Predictive Maintenance Case Study

by:

Philippe Béliveau (#11320035)

Elisabeth Lamarche (#11321255)

Submitted to Mamadou Yamar Thioub

on December 19, 2022

as part of MATH60603A.A22 - Statistical Learning

Table of Content

Executive Summary

With the objective to maximize annual profits, this study aims to identify which of the 100,000 pumps in our installations should be maintained at the 6 month mark. Based on data obtained last year of a sample of 200 pumps and the results of the first 5 months of this year, we determined that 4,046 pumps should be maintained to maximize the expected profit at $300M this year. This decision was based on a comparison of the expected profits if the pumps were maintained at the 6 month mark and the expected profits if they were only maintained at year end. Pumps whose 6 months profits were at least $10,000 more have been identified for maintenance.

Introduction

Our operations rely on the production of liquid by our 100,000 pumps in a remote setting. As access is difficult, we only have the opportunity to check on the pumps mid-year at a non-negligible cost. To find the optimal maintenance policy in order to maximize our profits, our approach is to build models to predict the annual net profits and run a few simulations to select a metric on which to base our decision to maintain each pump at 6 months or not.

Our profits are influenced by three main components: the predicted output of liquid from our pumps, the energy usage from each pump and the types of repairs to be expected. Each of these components are analyzed to understand their past behavior and modeled to predict their future values. Based on the probability of needing repairs and a comparison of the expected profits when a pump is maintained at 6 months with the expected profits when it is not, we aim to determine which pumps should be maintained at 6 months this year.

Data Manipulation

From a study done last year, we have monthly sensor information on the volume produced, the energy consumed and the vibrations of the pump at various frequencies as well as the maintenance schedule and repairs done for 200 pumps, 100 of which were maintained after 6 months.

Pre-processing

In a preliminary review of this data, there are no missing values. Moreover, there are no outliers, all types of repairs are realistic and their costs are in line with the associated repair type. As for extreme values, there are a few pumps in the study that have smaller or larger than expected sensor values, but nothing was significant enough to be removed.

In order to maximize the data available to train our models, the two separate 6 month periods obtained from the data for pumps that were maintained at 6 months were treated as two distinct periods. We assumed

independence in the operational results (volume output, energy usage and repair needs) of the same pump once it was maintained.

Moreover, to clearly identify which type of repairs was done when maintenance was performed a new variable, typeID, was created. This is a numerical categorical variable that takes on value 1 when only maintenance was done, 2 when only seals were replaced, 3 when only bearings were replaced, 4 when both seals and bearings were replaced and 5 when the entire pump needed replacement.

Finally, we separated the maintenance cost of $500 per maintenance from the repair costs to be able to model them separately.

Exploratory Data Analysis

With exploratory data analysis, we aim at building assumptions that would lead us to investigate the dataset in an efficient and structured prognostic of the pump.
Our leading questions were:

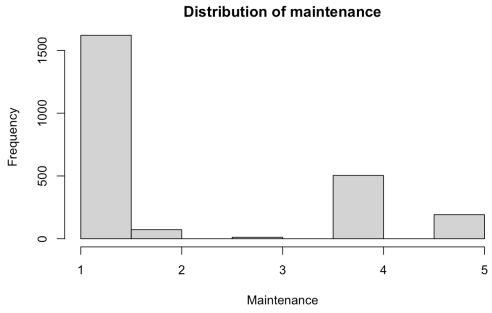
- How to utilize most efficiently the two studies made on 200 pumps to find the relationship between the activity of the sensors, the months, the volume and the energy on the type of maintenance, in order to infer the type of repairs that each of the 100k pumps will require?
- At what point in time should we assign the type of maintenance to a pump? In other words, could we assign, for a pump with a repair of type 5, a type 5 response variable to his 1, 2, 3, etc Month?

As for our leading assumptions:

- The label typeID should not be assigned to all the Months of a pump, as the mapping of the activity of the sensors, volume and energy probably doesn't correspond to the end repair of the pump.
- We assume that the volume and sensors were leading indicators in the health of the pump and that they are highly correlated.

Type of Maintenance

We see that the distribution of the repair is not well distributed, we could already consider using a technique for imbalance data, or simply identifying specific patterns corresponding to the maintenance. Although, we believe it is risky and difficult to infer anything on those pumps with such a little sample for each type of maintenance.



Insight:

The behaviour of the pumps representing the real mapping between input and output starts from months 3.

- For pumps not requiring repairs, there is constant behaviour in the sensor over the whole period.
- Type 2 maintenance show very low level of PSD2000 for the whole period.
- Although there are very few observations, pumps with repair type 3 show a higher range of activity in the sensors levels (high standard deviation)
- Study pumps with repair type 4 show a slight decreasing trend in the volume and energy sensors starting in the later months. This information could be utilized by creating features that capture the mean

volume extracted in the last 5 months. Also, pumps with repair type4 show very similar sensor activity with pump of type1 repair, as both of their sensor distributions follow normality.

