trained and tested. The F1-score of each model was recorded. Table 1 shows all the observed results of the model. Result shows that for KNN Classifiers, there are a lot of biases toward event 0 which leads to overfitting. Also, the Decision Tree model is overfitting on the dataset. Results are changing for Logistic Regression at every run which makes the model less generalized on an unseen dataset.

Second approach used is Down Sampling. In Down Sampling, the number of majority class records are made equal to the size of the minority class. K-Fold cross-validation is applied on this technique to make the model more generalized by removing biases. In K-Fold Cross-validation, we train our model using the subset of the dataset and then evaluate using the complementary subset of the dataset to reduce the overfitting. By using K-fold stratified validation, we fitted 3 models namely, KNN Classifier, Logistic Regression and Decision Tree. From table 2, we can observe that the Decision Tree model outperforms the other two models.

Also, the main reason to not use K-Fold cross validation in Up Sampling techniques is, due to more computation power requirement to fit the model on a large set of data.

Conclusion

After applying 3 classification models by using different approaches, it turns out to be that the Decision Tree with Down sampling approach gives the highest F1 score for predicting the machine events. Decision Tree is the best model not only for its performance but also for its ease to use in the production environment. Also, tree-based models are easy for the stakeholders understanding. However, we can reduce the number of the False Negative and False positive values by increasing the number of dataset observations.

Appendix

Machine Learning Model	F1-Score (event 0)	F1-Score (event 1)
KNN Classifier	0.67	0
Logistic Regression	0.75	0.61
Decision Tree	1.0	1.0

Table: 1 (UpSampling)

Machine Learning Model	F1-Score (event 0)	F1-Score (event 1)
KNN Classifier	0.84	0.81
Logistic Regression	0.75	0.61
Decision Tree	0.91	0.92

Table: 2 (DownSampling)