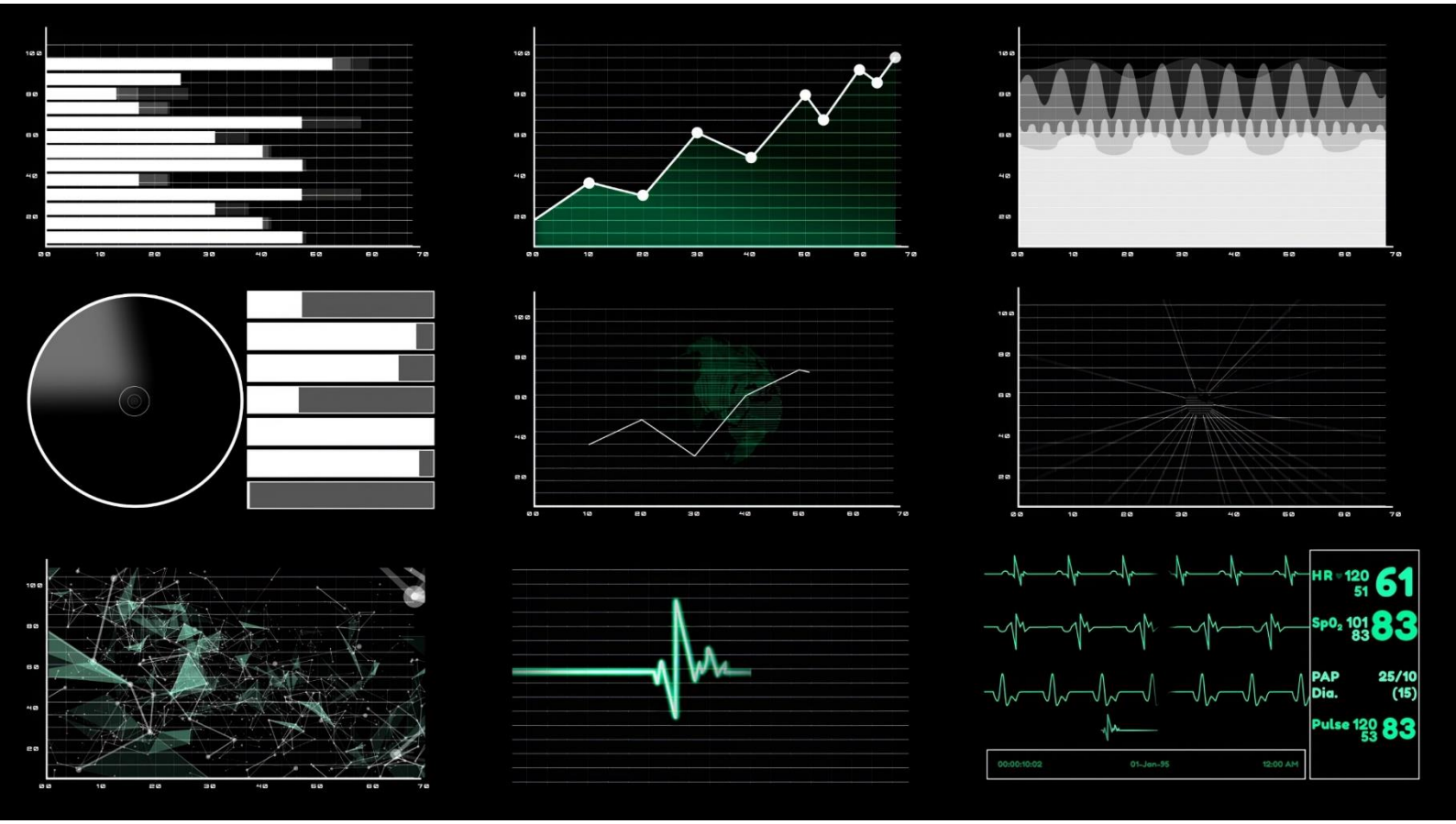
fermac risk

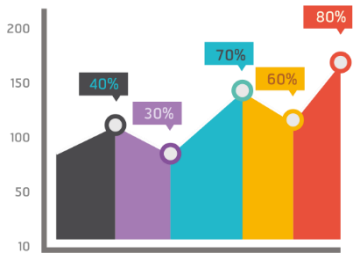
Synthetic Data Strategies for managing Low Default Portfolios

