**Data Science applied to maintenance planning optimization**

## **Summary**

[Summary](#_wtxxhkyqj45b)

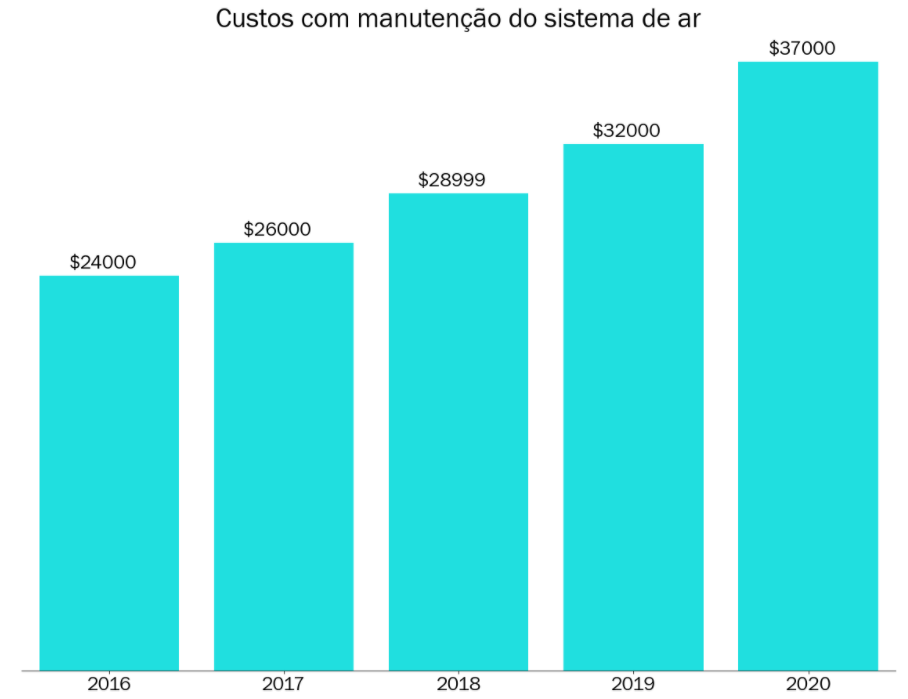
[Situation](#_fqqx5d7nemg)

[About the database](#_juhbkjxqloo1)

[Challenge Activities](#_gdorrczalfhr)

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

These points, according to them, are important to convince the executive board to embrace the cause and apply it to other maintenance systems during the year 2022.

## **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. Explore the datasets to understand the features and target variables so we can begin listing types of models to test (classification ones, regression etc.). That would include analyzing how large the dataset is, the types of data in it (text, numeric, date etc.), if there are null values etc.
   2. Once the type of the target value has been identified and the size of the dataset assessed, time to choose models considering its performance (such as training time) given data size.
   3. Then, data needs to be cleaned. Considering the present case, converting all text to numeric would be a must in order to use the data to train the models. Labels should be converted, as well as numbers that appears to be numeric but are actually text.
   4. ML models won’t accept null values, so we should consider removing certain rows/columns or filling the values. The problems is that some features may be relevant to the model performance even though they are full of null values. Considering the encoded column names, one strategy would be keeping all columns and filling null values with a relevant measurement of central tendence. Both mean and median values could be used. The first, when the values are normally distributed; the second, when they aren’t.
   5. In this case, 170 features may be a lot and might undermine models’ performance. One way to reduce the number of features is leaving just one of the identified correlated columns. In fact, this approach might reduce bias introduction in linear models once they consider the variables are independent from one another.
   6. While exploring data, other issues could arise such as features with huge differences between their variances and class imbalance, as for the provided dataset. Data must be scaled since some calculations are based on distances between points and the class values must be weighted once it is a lot rarer for a truck to be defective than not.
   7. After data has been preprocessed, comparing different suitable models would be a great approach. At this moment, the most important metric must be defined. In the case provided, that would be **sensitivity**. That is because a single corrective maintenance service is more expensive than the inspection of 49 non-defective trucks or the preventive repair service of 19 defective trucks. Therefore, getting most of true positive cases while reducing false negative ones is top priority given that sending a few non-defective trucks for maintenance is a lot cheaper than carrying out a single corrective maintenance.
   8. Once the model with the best metric score has been picked, the last steps to solve the problem would be tuning the model and getting the most important features, as it is stated that highlighting the main factors that point to a possible failure in the system is important to convince the executive board.
2. **Which technical data science metric would you use to solve this challenge? Ex: absolute error, rmse, etc.**