- For repair type 5, starting at month 3 we see the pumps declining to low volume of liquid extraction.
- For the various types of repairs, the sensor patterns are similar whether the pump is maintained at 6 months or at 12 months.
- The energy usage is highly correlated to volume extraction with a correlation coefficient of 0.91.

In conclusion, a rule base approach using created features coming from the EDA above could be a more suitable option in distinguishing the type of maintenance. We believe that by using 5 months moving average of the mean, standard deviation, minimum, would capture those normal and abnormal behavior (i.e. high standard deviation, very low value of volume, etc).

Models

Three different prediction models were developed: one for predicting the type of repairs, one to estimate liquid production and another to estimate energy usage.

Type of repairs

Our goal aims at predicting the probability of the type of maintenance for each pump, thus we treat the problem as a multi-class one.

Scaling

We scaled our values in order to bring the energy and volume on a similar scale to the sensors, in order to help our decision criteria utilizing all the values of the sensors. Even though we used a decision tree, which is known to deal well with unscaled data, we derived better results by scaling.

Feature engineering

We created 35 new features that correspond to moving averages of the last 5 months activity for each sensor for each month starting from 5 to 12. We believe that those features will capture the abnormal behaviour of the type of repairs of 2 and 3 due to the high standard deviation that they show and the very low and high value that they reached. We also want to capture the extremely low values of volume with min and use the mean and std to help identify constant behaviour i.e maintenance 1.

**We believe that our feature engineering has drawbacks to consider. As our inputs are the sensor activity of the study dataset from 5 to 12 months, this could compute decisions rules which would not be representative of the 5 months activity of the 100k pump. For example, our decision tree could compute this split criteria for the pumps of type 5 repairs with the volume of liquid extracted between 8 to 12 months, which will be close to zero. While in fact, from the 100k pump of the score dataset, the potential pump with a type 5 repair might not reach as low level of volume extracted as what the model has been trained on. Because, the volume won't have yet reached this level of volume extracted at month 5.*

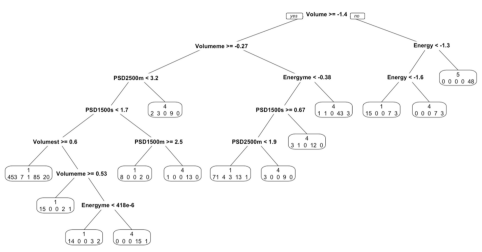
- Mean 5 Moving average for [5, 12] month
- Standard deviation 5 moving average [5, 12] month
- Min value 5 moving average [5, 12] month
- Max value 5 moving average [5, 12] month

Model selection

We then run a feature selection model using AIC where thee procedure chooses the best model given a set of features that minimize the AIC value. This came to improved significantly the accuracy of the model.
We end up choosing a set of 8 features: "Volume", "Energy", "Volumemean", "Energymean", "PSD1000mean", "PSD2000mean", "PSD500std", "Volumestd", "PSD1500std", "PSD2000std", "PSD2500std", "PSD1500max", "PSD2000max", "PSD2500max", "PSD1500min", "PSD2000min"

Train test split

In predictive maintenance problems, we can't randomly split the value, as some measures of one pump can end up in the train, valid and test set. This would lead to overly optimistic prediction, as the model will be seeing similar measurements between the 3 splits and learn from the training and validation the type of maintenance associated with those values. Thus, when predicting on the test set, the model will already have seen similar values and rightly classify the type of maintenance. In other words, the model could then learn to extrapolate between timesteps and make very accurate predictions on the test set. Thus, we split randomly the pump **per ID**. Due to this, Cross-validation is useless here, we then simply divide 70% train and 30% test.



Decision tree

Due to the low number of observations and attempting to establish the decision criteria with simple patterns in the activity of the sensors, we thought a rule base approach would be more robust. Our pruned tree is shown on the left, although, we have a high number of splits for a very few datapoints, which might bring difficulties in generalizing on the test set.

	Predicted				
Observed	1	2	3	4	5
1	576	0	0	10	0
2	11	0	0	5	0
3	4	0	0	0	0
4	112	0	0	108	0
5	27	0	0	7	48

Train set

Our results shows that the model has difficulty understanding the underlying pattern that distinguishes types 4 and 5, this could be coming from their very similar distribution in their sensors activity. On the other hand, our model never predicts 2 and 3 as there were to few observations for those type of repairs. We could think to create more personnalize features able to undercover their specific characteristic.

