# Data Science applied to maintenance planning optimization

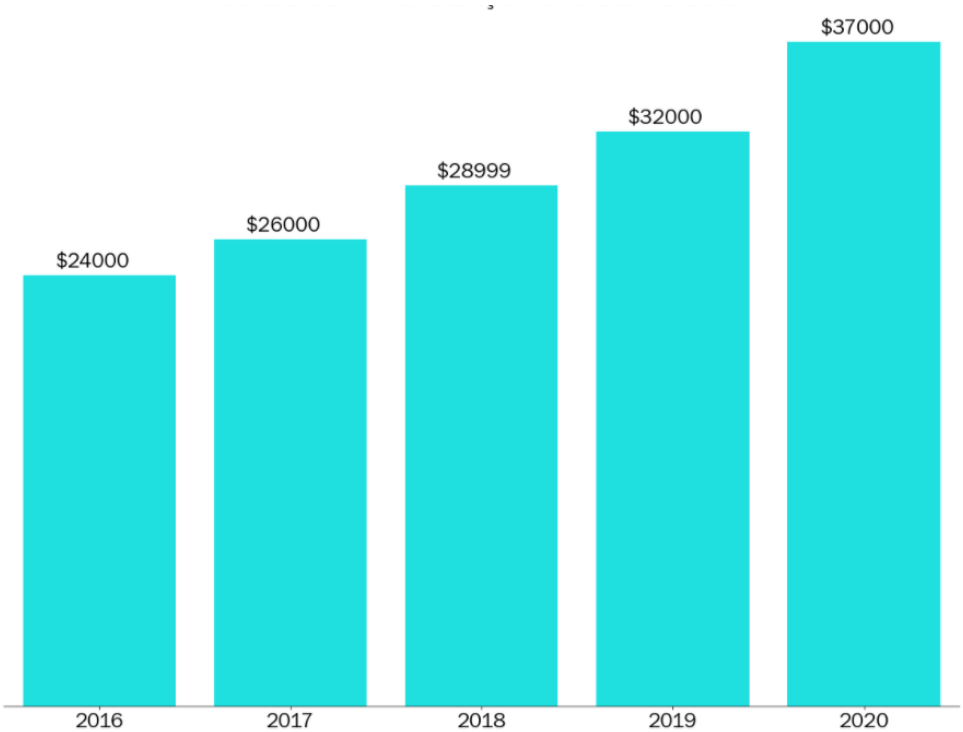
## Answers: by Leandro Souza

## Summary



## Situation

A new data science consulting company was hired to solve and improve the maintenance planning of an outsourced transport company. The company maintains an average number of trucks in its fleet to deliver across the country, but in the last 3 years it has been noticing a large increase in the expenses related to the maintenance of the air system of its vehicles, even though it has been keeping the size of its fleet relatively constant. The maintenance cost of this specific system is shown below in dollars:



Your objective as a consultant is to decrease the maintenance costs of this particular system. Maintenance costs for the air system may vary depending on the actual condition of the truck.

- If a truck is sent for maintenance, but it does not show any defect in this system, around $10 will be charged for the time spent during the inspection by the specialized team.
- If a truck is sent for maintenance and it is defective in this system, $25 will be charged to perform the preventive repair service.
- If a truck with defects in the air system is not sent directly for maintenance, the company pays $500 to carry out corrective maintenance of the same, considering the labor, replacement of parts and other possible inconveniences (truck broke down in the middle of the track for example).

During the alignment meeting with those responsible for the project and the company's IT team, some information was given to you:

- The technical team informed you that all information regarding the air system of the paths will be made available to you, but for bureaucratic reasons regarding company contracts, all columns had to be encoded.
- The technical team also informed you that given the company's recent digitization, some information may be missing from the database sent to you.

Finally, the technical team informed you that the source of information comes from the company's maintenance sector, where they created a column in the database called **class**: "pos" would be those trucks that had defects in the air system and "neg" would be those trucks that had a defect in any system other than the air system.

Those responsible for the project are very excited about the initiative and, when asking for a technical proof of concept, they have put forth as main requirements:

> Can we reduce our expenses with this type of maintenance using AI techniques?

Yes, we can reduce maintenance costs by using AI techniques, predicting failures and performing preventive maintenance, avoiding costly corrective maintenance.

- Can you present to me the main factors that point to a possible failure in this system?

Some "Best Features" remain consistent between the datasets of the two datasheets (columns: ah_000, bb_000, bu_000, bv_000, ci_000, and cq_000).

Features identified as important earlier (columns: aa_000, aq_000, bj_000, bt_000) do not appear among the best features in the current data.

New characteristics were identified in the current data (columns: an_000, ao_000, bg_000, by_000), which may have relevance in predicting system failures.

## About the database

Two files will be sent to you:

- *air_system_previous_years.csv*: File containing all information from the maintenance sector for years prior to 2022 with 178 columns.
- *air_system_present_year.csv*: File containing all information from the maintenance sector in this year.
- Any missing value in the database is denoted by *na*.

The final results that will be presented to the executive board need to be evaluated against *air_system_present_year.csv*.

## Challenge Activities

To solve this problem we want you to answer the following questions:

1. What steps would you take to solve this problem? Please describe as completely and clearly as possible all the steps that you see as essential for solving the problem.

    1.a. Fully understanding the problem and objectives of the project.

    1.b. Data exploration and preparation:

    1.b.a. Analyze the air_system_previous_years.csv and air_system_present_year.csv files to understand the structure, available variables, and missing data ('na').