Sensitivity (also known as recall), as stated above.

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

A financial one, such as the cost reduction rate. Showing the actual cost savings that could be achieved by using the model compared to the planned cost savings without it could highlight its potential and help convince the executive board.

1. **How do technical metrics relate to the business metrics?**

Technical metrics assess the performance of the processes executed in a given business, while business metrics reflect its overall success, such as revenue, customer retention rate etc. Once you improve your technical metrics, it would mean improving the processes inside the company, therefore achieving better results regarding its projects, which shall be indicated by business metrics. For instance, achieving a high sensitivity for the proposed model could be translated as a higher cost reduction rate.

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

Descriptive analysis, for better understanding data characteristics, aligned with exploratory analysis since visualizing the measurements and other details from descriptive analysis (such as data distribution, presence of null values, relationships plotted on graphs) is of great help. It would also be reasonable to perform such analysis based on which company site entered the location and when, since different guidelines might be followed in different sites or the way data is entered might change, respectively, what would lead to significant changes in data interpretation.

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

Two approaches could be calculating Person’s correlation coefficient between the features and eliminating one of a significantly correlated pair; another way could be via PCA, which transforms features into a new set of uncorrelated variables while preserving most of the variance.

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

Permutation importance is a good available estimator in sklearn. It randomly shuffles the feature values while keeping other features unchanged, calculates the model’s score, and compares it to the original. The drop in performance indicates how much the model relies on that feature.

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

K-Nearest Neighbors, Logistic Regression, and Decision Tree. All of them can be applied to binary classification situations, such as this one.

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

I would compare both their accuracy and sensitivity values, valuing those with higher sensitivity, as it is the most important metric for the challenge posed.

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

Permutation importance would allow me to explain which features are most important to the model (as it has).

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

Considering the cost reduction rate, as stated in question 3.

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

Two techniques could be GridSearchCV and RandomizedSearchCV. Both of them evaluate different combinations of hyperparameters within a predefined grid, such as combining a range of alpha values with different solvers . The difference is that the first one evaluates all possible combinations, while the combinations of the latter are not exhaustive, but randomly chosen.

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

Some considerations for the customer to take into account include:

* + Ensuring that the model is fed with accurate and representative data. Poor quality data can lead to unreliable predictions.
  + Once deployed, the model should be continuously monitored for performance and updated whenever necessary, as for adapting to new data or aligning to changes in the operating environment.
  + The fact that no errors have been detected in the system doesn’t necessarily mean it is working properly. This is another reason proper ML service is a must.

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

First step would be integrating the model into the existing maintenance system, which could involve APIs, microservices etc. This initial deployment environment would be isolated since testing would still be required to assess its functionality. After testing and validating the predictions, I would gradually roll out the model to a subset of trucks and monitor its performance.

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

Besides continuously tracking performance metrics such as recall and accuracy, it would be necessary to monitor data distribution changes over time. That is because the assumptions made when developing the model may no longer hold true, hindering its performance. Lastly, constant feedback from the maintenance team would be crucial since they are the ones directly dealing with the trucks and capable of assessing if the model is working or not. With their real-life expertise, new meaningful insights could arise.

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

Scheduled retraining is a good approach to improve performance using updated data while letting other departments know when the model would be available or not. Also, when significant changes such as new truck models, system upgrades occur or clear changes in performance metrics/data distribution are detected – always aligned with the ones who rely on the model.