Angewandte Digitalisierung in der Industrie

ML Practical Work

Robin Weiss, Tobias Grab*

September 2022

Data Acquisition

One could argue that a data scientist does not need to know how the data is acquired, since she/he has all the powerful tools to squeeze the relevant information out of the data the way she/he wants. While theoretically true, it can help a great deal if the acquisition process as well as the set-up are known. This enables to better identify and handle corrupt data. Since only meaningful data can lead to a meaningful result. In our example the bulk of the data, containing various temperature- and pressure variables, stems from the injection-molding-machine itself. The measured mass however comes from the Mettler-Toledo measurement device. Therefore we have in total two data sources. To match the data coming from these two sources, we can use the serial number per part. (potential error source)

Data Sources *Bilder von den zwei Quellen*

1 Preprocessing

1.1 VIF vs Correlation

In a setting where the number of observations n greatly exceeds the number of predictor variables m ($n \gg m$), we can use the so-called Variance Inflation Factor (VIF) to get a good indicator whether the variable is problematic in the sense of multicollinearity or not. However if we have the situation that $n = 2^m$ or even $n < 2^m$, we cannot rely on this measure since it could lead to too high VIF values coming from overfitted models in the calculation. Therefore we can use a simple correlation matrix to get a first look and manually drop or combine single features.

*with template material from ICAI.

Questions

1. **Correlation**
   a) What does a negative correlation coefficient mean?
   b) How would you deal with a correlation coefficient with say -0.96?

2. **Regression vs. Classification**
   a) Why is a regression model more fitted for this task?
   b) How could you alter the task to make it suitable for a classification task?
   Hint: Quality Control

3. **Categoricals vs. Numericals**
   a) Do we have some categorical input features?
   b) What could you take as a categorical feature?
   Hint: Could be interesting if we get more data

4. Why could multicollinearity pose a problem?

5. **Statistics**
   In the jupyter-notebook you can find the command `data.describe()` which generates a summary about the statistics of the data. As you can see you get some measures for every feature.
   a) Are these measures useful?
   b) What kind of a distribution do you expect for the feature *PowTotAct_Min* and for the feature *Seriennummer*?
   c) Why?

6. What are the different advantages and disadvantages of a linear model?

7. What is an outlier?

8. **Data Cleaning**
   We know we have some corrupt values in the data in the column *Ruecklauf_Temperiergeraet_AS* we know that we cannot use this particular data. Now have to decide whether we delete the entire column or the specific data entries (indexes).
   a) How should we decide?
   b) Why?

9. How many data points would we need if we wanted to fit a model with 10 predictor variables and have at least the same point density as the one we would have with one predictor variable and two points?

$2^{10} = 1024$

10. *We noticed that there are features like *Startzeit* and *Datum*. If we wanted to use them what would be a sensible transformation for them?

    First of all the format is wrong to do some reasonable calculation. Therefore we must transform the data into a float representation. This can be done with the `time`-module or manually. Also, absolute time is not really suited in the ML context. Therefore we need to make it relative. So we could take the time difference between the first "Schuss" of the day and the current "Schuss". In summary we can combine those two feature into a new feature *Timedelta*. With this feature we could detect daily trends in the data (start-up effects) and filter them out.

11. Why is PCA not always the best solution for the problem?