

# Dealing with data scarcity challenges when doing AI or ML experiments within Risk Management

---

Experimenting with AI or ML in the area of Risk management can lead to great results improving the effectiveness and or efficiency of your risk management processes. While it might look great from the outside, working on on such a project is not all glitter and glamour.

Before you can start with any modeling or experiment, you will need to obtain relevant data sets that are of high quality an quantity. As every modeler knows, the effort to collect, clean and prepare data far outweighs the time you spent on the actual modelling itself.

However, in certain areas of risk management, the availability of sufficient or complete data to create robust and predictive models can be already be a challenge by itself. While this might be less of an issue for credit risk related models, specifically areas such as operational risk and compliance can suffer from data that is scarce or incomplete.

Here are some common best practices for dealing with such situations.



## Best Practices for Dealing with Data scarcity challenges

---

### Data Augmentation:

In cases of sparse data, techniques like data augmentation can help enhance the volume and quality of data. This might involve synthesizing data based on existing patterns or using broader external datasets to provide context and fill gaps. Platforms such as MOSTLY.AI, BizDataX or Hazy can help you generate synthetic data while providing various levels of security and privacy protection.

### Utilizing External Data Sources:

Integrating external data, such as industry reports, market trends, commercial risk databases (e.g. Algo FIRST by IBM or OGD from SAS in the case of operational risk) or public financial data, can compensate for the lack of internal data and provide additional insights that enrich analysis.



---

**Advanced Data Imputation Techniques:** Employing sophisticated algorithms to impute missing data can improve the quality of datasets. Techniques like multiple imputation or machine learning-based imputation (using algorithms like k-NN, decision trees, or deep learning) can be effective.

**Partnerships and Collaborations:** Collaborating with other firms, industry groups, or research organizations can provide access to broader datasets and shared learning, which can be particularly useful in fields like (internal) frauds, operational risk events and cybersecurity where intelligence sharing is valuable. Examples of such consortiums are ORX or GOLD.

**Want to learn more? Let's connect.**



[www.redefi.cz](http://www.redefi.cz)



[info@redefi.cz](mailto:info@redefi.cz)