	Predicted				
Observed	1	2	3	4	5
1	184	0	0	12	0
2	15	0	0	1	0
3	2	0	0	0	0
4	44	0	0	12	0
5	5	0	0	5	12

Test set

Our metrics of interest are specifically towards type 4 and 5 repairs as they are the most costly in terms of repair and profit loss. We thus focus on the False Negative, as we will lose profit when not repairing a type 4 or 5 maintenance pump who actually requires it, because pumps of type 4 and 5 extract very low volume of liquid. In terms of False positive, it is costly to repair a pump that we believe are type 4 or 5, as it might not be advantageous to repair them. Thus, we want to reduce our FP and FN for maintenance 4 and 5. **FP_4: 6.4 %| FN_4: 78% | FP_5: 0% | FN_5: 45%** We see that we predicted 78% of the time that the pump doesn't require type 4 maintenance, while it does, this measure is high and unreliable. For pumps that require a type 5 maintenance, we are right every time we predict it, thus we minimize our

loss! But, 45% of the time we are predicting that the pump doesn't require a maintenance of type 5, while in fact the pumps does require it.

Probability of each type of repair at 12 months

The model allowed us to calculate the probability of requiring certain types of repairs immediately based on the current sensor information. To predict the impact of not maintaining a pump a deterioration factor was established to modify the immediate probabilities. This is to reflect the fact that if a pump isn't repaired in a timely manner, more expensive repairs will be required later.

To do so, we used the study information for the pumps that were not maintained at 6 months. First, we established the 6-month probabilities of repairs using the first 5 months of sensor information only and our model. Then, knowing what the repairs were at 12 months, we calculated the deterioration factor based on the difference of these probabilities. These deterioration factors were then applied to the probabilities for each type of repairs for each of the 100,000 pumps. The sum of the probabilities by pump was then normalized to one. The deterioration factor is approximately 25%, meaning that in our model, the average cost of repairs is 25% higher when waiting until year end to do the maintenance.

Liquid Extraction

When examining the liquid production of the pumps in the study, it was clear that the volume output was lower for pumps requiring more complex repairs. Unfortunately, there were no other identifiable trends found and the liquid production did not seem to follow any known distribution.

As the best prediction of the random variable liquid extraction is the conditional mean of the liquid knowing the type of repair, we made use of the bootstrap resampling simulation method to understand the underlying distribution of the mean liquid production by type of repairs required, where 1,000 simulations were done for each type of repairs. Due to the limited number of observations available for repairs where a replacement seal or bearings were required as well as the lack of clear differences in their volume output, these types of repairs were combined with the group not requiring any repairs. The table to the right presents the mean liquid extraction for the grouped repair types.

Type of repairs	Mean Liquid Extraction (Monthly cubic meter)	Standard Error of Mean Liquid Extraction
1: No repairs, 2: Seals only, 3: Bearings only	187,800	3,000
4: Seals and Bearings	153,000	4,900
5: Whole pump	60,300	8,000

Energy Usage

The energy usage was shown to be highly correlated with the volume of liquid extracted. In order to predict energy usage for the remaining months of the year, a simple linear regressions model, examining the impact of volume against energy usage, was built using 80% of the study data to train the model. The resulting R^2 measure of 83% as well as a mean square error on the test data of 0.055 (compared to 0.048 on the training data) gave us confidence that this model was appropriate to predict energy usage by month.

Profit Analysis

To estimate the expected profit for the year, the following formula was used: the liquid revenues less the sum of the repair costs, maintenance costs and the energy costs.

Liquid Revenues: The liquid that is extracted from the pumps is worth $0.03 per cubic meter. The revenues for the first 5 months was calculated using the pump data retrieved for all 100,000. For the following 7 months, a random liquid production estimate for each pump and for each duration (6-month or 12-month maintenance) was generated using the mean and standard error of the average liquid production by type of repairs. This output was multiplied by the probabilities of each type of repairs to obtain the expected revenues per pump.

Repair Costs: The repair costs occur up to twice a year, only when maintenance is done. The cost of repairs follows a fixed cost schedule as follows: $4,000 for whole pump replacement, $250 for replacing bearings and seals, $200 to replace only bearings and $100 for only seals. To predict the expected repair costs for each pump, we multiplied the probabilities of requiring the type of repair by the cost of each type of repairs. For pumps with maintenance at 6 months, a year-end repair cost was also established as the average cost of repairs after 6 months when no pump-specific information was available to predict the following 6 months. From our first 6-month predictions, this average repair cost is $370.

Maintenance Costs: The maintenance cost refers to the base cost of having a technician maintain the pump: it excludes the cost of repairs and is charged even if no parts need to be repaired. This cost differs based on when the maintenance is done. All pumps are maintained every 12 months at a cost of 500 $: for the 100,000 pumps, this is a cost of $5 million. For the pumps that are maintained at 6 months, the first 20,000 maintenance are at a cost of $500 per pump and every additional pump at a cost of $5,000 per pump.

