IJMRT: Volume (5), Issue 6, 2023
International Journal of Multidisciplinary Research Transactions
(*A Peer Reviewed Journal*)
www.ijmrt.in

ISSN (Print):2663-2381
ISSN(Online):2663-4007

Machine Failure Prediction using Supervised Machine Learning Technique

Anist. A[1*], Ganapathy Raja. M[2], Charles. M[3], Vignesh. V[4]

[1]*Assistant Professor, Department of Computer Science Engineering, DMI College of Engineering, Chennai, Tamilnadu, India.*
[2,3,4] *Student, Department of Computer Science Engineering, DMI College of Engineering, Chennai, Tamilnadu, India.*

[*]Corresponding author

DoI: https://doi.org/10.5281/zenodo.7908959

Abstract

We focus on machine failure prediction in industry 4.0.Indeed, it is used for classification problems on the reliability and quality of their machines and products. We compare machine learning methods applied to a difficult real-world problem: predicting machine failure using attributes monitored internally by individual parts. The problem is one of detecting rare events in a time series of noisy and non-parametrically-distributed data. We develop a new algorithm based on the multiple-instance learning framework and the Regression algorithm which is specifically designed for the classification problems, and is shown to have promising performance. Its implementation is modular and extensible to support changes in the underlying production processes and the gathered data. It involves; loading, exploratory data analysis, training and model evaluation. The primary algorithms used in the project is Logistic regression algorithm. It is predictive analysis that describes data and explains the relationship between variables. Our results suggest that nonparametric statistical tests should be considered for learning problems involving detecting rare events in time series data. As large-scale systems continue to grow in scale and complexity, mitigating the impact of failure and providing accurate predictions with sufficient lead time remains a challenging research problem. Developing an accurate failure prediction model requires a critical understanding of

the characteristics of real system failures. Experimental results indicates that the average prediction accuracy of our model using Logistic regression algorithm when failure is 90% accurate. The best performance overall was achieved with Logistic regression algorithm, although computational times were much longer and there were many more parameters to set.

Keywords: Machine Failure Prediction, Time series, cost-sensitive.

1. Introduction

In this paper, we propose a Machine Failure Prediction using the supervised machine learning technique. The failures occurs in the machines were predicted by using the logistic regression algorithm. For the prediction, we have to collect the data about the particular machine like the temperature, humidity, Measure the time of working of that machine. Then, analyze the prediction of that entire dataset to conclude the result. With the development of the Internet, an increasing number of services involve massive data transfer in optical networks. When an optical network suffers a failure, an immense loss of data will occur. To reduce the damage, many optical network protection algorithms have been proposed, including shared-path protection (SPP), best-effort shared risk link group (SRLG) failure protection, and others. However, these algorithms passively protect the optical network and reduce damage only after a failure occurs, which means the data are still lost on account of the time delay of protection and recovery. Therefore, early-warning and proactive protection is required. In [3], risk models are proposed in which high-risk services are switched to a low-risk path to prevent damage from disaster failures in optical backbone networks. In risk-aware models are presented to prevent data loss in data center networks.K-edge and k-node models are proposed to protect optical mesh networks and data center networks from multi-failures (e.g., disasters, massive power outages or mass destruction attacks). The above works provide a means of switching

the services or backing up the data when the risk exists for each link or node (mainly in disaster/attack scenarios); however, they do not consider how to forecast the risk. In fact, a means of predicting an equipment failure in an optical network and providing protective action before a failure occurs remain inadequately investigated. By predicting equipment failures in daily use, the aforementioned protection algorithms based on risk-aware models could thus be extended to daily equipment fault scenarios. Accordingly, the optical network would be more robust and the user quality of experience (QoE) would be greatly improved. Machine learning can be applied to advance the above efforts. Machine learning is a series of intelligent algorithms that can learn the inherent information of the training data. The inherent information is then abstracted into a decision model that provides guidance for further work. These algorithms can perform detection and decision-making in optical communications and improve the system performance. The present authors recently demonstrated their means of reducing nonlinear phase noise, overcoming system impairments in fiber communications, optical performance monitoring and performing data detection in visible light communications.