White Paper of Fermac Risk

April 2024



1. Credit Risk in Low default Portfolios



Modeling credit risk in low default portfolios (LDPs) presents unique challenges due to the scarcity of historical default events. LDPs are common in certain sectors, such as sovereign debt, high-quality corporate bonds, or specialized lending, where defaults are relatively rare. Despite the low frequency of defaults, it is crucial for financial institutions to assess and manage credit risk in these portfolios accurately. Here are some approaches and considerations for modeling credit risk in LDPs:

Bayesian Approaches:

- Bayesian methods can be particularly useful for LDPs as they allow the incorporation of prior knowledge and expert judgment into the modeling process.
- Prior information, such as industry experience, economic fundamentals, or qualitative assessments, can be combined with the limited available data to estimate the Probability of Default (PD) and Loss Given Default (LGD).
- Bayesian techniques, such as Bayesian network models or Bayesian hierarchical models, can help quantify the uncertainty associated with the estimates and provide a more robust assessment of credit risk.

Pooling and Data Augmentation:

- Pooling data from multiple sources or across different time periods can help increase the sample size and improve the statistical significance of the estimates.
- Techniques such as data augmentation or simulation can generate synthetic default events based on the characteristics of the LDP and the available historical data.
- However, care must be taken to ensure the relevance and representativeness of the pooled or augmented data to the specific LDP being modeled.

Expert Judgment and Qualitative Factors:

- Given the limited historical data, incorporating expert judgment and qualitative factors becomes crucial in assessing credit risk in LDPs.
- Experts with deep domain knowledge can provide valuable insights into borrowers' creditworthiness, industry trends, and potential risk factors.
- Qualitative assessments, such as management quality, competitive positioning, or country risk, can be systematically integrated into the credit risk models through scoring or rating frameworks.

Stress Testing and Scenario Analysis:

- Stress testing and scenario analysis play vital roles in assessing the potential impact of adverse events on LDPs.
- Institutions can evaluate the LDP's resilience under extreme conditions by defining plausible stress scenarios, such as economic downturns, industry-specific shocks, or geopolitical events.
- Stress testing helps identify potential vulnerabilities, estimate potential losses, and inform risk mitigation strategies.

External Data and Benchmarking:

- Leveraging external data sources, such as credit rating agencies, industry databases, or peer group benchmarks, can provide additional insights and support the modeling process.
- External data can help calibrate internal models, validate assumptions, and assess the relative risk of the LDP compared to similar portfolios or market segments.
- However, the applicability and relevance of external data should be carefully evaluated, considering the LDP's specific characteristics and the institution's risk profile.

Granularity and Segmentation:

- Modeling credit risk at a granular level, such as individual obligor or facility level, can help capture the idiosyncratic risk factors and improve the accuracy of the estimates.
- Segmenting the LDP based on relevant risk drivers, such as industry, geography, or credit quality, can help identify higher or lower risk pockets and tailor the modeling approach accordingly.

- Granular modeling and segmentation enable a more precise credit risk assessment and facilitate targeted risk management strategies.

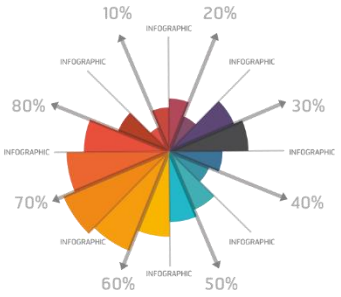
Ongoing Monitoring and Validation:

- Given the limited default history, ongoing monitoring and validation of the credit risk models are essential for LDPs.
- Institutions should establish robust processes to track the models' performance, compare predicted versus actual outcomes, and assess the stability of the risk estimates over time.
- Regular validation, including back-testing and benchmarking, helps identify potential model limitations, calibration issues, or LDP risk profile changes.