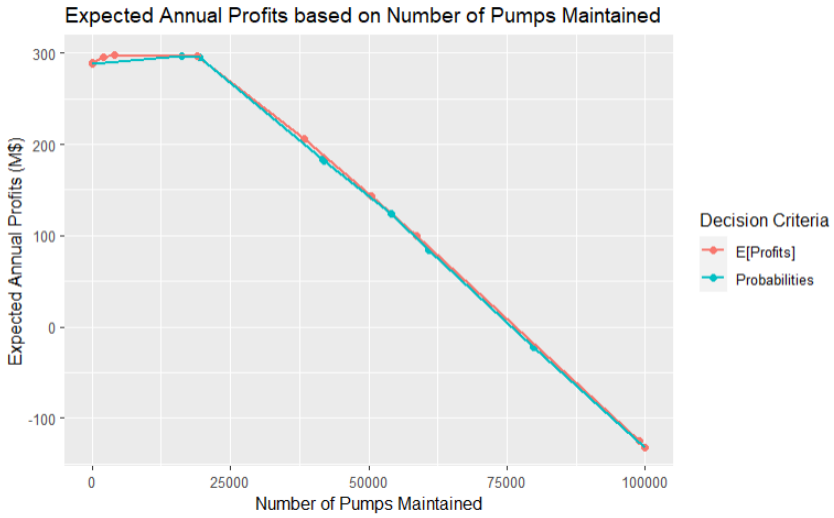
Energy Costs: The electrical energy used costs $0.10 per kwh. For the first 5 months, the actual consumption was used to calculate the costs. For the 7 additional months, the energy usage was estimated using the linear regressions model based on the predicted liquid extraction volume per month. Again, this was done for both scenarios, one where the pump was maintained at 6 months and one where the pump was only maintained at 12 months.

Profits Simulation

With the profit predictions methodology, for each pump we obtained two values: the expected profit for the year if the pump was maintained at 6 and 12 months and the expected profits for the year if the pump was maintained only at 12 months. In order to find the optimal policy, two simulations were done using different decision criteria to decide which pump should be maintained.

Decision criteria 1: Maintain the pumps at 6 months if the probability of requiring a repair as 6 months is higher than the designated cutoff. This was the most simple approach and did not consider the costs of repairs or the lost revenue. Several cutoffs were tried.

Decision criteria 2: Maintain the pumps whose expected profit from doing a maintenance at 6 months was higher than the expected profits from not doing a maintenance is more than the cutoff. Again, the several cutoffs were simulated.



Overall, both approaches give the same overall trend in the number of pumps to be maintained. Profits are maximized at approximately $300M when 4,046 pumps are maintained at 6 months. Because of the lack of precision from establishing the 6 month repair type probabilities, it is difficult to obtain more precision from that method. The ideal maintenance policy is based on the expected profits. Pumps will be maintained if their expected profits, when they are maintained at 6 months, is at least $10,000 more than if we wait and only proceed with the maintenance and possible repairs until the end of the year.

Decision Criteria Approach	Cut-off	Expected Profit (M$)	Nb of pumps maintained
Probabilities	1	$289	0
Probabilities	0.979	$297	16,257
Probabilities	0.857	$295	19,398
Probabilities	0.75	$183	41,657
Probabilities	0.4	$182	41,961
Probabilities	0.263	$123	54,092
Probabilities	0.2	$84	60,783
Probabilities	0.167	-$22	79,747
Probabilities	0	-$131	100,000
E[Profits]	50,000	$289	0
E[Profits]	40,000	$289	87
E[Profits]	30,000	$296	2,089
E[Profits]	10,000	$298	4,046
E[Profits]	5,000	$297	18,943
E[Profits]	2,500	$206	38,402
E[Profits]	1,000	$143	50,495
E[Profits]	0	$100	58,636
E[Profits]	-5,000	-$125	98,973
E[Profits]	-10,000	-$131	100,000

Conclusion

Using information from the 200 pump study done last year, we built three models to predict the repair costs, the liquid extraction and the energy usage for this year in order to estimate the profits. According to this methodology, 4,046 pumps should be maintained at mid-year.

However, this model isn't perfect. When looking at the results in table format, we notice that expected profits stay within $5M of the maximum if we maintain between 2,000 and 19,000 pumps. This is quite a range and indicates to us that ideally more time should be spent understanding trying to enhance the models. One of the ways to do this would be to obtain more data to get better predictions of the type of repairs required. This model drives many of the costs and of the potential profit and would greatly impact the quality of the prediction.