2. Motivation

Before delving deeper into this problem, we need to understand why it is important to address this problem, where it is used, or where it can be used.

- **Industries-** As we discussed, in industry, it is very important to predict machine failure. They had a system called SCADA which monitors signals and helps to predict the failure of the machine. But when there huge data or the anomaly pattern in data is very hard to detect then SCADA can’t work. Then ML take a step and give a prediction of failure effectively and efficiently. Once a future failure is detected, they provide maintenance to

the machine, it reduces maintenance cost because it provided only when there is a future failure and it also increases the life of the machine.

- **Electricity board-** We can monitor signals taken from the various distribution points of electricity and we can predict failure. It will help to avoid all problems causes due to electricity disconnection in the industry, hospital, etc.

- **Hard Drive Failure Prediction-** Modern hard drives are reliable devices, yet failures can be costly to users and many would benefit from a warning of potential problems that would give them enough time to backup their data. There are various researchers are working on this problem, in this paper they have provides one of the solution to this problem and there are many more solutions are available.

3. Problem Sefinition

I am solving a problem of predicting the failure of a water pump which causes a water supply disconnection. There is a water supply system to provide water to a big town and located far from that town. I have an observation of 5 months in which the water pump get failed 7 times. Those failures cause a huge problem for many people and also lead to some serious living problems for some families. Some people are taking care of that water pump, they tried to analyze all the readings taken from the sensors mounted on a water pump but they failed to make sense out of it to predict the next failure. Hence they proposed this problem to solve by Machine Learning. We have to train a model on the given data and give warning of failure as soon as possible to the person who is taking care of that water pump so that he can take the required step. It is a **binary classification problem** where we have to predict the state of the water pump, is it working normally or it is broken.