Modeling credit risk in LDPs requires combining quantitative techniques, expert judgment, and a thorough understanding of the portfolio's specific characteristics and risk drivers. Institutions should adopt a comprehensive and pragmatic approach, leveraging available data, industry best practices, and domain expertise to develop robust and reliable credit risk models for LDPs.

It is important to note that the choice of modeling approach and the level of sophistication will depend on the LDP's size, complexity, materiality, and the institution's risk appetite, regulatory requirements, and available resources. Engaging with experienced professionals, staying updated with evolving industry practices, and promoting a strong risk culture are key to effectively managing credit risk in LDPs.

2. Classical and generative AI approaches for generating synthetic data



Classical Approaches:

Statistical Modeling:

- Statistical modeling involves analyzing real data's distributional properties and relationships and using statistical techniques to generate synthetic data that mimics those properties.
- Common statistical models include Gaussian mixtures, Bayesian networks, and copulas.
- The generated synthetic data aims to preserve the real data's statistical characteristics, such as means, variances, and correlations.
- A financial institution can use a multivariate Gaussian mixture model to generate synthetic financial transaction data, capturing the distributions and dependencies among variables like transaction amounts, timestamps, and customer segments.

Simulation-based Methods:

- Simulation-based methods create synthetic data by simulating the underlying processes or systems that generate the real data.
- These methods often involve domain-specific models, such as agent-based models, discrete-event simulations, or stochastic processes.
- The synthetic data is generated by running simulations based on predefined rules, constraints, and probabilistic distributions.
- Example: An e-commerce company can use an agent-based simulation to generate synthetic customer behavior data, simulating customer interactions, preferences, and purchase decisions based on predefined rules and probabilistic models.

Generative AI Approaches:

Generative Adversarial Networks (GANs):

- GANs are a class of deep learning models that consist of two competing neural networks: a generator and a discriminator.
- The generator learns to create synthetic data that closely resembles the real data, while the discriminator learns to distinguish between real and synthetic data.
- Through an adversarial training process, the generator progressively improves its ability to generate realistic synthetic data.
- Banks can use GANs to generate synthetic credit card transaction data, capturing patterns and anomalies similar to real transaction data for fraud detection model development and testing.

Variational Autoencoders (VAEs):

- VAEs are generative models that learn a compressed representation (latent space) of the real data and can generate synthetic data by sampling from the latent space.
- VAEs consist of an encoder network that maps the real data to the latent space and a decoder network that reconstructs the data from the latent space.
- By sampling from the latent space, VAEs can generate new synthetic data points that capture the essential features and variations of the real data.

Generative AI approaches offer powerful capabilities for creating realistic and diverse synthetic data by learning real data's underlying patterns and distributions. They can capture complex relationships, handle high-dimensional data, and generate synthetic samples that are difficult to distinguish from real data.

However, generative AI models also have challenges, such as potential biases inherited from the training data, the need for large amounts of representative data for training, and the difficulty in controlling specific attributes or constraints in the generated data.

Both classical and generative AI approaches have their strengths and limitations, and the choice of approach depends on the specific requirements, data characteristics, and available resources of the synthetic data generation task at hand. In practice, a combination of techniques may be used to generate comprehensive and reliable synthetic data.

3. Synthetic Data for Credit Risk

Synthetic data can be a valuable tool for credit risk modeling, particularly in situations where real data is limited, sensitive, or not readily available. Synthetic data is artificially generated data that mimics real data's statistical properties and patterns. Here are some examples of how synthetic data can be used in credit risk modeling:

Synthetic Data in Credit Scoring

- Synthetic data can be effectively used to develop and validate credit scoring models.
- Credit scoring involves assessing the creditworthiness of individuals or businesses by analyzing various factors such as credit history, financial information, and demographic data.

