Question 1

Let's assume you are responsible for monitoring a simple machine learning model, which is running on a productive cloud cluster. The model is scheduled to run every hour and predicts the average power price for the following day.

- Which would be the most important KPI's to monitor that model?

The best metric or KPI for evaluating a machine learning model vary from each situation in regards of the model type, the data used, and the actual application of the model. Is the model supervised or unsupervised? Is the model trying to predict prices or is it trying to diagnose a disease? What kind of data do we have to build the model?

Regarding supervised regression models the following metrics should be considered:

- Mean squared error. The average squared error between predicted and actual values. Essentially compares each predicted value with the actual one, adding up each of all errors committed and dividing it by the total number of values. Applying a square root to the MSE it is obtained the RMSE which provides with a more standardized metric values.

- Mean absolute error. Calculated as the average in absolute values between predicted an actual values. This metric provides with more resilient results as it is less impacted by outliers among the sample.

Regarding supervised classification models the following metrics should be considered:

- Classification accuracy. Simple proportion of correctly predicted elements and total number of predictions made
- Confusion matrix. In order to analyse True/False positives/negatives, as Classification accuracy does not consider the class distribution.
    - Precision. The true positive ratio (TP / TP +FP)
    - Recall. The true positive class ratio (TP/ TP+FN)

As mentioned earlier, depending on the actual application of the model we may give different priorities between Precision or Recall.

- F1 Score. A combined metric of Precision and Recall (2*Precision*Recall / Precision + Recall)
- Specificity. The true negative ratio (TN/TN+FP)
- ROC Curve. The plot of TPR against FPR of the predicted probability values. This metric allows to set a threshold to define a "requirement" for class label calcifications in which lower values won't be labelled.

Regarding unsupervised models the following metrics should be considered:

- Pearson Correlation Coefficient
- Coefficient of Determination (R^2)

- What part of the monitoring could be automated in order to have the model running in operation 24/7, and how would you do that?

Assuming that the monitoring process is based on keeping track of the metrics and KPI's the main model produces, it is possible to design a secondary ML model which is fed off of the main model's predictions and provide these secondary model's predictions as test/train input to the main model.

Question 2

Imagine that you are responsible to monitor a set of ETL pipelines, which load structured time series data from external sources into the cloud. All of this data is needed as input for ML models, that are running in production.

- What would you do if one of those ETL pipelines starts to fail?

  First of all, I believe the right attitude to approach these situations is to be prepared for the system to fail. Failures should be assumed beforehand in order to properly design the pipeline and anticipate the type of errors the specific ETL in question may arise or the specific problems we want to avoid.

  The system should be designed as error-tolerable while having a parallel monitoring and backup system that analyses why is it failing and why, so a set of automated actions are prepared to be executed when facing an ETL failure.

- What measures could be implemented to make those ETL pipelines as robust as possible?

  A set of monitoring indicators of every technical aspect of the ETL design, i.e. considering node capacity, connection stability, error prompts, response times, etc.

- Which characteristics of the different data sets could be used to evaluate the particular data quality?

  Many problems may arise along the ETL process and of different types, such as schema's (integrity, schema design, embedded values…), data (duplicates, nulls, variety of types, naming or syntax conflicts…)

  The following characteristics are in my opinion the most important to consider while manipulating a dataset:

  - Duplication. Degree of duplicate values along the dataset and the information available regarding this duplicity (i.e. the degree of knowledge we have on why this duplicity happens, how it should be treated, etc)
  - Completeness. Degree of missing values / required data available
  - Accuracy. Degree of confidence about the data source. i.e. the degree of knowledge we have on how was the data treated, manipulated and classified before we obtained it.
  - Readability. The degree of interpretation/information available of the types, measures, features of the data we obtained
  - Consistency. The degree of which the data we obtain is presented in the same format.