4. Data information

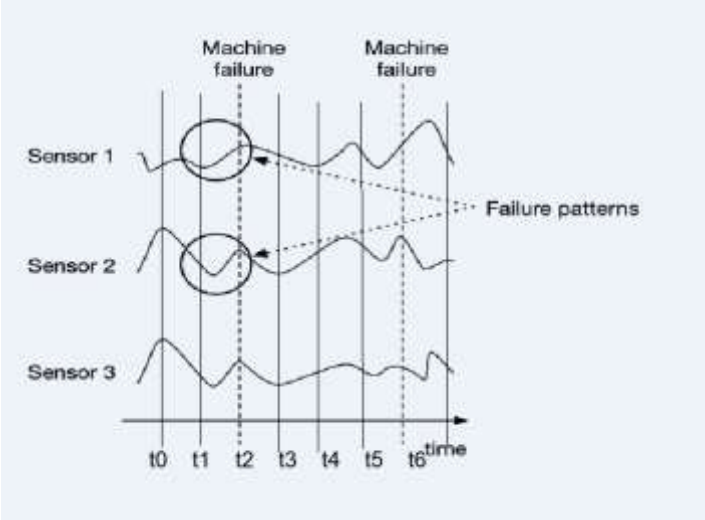


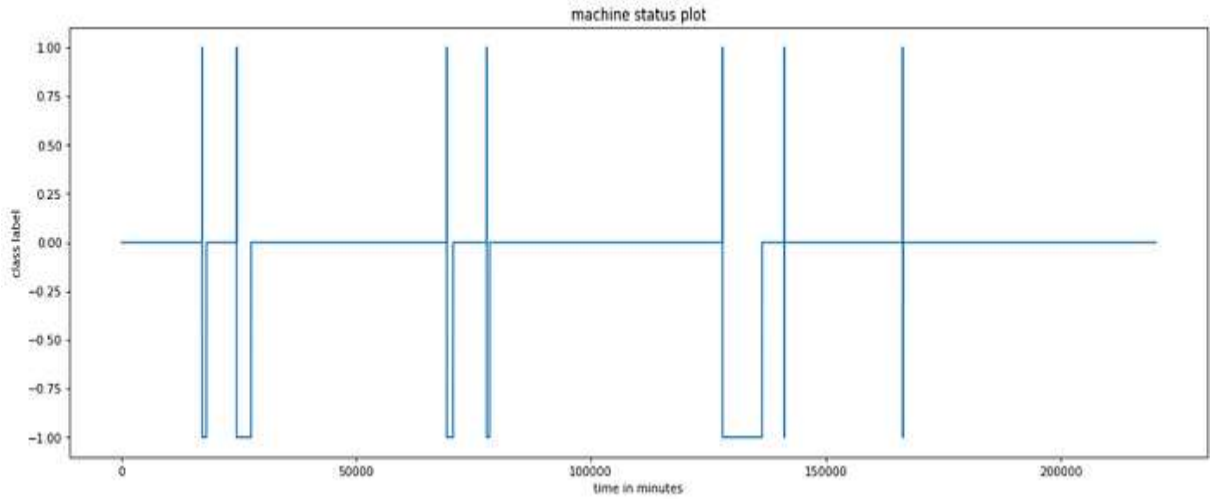
Figure.1. Data Information

As we can see in the figure, we have time from t0 to t6, for each time t we are taking readings from 3 sensors. Each sensor reading varying differently with time. At t2 and t6, the pattern we have recorded is deferent from normal hence there is failure detected. If we tried manually, we can't catch such a pattern, but ML can.

In our dataset, we have 52 such sensor readings along with the corresponding status of the water pump, all readings are taken with an interval of 1 minute. The water pump stays in any one state from 'NORMAL', 'BROKEN' or 'RECOVERING'. 'NORMAL' state means the water pump working properly, 'BROKEN' means the pump got failed and it stopped working, and 'RECOVERING' means a pump is not working and it is under-recovery. There are only 7 points for class 'BROKEN' hence there is a huge imbalance in data.

5. Exploratory Data Analysis

EDA is a way of visualizing, summarizing and interpreting the information that is hidden in rows and column format in data. In our case we should understand things like, how our data actually looks, how sensor reading differ in each state of machine, which pattern appear when there is failure, and so on.




```
for failure 1 at 2018-04-12 21:55:00, recovering time is 15.733333333333333 hours
for failure 2 at 2018-04-18 00:30:00, recovering time is 51.833333333333336 hours
for failure 3 at 2018-05-19 03:18:00, recovering time is 21.866666666666667 hours
for failure 4 at 2018-05-25 00:30:00, recovering time is 10.083333333333334 hours
for failure 5 at 2018-06-28 22:00:00, recovering time is 139.83333333333334 hours
for failure 6 at 2018-07-08 00:11:00, recovering time is 0.6833333333333333 hours
for failure 7 at 2018-07-25 14:00:00, recovering time is 1.25 hours
```

Figure.2. Exploratory Data Analysis

The time period between each failure is different and the time to recover after each failure is also different. We have printed the exact time of failure and time taken by the water pump to recover after that failure. The maximum time for the water pump to recover is 139 hours, 6 days approximately and the minimum time is 41 minutes. We can say that from the different recovery time of the water pump in each failure, the reasons for each failure may vary, so we should not remove any feature, even if we did not find it important.

6. Plotting Some Features and Class Labels Over Time

Here we are going to see how the features are changing with the change in status of water pump. To achieve that view we will plot the feature and 'machine_status' in the same plot with respect to time. We have to normalize our feature because the range of 'machine_status' is -1 to 1 and we do not know about the range of the feature if the range is high then the plot will not be seen properly.

First we have to plot a graph in which we put ‘machine_status’ on y-axis and time on x-axis. After that we will plot any feature eg. ‘sensor_04’.

