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, we have to make sure that our data is being read correctly. In this example, we need to read the 'na' values as NaN values, not as strings. Second, we need to ensure that our numeric variables are recognized as such, and the same goes for the categorical ones. Once this is done, we have to look for other mistakes in the dataset. For example, we could look for extreme outliers, but that would depend on the nature of the variable. In this example, we don't know anything about the numeric variables, so this step might not be feasible.

Once the dataset is considered reliable, we perform an exploratory data analysis (EDA). We look for correlated variables, variables with many missing values, low variance, whether the groups of interest are unbalanced, and if the response variable is correlated with the occurrence of missing data (as was the case in the assignment). Even though this step is separate from the initial data cleaning, it's important to note that mistakes in the dataset are often found during EDA. At this stage, it's essential to create many graphs to understand the data and the shapes of the groups. Techniques like Multidimensional Scaling (MDS), Principal Components Analysis (PCA), and t-Distributed Stochastic Neighbor Embedding (t-SNE) can be very useful here.

With a clean dataset and a good descriptive analysis, we can start to address the missing data. The amount of missing data, whether it is correlated with certain variables, and the balance of the groups all impact this decision. Generally, we have four options: imputing data to replace the missing values, removing observations with missing data, removing variables with missing data, or ignoring them and using models that can handle missing values. There is no general answer, it all depends on your dataset. It is not possible to generalize how to evaluate the decision, but it is important to know how well that was done.

Then, you can train some models. You have to be careful on how to choose them. If you want to have a good interpretation on the importance of the variables, you should not choose a model such as Support Vector Machine (SVM). But if you only want to make good predictions, then that could be suitable. To choose the model, you also need to consider the amount of information you have, the relationship between the number of observations and predictors, the types of variables, whether the predictors are independent, and many other factors. Another worry is on the tuning of the hyperparameters, which usually is made by cross validation, but each case has its own solution. You also have to be careful on the selection of metrics to determine what makes a model good. On our task, the main problem is evaluating that a truck is in a good state when actually it needs repair, so our most important metric is something like recall, but if that was not the case we should use other metrics.

Finally, we evaluate our models using the test dataset. By comparing them, for example, using recall, we can select the best-performing model. It’s also important to check our hyperparameters, ensure there is no overfitting, and address any other issues that arise.

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

If we consider the null hypothesis as being H_0 = The truck is in well state and the alternative being H_1 = The truck presents malfunction, we can say that the cost of type 1 error is $10 (saying that a truck needs to repair when it is actually perfect), the cost of type 2 error is $500 (saying that a truck is perfect when it actually needs to repair), so it’s really clear that the type 2 error has much more impact on the operation cost. That being saying, we conclude that we should maximize metrics such as recall/sensibility.

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

We can propose some business metric, for example:

- Rate of trucks in repair, that is, number of trucks sent to repair divided by number of defective trucks, which is useful to know if more trucks are being sent to repair than necessary.

- Rate of flaw, that is, number of trucks correctly sent to repair divided by the number of defective trucks, which is useful to know if many broken trucks are being considered as in good condition.

- Money efficiency, that is, how much money needed to be used divided by the total amount of money spent. The numerator is calculated by the formula $25 * number of broken trucks and the denominator is $500 * number errors of type 2 + $10 * number error of type 1 + $25 number of correctly classified as broken trucks

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

The best model has a money efficiency close to 1. As it gets closer to 0, that means we are spending much more money than needed. The rate of flaw is almost the same as recall, but it doesn’t take into account the number of flawless trucks sent for repair. For that measure, we have the rate of trucks in repair, which increases as more flawless trucks are incorrectly classified.

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

   First I would like to know how the data was collected. Then, I would see if the groups are balanced, how much missing data there is, the meaning of each variable, their nature, if there is a variable that we would like to know its effect on the model even if it is not statistically significant and if there is multicollinearity.

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

   As mentioned, we want to identify which variables are more important to classify a truck as defective or not, so we can not use techniques such as PCA. In that sense, if there are many predictors, I would remove the ones that don't have much variance, columns in which the most common variable appears much more frequently than the second most variable, variables with a lot of missing data (more than 10%).

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

   Could be a stepwise selection for logistic regression, or apply penalties such as SCAD or L1 depending on the model

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

   Random forest, logistic regression and XGBoost

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

   The model that, applied to the test data, brings the biggest money efficiency is the best. It's also important to consider how many false negatives our model produces.

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

    For all models that can be done differently depending on what 'most important' means to you. In linear regression, we will consider as most important the variables with lesser p-value. For Random Forest and XGBoost it will be the ones with most impact on Gini coefficient.

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

    By calculating the Money Efficiency, or comparing the cost of the proposed model compared to a baseline model such as a random classifier.

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

    Mostly cross validation, except if there are already knowledge from previous studies indicating what we should use.

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

    First, it would be important to know what the variables mean, as well as how the data was collected and how they plan to collect next. Then, we should compare the efficiency of our model to how they used to detect the broken trucks.

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

    Knowing which variables are most important, we should deploy our model by creating an API.

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

    We can create a threshold for our business metrics and evaluate if our model crosses that

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

    If the metrics are beyond the threshold, then we should retrain it.