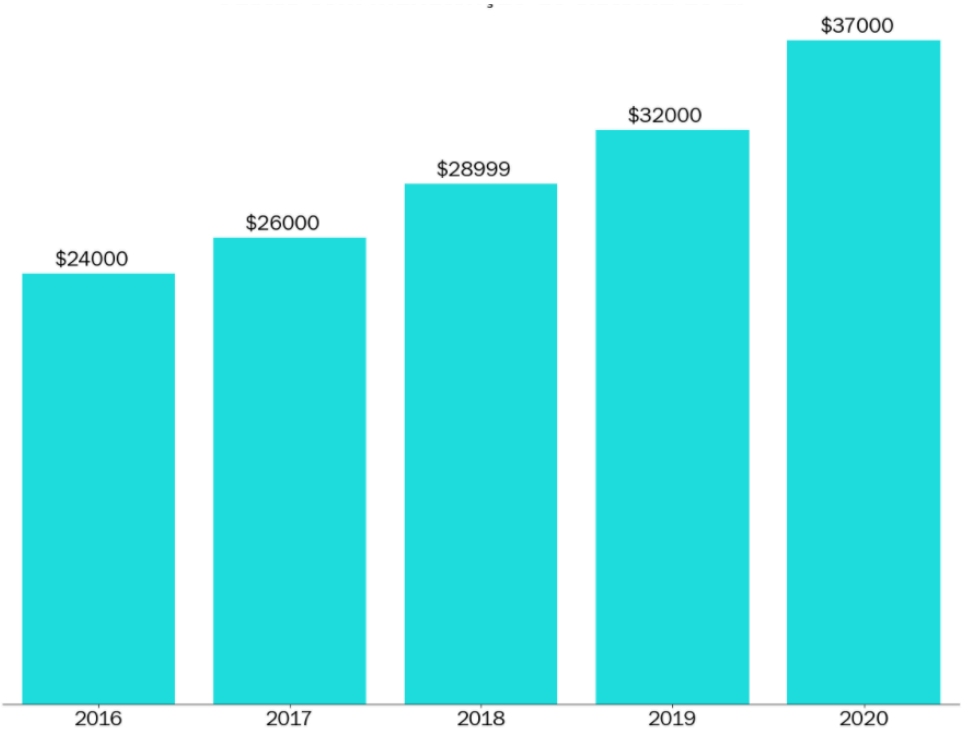
Data Science applied to maintenance planning optimization

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?

- 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.

   First I took a quick look at the data and saw that the class variable had a big imbalance between neg and pos. Also saw that the rest of the data was numerical.

   Then I did some basic data treatments such as encoding the class column, changin the na strings to np.nan, setting the types of the other columns to numerical and lastly dropping columns with too many null values.

   With the clean data, I started searching for the most useful features by checking which distributions changed from class pos to neg. Not only that, but also which ones changed with time, since the number of failures increased even with the same amount of trucks.

   Knowing where to look I then created different classification models including Logistic Regression, Decision Tree, Random Forest and XGBoost and tuned their hyperparameters. I decided to take the XGBoost forward because it showed better results in the test data and kfold cross validation tests.

PS: I defined the penalty in the models according to the different costs of check up and emergency maintenance.

Lastly, I trained my XGBoost model with the tuned hyperparameter and the whole set of data.

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

   I used Gini Impurity and Log Loss as loss functions to train the models, with different weights for false negatives and false positives. But I used a custom loss function based on the confusion matrix and the maintenance costs to tune the hyperparameters and select the thresholds.

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

   I used the maintenance cost, because reducing it would reduce company costs, improve efficiency and most likely help with customer satisfaction.

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

   The false negatives and false positives are unnecessary costs for the company, but a false negative is 50 times more costly than a false positive, so the loss function needs to reflect this scenario.

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

   I basically did KS tests to find out which random variables were related to the classification and which ones were varying over time to try to explain why the trucks were breaking down more frequently.

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

   I dropped variables with too many NAs. Then I dropped variables that were very linearly correlated to another but had a lower correlation to the class variable. Lastly, I dropped variables that

didn’t behave differently depending on the class or dataset (previous or present).

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

   I used ks tests to find out which ones had different behaviors depending on the class and dataset.

8. What predictive models would you use or test for this problem?
   Logistic Regression

   Decision Tree

   Random Forest

   XGBoost

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

   I rated XGBoost because it had a lower loss function value on test data and kfold cross validation using fewer variables,

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

    The loss function was 10 times the number of false positives plus 500 times the number of false negatives. The model had approximately 97.5% accuracy for negative and 94.5% accuracy for positive. Yes, the most important variables were am_0, az_003, ci_000, ai_000.

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

    I would ask for the company’s number of emergency maintenance needed and unnecessary checkups regarding the air system of the trucks and calculate how much money would be saved considering the models accuracy on defining when to do predictive maintenance.

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

I used the GridSearchCV from scipy which tests every combination of a predefined candidate values for each hyperparameter.

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

    I would present the errors to show that it's not 100% effective and there would still exist unnecessary checkups and emergency maintenance, but even though it's not perfect it could still help the company save a lot of money.

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

    I would train it with all the data that the company could provide and deploy it with feedback to monitor if the model is achieving the same efficiency as it was expected. Further details of the deployment would depend on the company's system, hardware and access to the trucks' data, but the main objective would be to use the model to provide actionable insights in time for the trucks to perform predictive maintenance before emergency maintenance is needed.

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

    I would collect environment data such as connection, computation usage, memory availability etc. And also triggers to store data of wrong classifications to try to find patterns in the mistakes in order to further improve the model.

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

    If the loss function increases for over 10% of the baseline or if the accuracy for false negatives drops over 5% and maybe if the independent variables distribution changes.