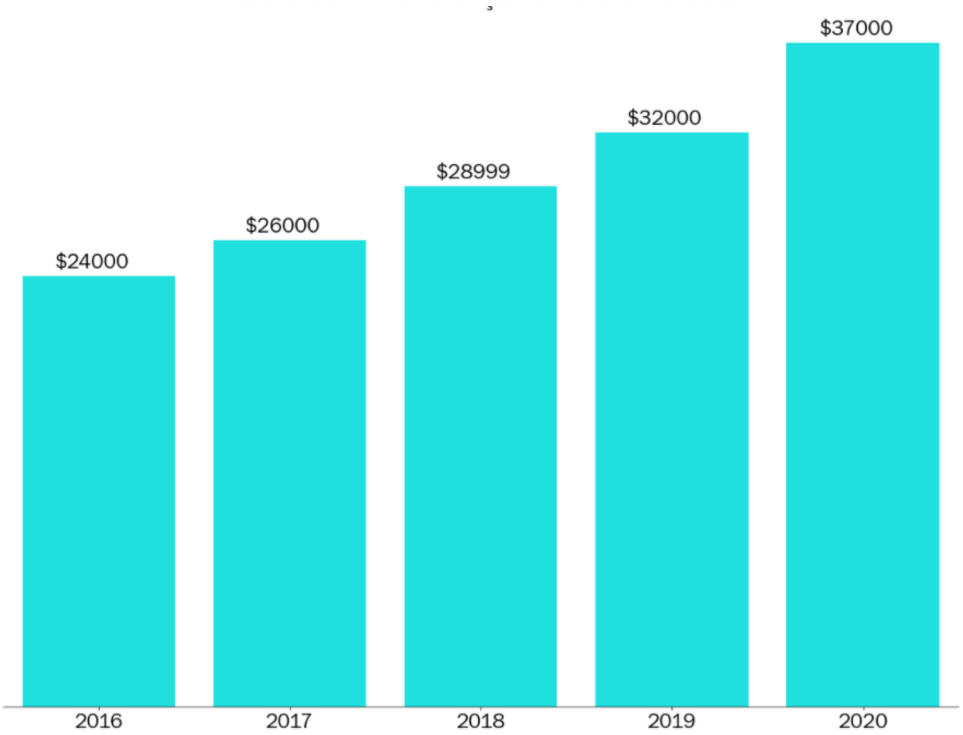
Data Science applied to optimizing maintenance planning

Summary

Situation

A new data science consulting company was hired to solve and improve the maintenance planning of a third-party transportation company. The company maintains an average number of trucks in its fleet to make deliveries throughout the country, but in the last 3 years it has noticed a large increase in expenses related to the maintenance of the air system of its vehicles, even though the size of its vehicle has been maintained. relatively constant fleet. The maintenance cost for this specific system is shown below in dollars:



Your objective as a consultant is to reduce maintenance costs for this specific system. Air system maintenance costs may vary depending on actual truck condition.

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

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

- The technical team informed you that all information regarding the aerial system of the paths will be made available to you, but due to bureaucratic issues in relation to the company's contracts, all columns had to be coded.
- The technical team also informed that given the company's recent digitalization, some information may be missing from the database sent to it.

Finally, the technical team informed that the source of the 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 enthusiastic about the initiative and, when requesting a technical proof of concept, the main requirements were:

- Can we reduce our spending on this type of maintenance using AI techniques?
- Can you present me with 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 2022.

About the database

Two files will be sent to you:

- *air_system_previous_years.csv* : File containing all maintenance sector information from years prior to 2022 with 178 columns.
- *air_system_present_year.csv* : File containing all information from this year's maintenance sector.
- Any missing value in the database is indicated 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 resolve this issue, we want you to answer the following questions:

1. What steps would you take to resolve this issue? Describe as completely and clearly as possible all the steps that you consider essential for resolving the problem.

Steps to Solve the Problem:

Understanding the Problem:

- In-depth understanding of the business.
- Analyze the problem description to understand the costs involved in maintenance (inspection, preventive and corrective).
- Identify the main objective: reduce air system maintenance costs.

Data Collection and Preparation:

- Load the provided datasets (air_system_previous_years.csv and air_system_present_year.csv).
- When there is missing data, the ideal is **to check with the requesting area the reason for the missing data** , and if the lack still persists, we can:
  - Fill in the missing data with:
  - Fixed values (SimpleImputer)
  - Mean, Median and Mode Values (reduces variation in the data set)
  - Prediction (KNN), where data is imputed using neighbors with similar characteristics as a reference .
- class column to numeric values (1 for ' pos ' and 0 for ' neg ').
- Analyze and deal with possible inconsistencies in the data.

Exploratory Data Analysis (EDA):

- Calculate descriptive statistics.
- Visualize distributions and correlations between variables.
- Identify outliers and possible redundant variables.

Choice of features :

- **Feature Engineering:**
    - Create new features relevant to failure prediction, if there is little correlation between features and target.
    - Select important variables using feature selection techniques.
        - In this case, feature importance of the Random Forest decision tree
        - shape values

- **Dimensionality Reduction:**
    - Apply PCA (Main Component Analysis) to reduce the dimensionality of the problem.
    - Use TruncatedSVD if data is sparse.