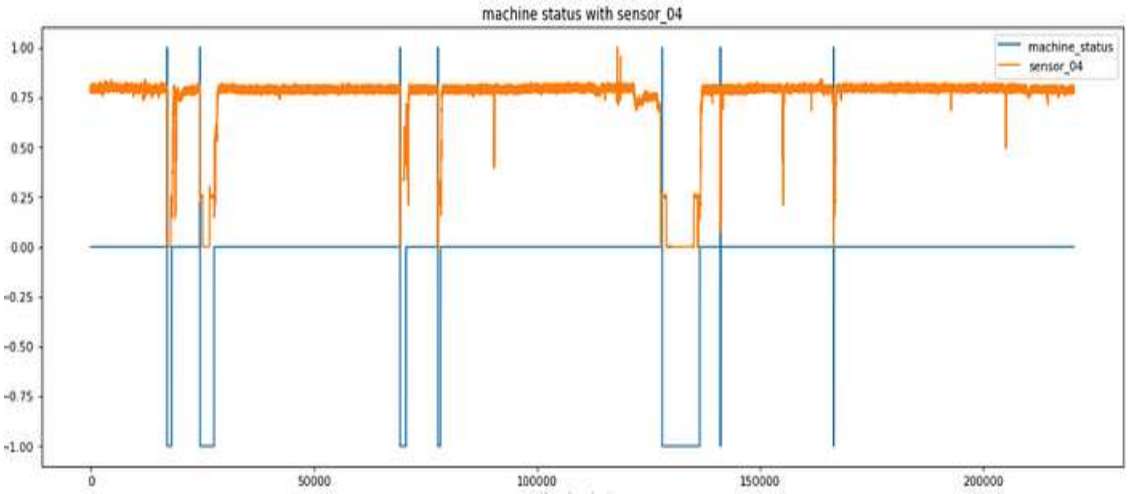


Figure.3. Response

We can see wherever the ‘machine_status’ become 1 means there is a failure, the sensor_04 values suddenly fall to a minimum and it stayed minimum for whole recovering time. It means feature sensor_04 is changing with status of water pump.

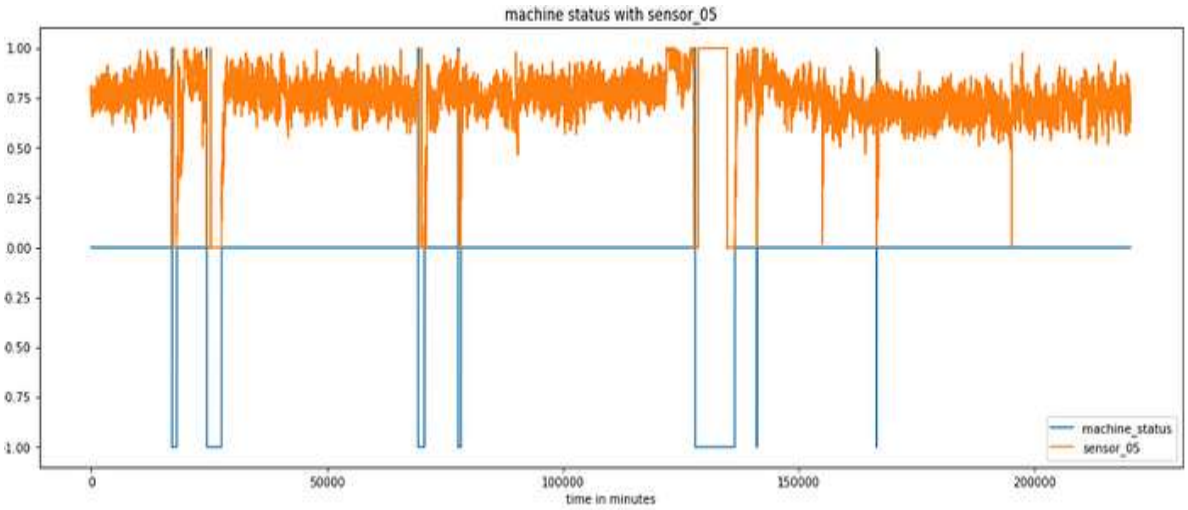


Figure.4. Response

Here for sensor_05 the values fall when a failure occurs and for the time of recovering it stayed at a maximum constant value. Some fall appears without failure still this feature is affected with status of water pump.

There are some features that are not changing with the change in state of the water pump. Their behavior is completely random. Some of them are shown below.