Augmenting Limited Default Data:

- Historical default events may be scarce in portfolios with low default rates, such as high-quality corporate bonds or specialized lending, making building robust credit risk models challenging.
- Synthetic data can be generated to augment the limited real default data by simulating additional default events based on the portfolio's characteristics and expert judgment.

Addressing Data Privacy and Confidentiality:

- Credit risk modeling often involves sensitive and confidential information, such as personal financial data or proprietary business information.
- Synthetic data can be used to create realistic but anonymized datasets that protect the privacy and confidentiality of the underlying individuals or entities.

Testing and Validating Credit Risk Models:

- Synthetic data can be used to test and validate credit risk models under different scenarios and stress conditions.
- By generating synthetic data with specific characteristics or risk profiles, institutions can assess the performance and robustness of their credit risk models under various hypothetical situations.

Benchmarking and Comparative Analysis:

- Synthetic data can be used to create benchmark portfolios or compare the risk profiles of different institutions or market segments.
- By generating synthetic portfolios with specific risk characteristics, institutions can assess their relative risk positions and make informed decisions about risk management strategies.

Generating Rare Event Scenarios:

- Credit risk models often struggle to capture and predict rare but high-impact events, such as extreme default scenarios or tail risks.
- Synthetic data can be used to generate plausible rare event scenarios that are not well represented in the historical data, helping to stress-test and improve the robustness of credit risk models.

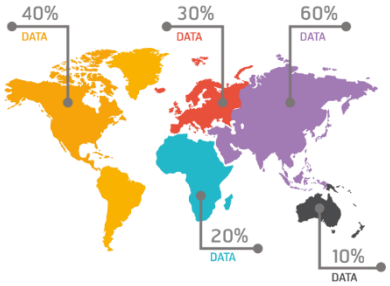
When using synthetic data for credit risk modeling, it is crucial to ensure that the generated data is realistic, statistically sound, and aligned with the characteristics and risk drivers of the actual portfolio. The quality and reliability of the synthetic data depend on the underlying assumptions, models, and expert judgment used in the generation process.

Institutions should validate the synthetic data generation process, regularly review and update the models, and use synthetic data in conjunction with real data when possible. Synthetic data should be considered a complementary tool to enhance credit risk modeling rather than a complete substitute for real data.

It is also important to consider the regulatory and compliance implications of using synthetic data, ensuring that the use of synthetic data aligns with the institution's risk management framework, model governance practices, and applicable regulations.

Overall, synthetic data can provide valuable insights and improve the robustness of credit risk modeling, particularly in situations where real data is limited or sensitive. By carefully generating and leveraging synthetic data, institutions can enhance their credit risk assessment, stress testing, and decision-making processes.

4. Our team at Fermac Risk has extensive experience working with synthetic data for credit risk modeling



A European bank's case study on using synthetic data to improve credit scoring models for personal loans.

Objective:

- A bank wants to develop a credit scoring model to assess the creditworthiness of individuals applying for personal loans.
- The bank has a limited dataset of historical loan applications and repayment records, which may not be sufficient to build a robust and reliable credit scoring model.

Synthetic Data Generation:

Data Analysis:

- The bank analyzes its existing dataset to identify the key variables and relationships influencing credit risk, such as income, employment status, credit history, and debt-to-income ratio.
- Statistical analysis is performed to determine the distributions, correlations, and patterns among the variables.

Synthetic Data Modeling:

- Based on the analysis, the bank develops a synthetic data model that captures the statistical properties and relationships of the real data.
- The model generates synthetic individual profiles with attributes such as age, income, employment status, credit history, and loan characteristics.
- The synthetic data model incorporates realistic variability and ensures that the generated data aligns with the observed patterns and distributions in the real data.

Synthetic Data Generation:

- Using the synthetic data model, the bank generates a large dataset of synthetic loan applicants, including their personal and financial information.
- The synthetic dataset covers a wide range of credit risk profiles, from low-risk to high-risk applicants.
- The bank ensures that the synthetic data is balanced and representative of the target population, considering factors such as demographic diversity and loan characteristics.