    1.b.b. Clean and pre-process data, treating missing values and making necessary adjustments.

    1.c. Descriptive analysis:

    1.c.a. Investigate the distribution and identify patterns and anomalies of the cost data.

    1.d. Resource Engineering (Optional): Note: it will not be applied in this exercise.

    1.d.a. Create new variables relevant for analysis (age of trucks, mileage, etc.).

1.e. Predictive Modeling:

1.e.a. Select and train a model to predict future costs or identify potential problems. Validate the model using current year data.

1.f. Optimization:

1.f.a. Pre-processing

Scaling and dimensionality reduction to train and evaluate a model.

1.f.b. Balancing of classes

Class balancing is an important technique in optimizing machine learning models, especially when there are unbalanced datasets. Techniques like SMOTE and SMOTEENN can help improve performance.

1.f. Model Evaluation:

1.f.a. Evaluate model performance and adjust parameters to improve accuracy. Use appropriate metrics (accuracy, recall and F1-score) for evaluation.

1.g. Interpretation and reporting of results:

1.g.a. Interpret the results to identify critical factors related to maintenance costs (Confusion Matrix; Bar Diagram).

1.h. Critical Factors:

1.h.a. Evaluate the performance of the model showing the relationship between the rate of true positives and false positives for future work (ROC Curve).

2. Which **technical** data science metric would you use to solve this challenge? Ex: absolute error, rmse, etc.

I would use Precision, Recall and F1 Score to evaluate the model. Accuracy indicates the accuracy of positive predictions, Recall

measures the ability to identify all positive instances, and the F1 Score combines both, balancing Accuracy and Recall.

3. Which business metric ***would*** you use to solve the challenge?

   I would use the "Cost Reduction Percentage", which measures the savings in relation to previous costs, and the "Absolute Economy", which calculates the cost reduction in monetary values. These metrics provide a clear view of the financial impact of the improvements implemented.

4. How do technical metrics relate to the business metrics?

   Technical metrics help improve accuracy in identifying failures, which in turn directly impacts business metrics by ensuring interventions are more effective and unnecessary costs are avoided.

5. What types of analyzes would you like to perform on the customer database?

   I would like to perform predictive analysis to identify faults in the air system and financial analysis to compare preventive and corrective maintenance costs.

6. What techniques would you use to reduce the dimensionality of the problem?

   PCA, LDA or t-SNE.

7. What techniques would you use to select variables for your predictive model?

   I would use statistical tests, such as the F test of ANOVA or the t-test, to select variables with significant relation to the target.

   I would use Random Forest to select the most relevant variables.

8. What predictive models would you use or test for this problem? Please indicate at least 3.

   8.1. Random Forest Classifier (RFC):

RFC is a robust choice for classification problems, known to handle large datasets and heterogeneous characteristics well. It can capture complex relationships between variables and is less prone to overfitting due to the average of multiple decision trees.

8.2. Gradient Boosting Machines (GBM):

GBM is a boosting technique that builds decision tree models sequentially, where each new model will correct the errors of the previous model. It is effective in dealing with unbalanced data and can capture nonlinear interactions between variables.

8.3. Support Vector Machines (SVM):

SVM is suitable for binary classification problems and can work well in datasets with high dimensionality. It tries to find an optimal separation hyperplane between classes, maximizing the margin between data points of different classes.

9. How would you rate which of the trained models is the best?

Evaluation of Performance Metrics: It would use metrics such as accuracy, recall, F1-score, roc curve and the accuracy itself for each trained model. These metrics provide a detailed view of how each model is performing in terms of predictive capability.

10. How would you explain the result of your model? Is it possible to know which variables are most important?

To explain the results of your model, i can use performance metrics such as accuracy, recall and F1-score, which indicate how well the model is predicting failures in the truck’s air systems. In addition, it is feasible to identify which variables are most important for these predictions through techniques such as analysis of importance of variables.

11. How would you assess the financial impact of the proposed model?

To evaluate the financial impact of the proposed model, we used the confusion matrix to calculate the inspection and maintenance costs associated with model errors and compared these costs with the costs before implementing the model.

12. What techniques would you use to perform the hyperparameter optimization of the chosen model?

I would use grid search (Grid Search) and random search (Random Search) techniques to find the ideal combination of hyperparameters. In addition, cross-validation can be employed to evaluate model performance with different configurations.

13. What risks or precautions would you present to the customer before putting this model into production?

I would tell the customer that it is essential to ensure data quality, interpret the results with caution, test the model in different data sets, continuously monitor its performance, and respect ethical and privacy standards. It is also important to have backup and recovery plans and provide appropriate training to users.

14. If your predictive model is approved, how would you put it into production?

If approved, I would set up the production infrastructure, integrate the model with the company’s data systems, develop data pipelines, and establish continuous monitoring to ensure performance and adjust as needed.

15. If the model is in production, how would you monitor it?

I set up automatic alerts to detect significant drops in performance or changes in data by monitoring accuracy, recall and F1-score, as well as business-specific indicators such as cost reduction.

16. If the model is in production, how would you know when to retrain it?

Model Performance: i would regularly track model performance metrics such as accuracy, recall, F1-score, and other metrics relevant to the business problem. Significant falls in these metrics may indicate that the model is losing accuracy and needs to be re-evaluated.

Automated Monitoring: i would set up automated monitoring systems that alert you to the need for Retraining based on predefined metrics or changes detected in the data.