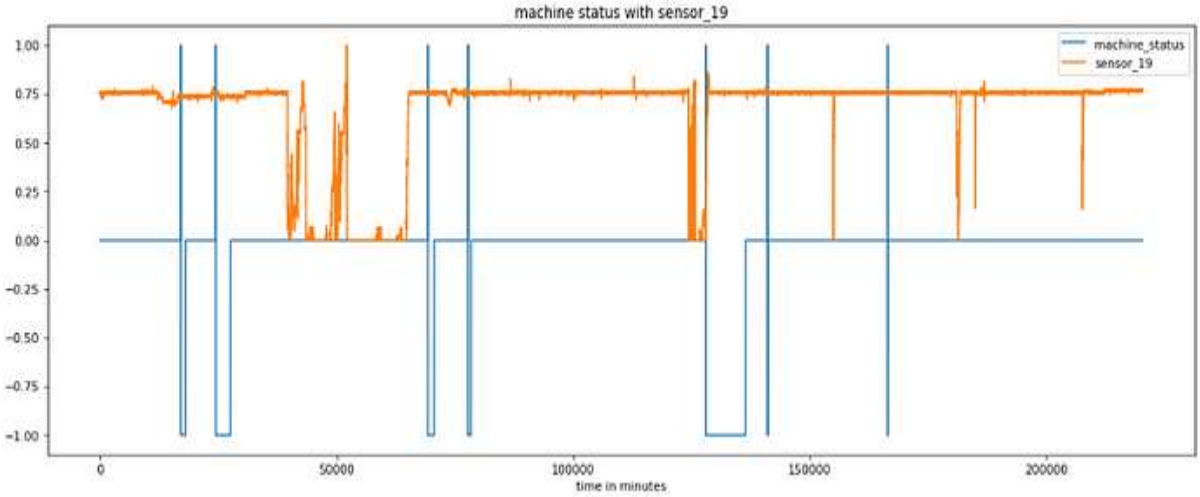


Figure.5. Response

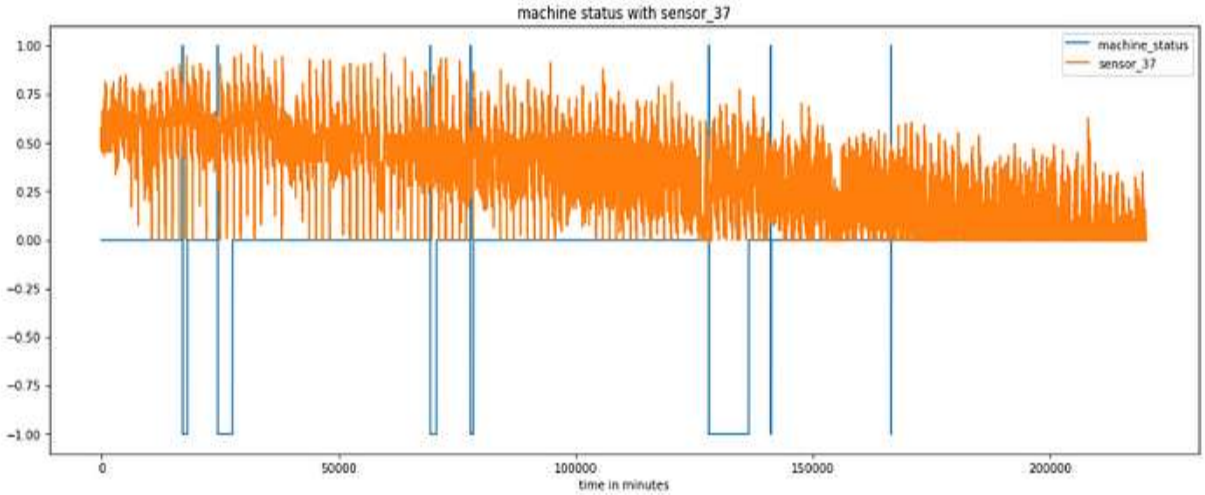


Figure.6. Response

I have plotted this graph for every feature and made a list of features that are not changing with the changing status of the water pump.