Data Division:

- Split the data into training and testing sets.

Model Training and Assessment:

- Test different predictive models: Logistic Regression, Random Forest, Gradient Boosting , SVM, for example.
- Evaluate model performance using appropriate metrics.

Hyperparameter Optimization:

- Perform hyper parameter optimization using Grid Search or Random Search, for example.
- There are Auto ML tools that generate several already hyper-parameterized models, simply analyzing their metrics using ML Flow.

Best Model Selection:

- Compare models based on evaluation metrics.
- Select the model with the best performance.

Implementation and Monitoring:

- Deploy the model into production.
- Monitor model performance continuously.
- Define criteria for retraining the model.

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

   Technical Metrics:

   - **AUC-ROC:** To evaluate the model's ability to distinguish between classes.
   - **Precision (false positives), Recall (false negatives) and F1-score (precision X recall relationship):** To balance the trade- offs between false positives and false negatives.
   - **Confusion Matrix**

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

   Business Metrics:

   - **Reduction in Corrective Maintenance Costs:** The main business metric will be the reduction in costs associated with corrective maintenance.
   - **Total Maintenance Cost:** Analyze the reduction in the total maintenance cost (adding inspection, preventive and corrective costs).

4. How do technical metrics relate to business metrics?

   Relationship between Technical and Business Metrics:

   - **AUC-ROC** and other technical metrics help ensure that the predictive model correctly identifies faults in air systems. Better technical performance (higher AUC-ROC, greater precision) leads to more accurate identification of faults, resulting in less corrective maintenance and, therefore, reducing maintenance costs.

   - Considering that the cost of maintaining the air system is R$25.00 and for others it is R$10.00, it is important that the model is accurate in pointing out when the error is in the air systems, so that there are no unnecessary costs. Therefore, it is important that there are **fewer false positives possible (high precision)**

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

   Analysis to be carried out:

   - **Correlation Analysis:** To identify which variables are most correlated with air system failures.
   - **Distribution Analysis:** To understand the distribution of data and identify outliers.
   - **Temporal Analysis:** To check whether there are temporal patterns in air system failures.
   - **Cost Analysis:** To evaluate the relationship between different types of maintenance and the costs involved.

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

   Dimensionality Reduction Techniques:

   - **Main Component Analysis (PCA):** To reduce the dimensionality while maintaining the variance of the data.
   - **TruncatedSVD :** To handle sparse data without centralizing it.

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

   Variable Selection Techniques:

   - **Tree-based Feature Selection :** Using tree models to evaluate the importance of features. Ex : Feature Importance of Random Forest

   - **shape Values** : Help understand the contribution of each feature to a model's prediction

   - **Pearson Correlation:** To identify highly correlated variables and reduce multicollinearity . Example: 2 or more columns with similar characteristics and highly collinear to the target, theoretically it is enough to choose one of them for the model, without loss of performance.

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

   Predictive Models:

   - **Logistic Regression**
   - **Random Forest**
   - **Gradient Boosting Machines (GBM)**

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

Model Assessment:

- **Comparison of Technical Metrics:** Using AUC-ROC, precision, recall, F1-score.
- **Cross Validation:** To ensure the robustness of the models.
- **ROC Curve and Confusion Matrix:** For a detailed analysis of model performance.

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

I would explain by talking about what was measured in the metrics, how many False Positives and Negatives the model predicted, and how many True Positives and Negatives as well. Of course, the lower the False PV, the better the performance of this model. Depending on the project and business model , some False Positive or False Negative has more weight.

- **SHAP Values (SHapley Additive exPlanations):** To explain the contribution of each variable in the prediction.
- **Feature Importances in Tree Models:** To identify the most important variables.

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

- **Scenario Simulations:** Compare maintenance costs before and after model implementation.
- **Return on Investment (ROI) Analysis:** Calculate the ROI of implementing the predictive model.
- **Cost Reduction Estimation:** Project savings in corrective maintenance costs.

12. hyperparameter optimization of the chosen model?

Hyperparameter Optimization Techniques :

- **Grid Search:** To systematically explore a range of hyperparameters .
- **Random Search:** To more efficiently explore a hyperparameter space .

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

Risks and Cautions:

- **Data Quality:** Ensure that data quality and integrity are maintained.
- **Model Bias:** Evaluate whether the model introduces any type of bias.
- **Operational Impact:** Assess the impact on the company’s operations.
- **Model Stability:** Monitor the stability and ongoing performance of the model.

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

Implementation in Production:

- **Production Pipeline:** Develop a robust production pipeline.
- **Continuous Monitoring:** Implement monitoring systems to evaluate model performance.
- **Feedback Loop:** Create a feedback loop to update the model with new data.

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

Model Monitoring:

- **Performance Metrics:** Monitor metrics such as precision, recall, F1-score, AUC-ROC.
- **Automatic Alerts:** Configure alerts for significant drops in performance.
- **Periodic Reviews:** Review model performance regularly.

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

Criteria for Re-training :

- **Performance Monitoring:** Drops in precision, recall, F1-score, or AUC-ROC would indicate the need for re-training .
- **Data Changes:** Significant changes in input data patterns.
- **Feedback Assessment:** Analysis of end-user feedback and impact on operations.