Model Development and Validation:

Model Training:

- The bank uses the synthetic dataset to train its credit scoring model, which can be based on various techniques such as logistic regression, decision trees, or machine learning algorithms.
- The model learns the patterns and relationships between the input variables and the credit risk outcomes (e.g., loan default or repayment).

Model Testing and Validation:

- The bank validates the credit scoring model using a combination of synthetic and real data.
- The model's performance is evaluated using metrics such as accuracy, precision, recall, and area under the receiver operating characteristic curve (AUC-ROC).
- The bank assesses the model's ability to discriminate between low-risk and high-risk applicants and its robustness to different scenarios and data variations.

Threshold Determination:

- The bank uses synthetic data to simulate different credit risk scenarios and determine appropriate credit score thresholds for loan approval or rejection.
- The thresholds are set based on the bank's risk appetite, business objectives, and regulatory requirements.

Model Monitoring and Updating:

- The bank continuously monitors the performance of the credit scoring model using real data from new loan applications and repayment records.
- The synthetic data is periodically updated to reflect market conditions, customer behaviors, or loan product changes.
- The credit scoring model is regularly validated and recalibrated using a combination of real and synthetic data to ensure its ongoing effectiveness and reliability.

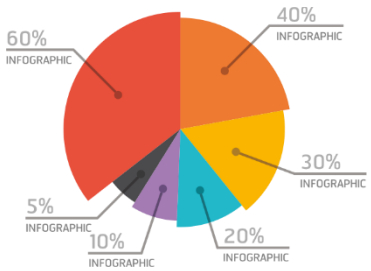
Benefits of Using Synthetic Data:

The bank developed a more robust and reliable credit scoring model by leveraging synthetic data to augment its limited real dataset.

- Synthetic data allows the bank to test and validate the model under various scenarios and stress conditions, improving its resilience and adaptability.
- Using synthetic data helps protect the privacy and confidentiality of real customer information, as the model is developed and tested using artificially generated data.
- Synthetic data enables the bank to generate a wide range of credit risk profiles, ensuring that the model is trained on a diverse and representative dataset.

In summary, synthetic data can be a valuable tool for credit scoring, allowing financial institutions to develop and validate robust credit scoring models even in the presence of limited or sensitive real data. By carefully generating and utilizing synthetic data, banks can improve their credit risk assessments' accuracy, fairness, and reliability while protecting customer privacy.

5. We present important case studies and practical exercises in the course



- **Exercise in LDP:** A banking portfolio of high-quality corporate bonds that has experienced minimal

instances of default in the past. To strengthen the credit risk models and improve the accuracy of Probability of Default (PD) estimates, the bank can create synthetic default data by factoring in various aspects such as credit ratings, industry sectors, and financial ratios. This synthetic data can then be used alongside real data.

- **Exercise in Credit Scoring:** A credit bureau wants to develop a credit scoring model but cannot share the actual consumer credit data due to privacy regulations. Instead, the credit bureau can generate synthetic data that maintains the statistical properties and relationships of the real data without revealing the individuals' actual identities or specific details. The synthetic data can be used to build and validate the credit scoring model while ensuring compliance with data privacy requirements.
- **Exercise in Stress Testing for Credit Risk:** A bank wants to evaluate the sensitivity of its credit risk models to changes in macroeconomic factors. The bank can generate synthetic data scenarios that reflect different economic conditions, such as high unemployment rates, low GDP growth, or increased interest rates. By running the credit risk models on the synthetic data scenarios, the bank can assess how well the models perform under stress and identify potential weaknesses or areas for improvement.
- **Other Exercise in Stress Testing for Credit Risk:** A bank wants to assess the potential impact of a severe economic downturn on its loan portfolio. The bank can generate synthetic data that simulates extreme default scenarios, considering factors such as high default rates, decreased collateral values, and reduced recovery rates. By incorporating these synthetic rare event scenarios into its credit risk models, the bank can better understand its potential losses and develop appropriate risk mitigation strategies.