7. Filling NULL Values

As we have seen in EDA that, we have NULL values in each feature. We can simply remove those rows which contain NULL values but removing that rows might create loss of information. We have to replace these values but the question is by which we should replace it? If we replace it by 0, but we don't know the meaning of 0 for that feature, hence it is not a good option. We can replace these NULL values with the mean of that feature. I have computed the mean of each feature and replaced NULL values in that feature by the mean of that feature.

8. Performance Metrics

Performance metrics is the term we used to evaluate the ML algorithm. The next step after implementing a machine learning algorithm is to find out how effective is the model based on metrics and datasets. Different performance metrics are used to evaluate different Machine Learning Algorithms.

9. Modelling

I have split data in 50% train 25% cv and 25% test. I kept older data to train and newer to test. By doing simple time-based splitting I got 4 failure points in a train, 2 in cv, and 1 in a test.I have normalized data and train random forest model on train data set and tested cv and test data set.

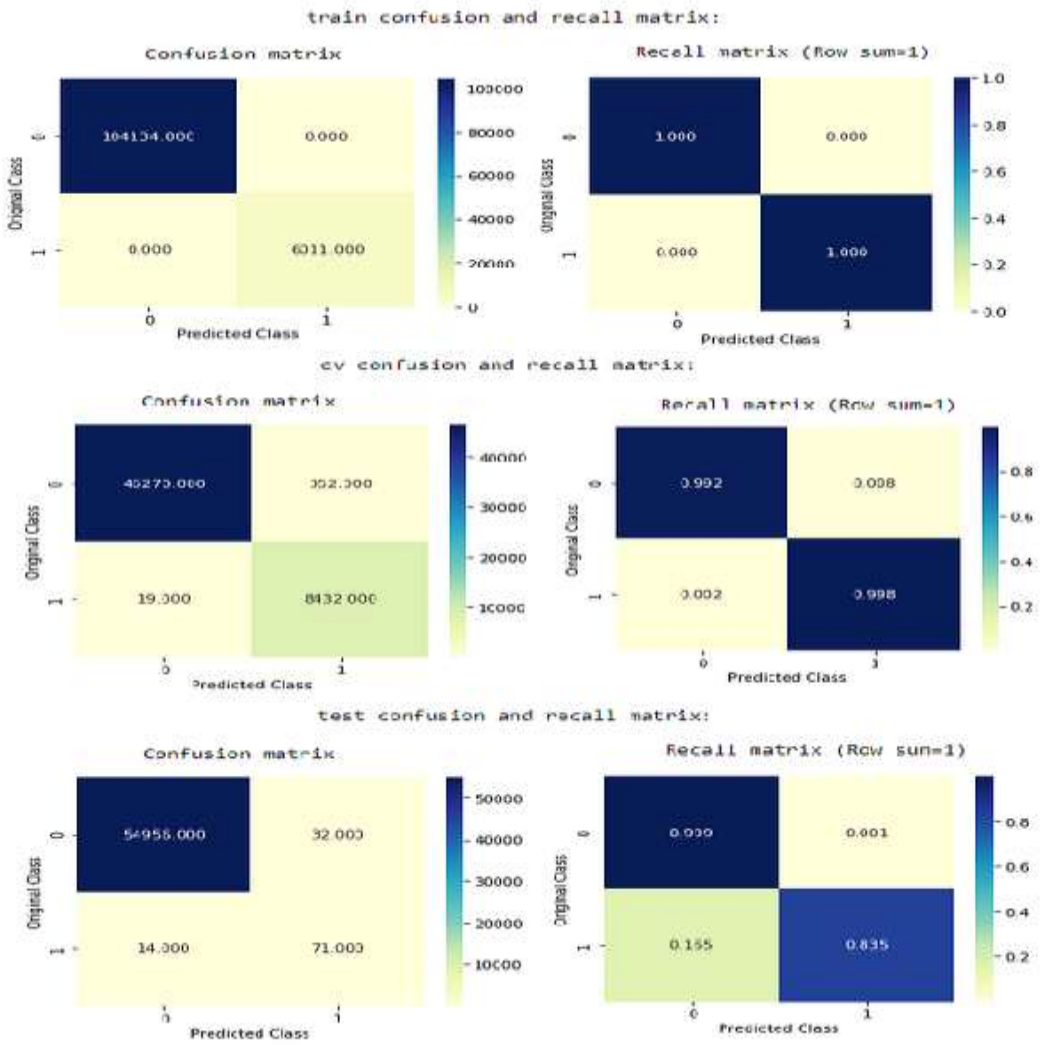


Figure.7. Response

You can see the confusion matrix and the recall matrix I am getting for this model. For train recall is 1, for cv is 0.99 and for a test, it is 0.83. If we observe false positive, it is also low but there is some false negative in cv and test. Since I am using a window of size 10, I got 10 points

