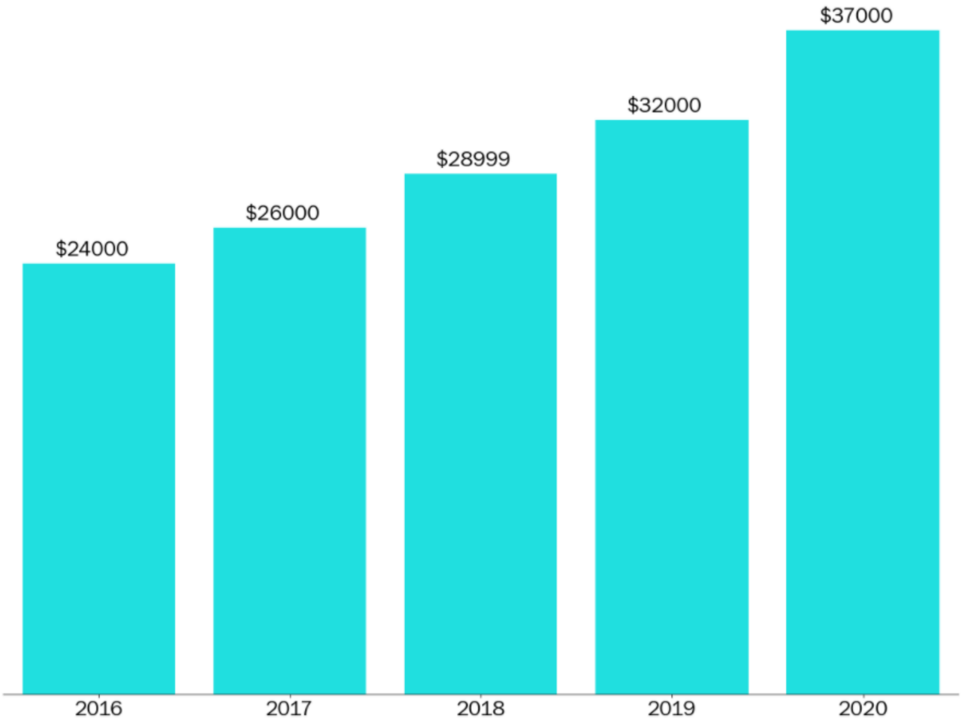
Data Science applied to maintenance planning optimization

Solved Challenge Activities
Postulant: Andrés Gallegos Aguilar

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.**

   To solve this problem, I would start with data exploration by loading the data and reviewing its structure. I would conduct an exploratory data analysis (EDA) to identify patterns and distributions and handle missing values by filling numerical columns with their means. In the data preprocessing step, I would encode the `class` column where "pos" = 1 and "neg" = 0 and normalize the numerical data to ensure all features have the same scale. I would use techniques like PCA to reduce the dimensionality to a manageable number of columns (in this case, 20 components). Then, I would split the data into training and test sets to evaluate the model's performance. I would test different machine learning algorithms such as Random Forest, Gradient Boosting, and SVM. After evaluating the models using metrics like accuracy, recall, and F1-score, I would select the best-performing model. I would analyze the important features of the selected model to understand which variables most influence the predictions. I would calculate the financial impact of implementing the model in terms of preventive and corrective maintenance costs. Finally, I would use hyperparameter optimization techniques like Grid Search to improve the selected model's performance, implement the model in production, and establish a monitoring and retraining system.

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

   I would use classification metrics such as accuracy, recall, and F1-score since we are dealing with a binary classification problem. These metrics will help evaluate how well the model predicts failures in the truck air system.

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

   I would use the reduction in the total maintenance cost of the air system as the business metric. This includes reducing unnecessary inspection costs and corrective repair costs. The goal is to minimize operational expenses related to air system maintenance.

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

   I would perform distribution analysis of the features to better understand the data. I would also conduct correlation analysis between features to identify relationships between variables. I would analyze missing values to ensure they are properly handled. Additionally, I would analyze the importance of features for prediction, identifying the most influential variables.

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

   I would use Principal Component Analysis (PCA) to reduce the number of features to a manageable amount. PCA helps retain as much variance as possible from the original data while reducing the number of dimensions, facilitating modeling and improving computational performance.

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

   I would use Principal Component Analysis (PCA) to reduce the number of features to a manageable amount. PCA helps retain as much variance as possible from the original data while reducing the number of dimensions, facilitating modeling and improving computational performance.

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

   I would use correlation analysis to identify redundant variables. Feature selection based on importance using models like Random Forest is also an effective technique. Additionally, I would employ regularization techniques like LASSO to select relevant features and eliminate those with little or no impact on predictions.

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

   I would test several predictive models for this problem, including Random Forest, Gradient Boosting, and Support Vector Machine (SVM). These models are known for their effectiveness in binary classification problems and can handle datasets with multiple features.

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

   I would evaluate the models using a combination of technical metrics such as accuracy, recall, and F1-score. Additionally, I would perform cross-validation to assess the model's performance consistency. This combination of metrics and evaluation techniques ensures that the selected model not only performs well on the test set but is also robust and generalizable.

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

    I would use model interpretability techniques such as SHAP (SHapley Additive exPlanations) to identify the most important features and explain their impact on the predictions. These techniques allow understanding how each variable contributes to the model's final prediction, providing transparency and explainability for the model's decisions.

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

    I would evaluate the financial impact by calculating the total maintenance cost with and without the model. I would compare the costs of unnecessary inspections and corrective repairs before and after implementing the model. This financial analysis helps determine the potential savings the model could generate in terms of reduced operational costs.

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

    I would use various hyperparameter optimization techniques such as Grid Search, Random Search, and Bayesian Optimization. These techniques help find the optimal combination of hyperparameters that maximize the model's performance, improving its predictive capabilities.

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

    I would present risks such as model overfitting, the need to maintain and update the model periodically, and potential data errors that could affect predictions. It is important for the client to understand these risks to manage them appropriately.

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

    I would deploy the model on a server, integrating it with the company's existing systems. I would establish data pipelines to ensure a continuous flow of information to the model, enabling real-time predictions and periodic updates.

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

    I would monitor the model's predictions and performance in real-time. I would implement alerts to detect potential degradations in the model's performance. Continuous monitoring is crucial to maintaining the model's effectiveness over time.

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

    I would establish performance thresholds to determine when the model needs to be retrained. I would monitor accuracy and other metrics on new data and conduct data drift analysis to identify changes in data distribution that may affect the model's performance.