for each failure. Out of 10 points, even a single point gets predicted correctly for each failure still we can say, all failures predicted correctly.

10. Table of Results

Whatever the experiments I have done, all are mentioned in this table. I have tried two feature engineering approaches you can see a column of 'feature set approach'. Column 'prediction before' gives the information that how many minutes before I have tried to predict failure. Some model predict all 7 failure but false positive is high. All the cases are mentioned here.

Table.1. All the cases are mentioned

Feature set approach	Prediction before	ML model	Actual failure predicted correctly (out of 7)	AUC score			Recall score			False positive		
				train	cv	test	train	cv	test	train	cv	test
1	60 minutes	Random forest	4	1.0	0.469	-	1.0	0.0	-	0.0	0.0	-
1	60 Minutes	Isolation forest	-	-	-	-	-	0.5	0.0	-	0.106	0.079
1	60 minutes	SVM	4	1.0	0.49	0.5	1.0	0.0	0.0	0.0	0.005	0.0
1	60 minutes	Multiclass random forest	7	-	-	-	1.0	0.73	-	0.0	0.248	-
1	30 minutes	Multiclass random forest	4	-	-	-	1.0	0.0	-	0.0	0.132	-
2	60 minutes	Random forest	5	1.0	0.98	0.46	1.0	0.88	0.081	0.0	0.010	0.003
2	40 minutes	Random forest	5	0.99	0.98	0.51	1.0	0.89	0.32	0.0	0.016	0.002
2	5 minutes	Random forest	7	0.99	0.99	0.99	1.0	0.99	0.83	0.0	0.008	0.001

11. Conclusion and Future Work

Thanks a lot if you have reached here. This is my first attempt in blogging so I expect the readers to be a bit generous and ignore the minor mistakes I might have made.

I have predicted all failure before 5 minutes, with false positive points on 3 dates. 1 in the cv dataset and 2 in the test dataset. For other false positive I have given proper justification.

REFERENCES

[1]. Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Support vector machines for multiple-instance learning. In Advances in Neural Information Processing Systems 15 (NIPS*2002), pages 1–8, Cambridge, MA, 2003. MIT Press. Julius S. Bendat and Allan G.

[2]. Piersol. Random Data. Wiley, New York, 3rd edition, 2000. P. J. Bickel and K. A. Doksum. Mathematical Statistics. Holden-Day, San Francisco, 1977.

[3]. P. D. Bridge and S. S. Sawilowsky. Increasing physicians' awareness of the impact of statistics on research outcomes: Comparative power of the t-test and Wilcoxon rank-sum test in small samples applied research. Journal Of Clinical Epidemiology, 52(3):229–235, March 1999.

[4]. Edgar Brunner, Ullrich Munzel, and Madan L. Puri. The multivariate nonparametric Behrens-Fisher problem. Journal of Statistical Planning and Inference, 108:37–53, 2002.

[5]. Christopher J. C. Burges. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, 2:121–167, 1998. J. Catlett. On changing continuous attributes into ordered discrete attributes.

[6]. In Y. Kodratoff, editor, Proceedings of the European Working Session on Learning, pages 164–178, Berlin, 1991. Springer-Verlag. P. Cheeseman and J. Stutz. Advances in Knowledge Discovery and Data Mining, chapter Bayesian Classification (AutoClass), pages 158–180. AAAI Press, Menlo Park, CA, 1995. Vladimir Cherkassky and Filip Mulier. Learning from Data: Concepts, Theory, and Methods. Wiley, New York, 1998.